Feature Relevance is Contextual: A Class-Specific and Cross-Dataset Analysis for Network Intrusion Detection

Daniel M. Martins[1], **Ronaldo R. Goldschmidt**[1], **Ronaldo M. Salles**[1,2]

[1]Defense Engineering Program
Military Institute of Engineering, Rio de Janeiro, RJ, Brazil

{mourao.daniel,ronaldo.rgold,salles}@ime.eb.br

[2]CIICESI ESTG
Polytechnic of Porto, Porto, Portugal

rmo@estg.ipp.pt

Abstract. *Current Network Intrusion Detection Systems (NIDS) often rely on global feature-selection strategies, assuming feature relevance remains stable across attack classes and environments. This paper challenges this assumption, demonstrating that feature importance is contextual, sparse, and class-dependent. A class-wise and cross-dataset analysis investigates feature specialization, semantic consistency, and relevance stability. Results reveal strong intra- and inter-dataset variability, where class-specific relevance discrepancies can exceed 40x for the same attribute, demonstrating that globally dominant attributes fail to preserve discriminative consistency. Preserving class-specific discriminative structures is therefore essential for next-generation NIDS.*

1. Introduction

The rapid expansion of IoT devices and high-speed communication networks has dramatically increased network traffic volume and complexity [Awad et al. 2022, Zhao et al. 2022]. Although this digitalization improves automation and connectivity, it also enlarges the cyberattack surface, exposing infrastructures to increasingly sophisticated threats such as ransomware, zero-day exploits, and Distributed Denial of Service (DDoS) attacks [Almomani 2020].

To mitigate these threats, Network Intrusion Detection Systems (NIDS) analyze network behavior in real time [Almomani 2020, Awad et al. 2022, Göcs and Johanyák 2024]. However, modern network datasets present critical bottlenecks for supervised machine learning models, being inherently characterized by high dimensionality and severe class imbalance [Zhang et al. 2022, Kurniabudi et al. 2020]. These characteristics increase computational cost and often reduce detection performance, as irrelevant or noisy attributes can mask subtle signatures of low-frequency attacks [Tripathi et al. 2024, Göcs and Johanyák 2024, Disha and Waheed 2022].

This challenge is further intensified by the non-stationary nature of network traffic, which continuously evolves due to new applications, protocols, and evasion techniques [Liu et al. 2023, Han et al. 2024]. Consequently, signature-based approaches struggle to detect previously unseen attacks [Scheirer et al. 2013, Li et al. 2024]. Open Set Recognition (OSR), introduced by Bendale and Boult [Bendale and Boult 2016], addressed this

limitation by enabling models to reject unknown classes using Extreme Value Theory (EVT). Nevertheless, while classification architectures have evolved, most NIDS preprocessing strategies still rely on static and globally aggregated feature representations.

Most existing NIDS frameworks evaluate feature relevance globally, assuming that an attribute has uniform importance across all attack categories. Under this perspective, feature importance is estimated from the marginal distribution $P(x)$ of the entire dataset. However, network traffic attributes are inherently class-dependent: features effective for SQL Injection detection may be irrelevant for volumetric DDoS attacks [Göcs and Johanyák 2024, Sarhan et al. 2022]. In contrast, class-specific analysis considers the conditional distribution $P(x|c)$, preserving the statistical behavior and discriminative geometry of each attack class [Niño-Adan et al. 2021]. As a result, global aggregation may dilute localized discriminative patterns, especially in multiclass and imbalanced scenarios [Ning et al. 2026].

Another overlooked issue is feature sparsity. In intrusion detection datasets, several attributes become active only under specific attack conditions, producing highly sparse and event-driven distributions. Although such features may appear globally relevant due to spurious correlations, they often lack generalized discriminative behavior.

To address these limitations, this paper presents a statistical and qualitative analysis of class-specific feature behavior across modern NIDS datasets. Existing studies predominantly rely on global relevance assumptions and often overlook the semantic and contextual properties that shape feature distributions and discriminative behavior. Consequently, feature importance is frequently treated as a purely numerical measure rather than a context-dependent characteristic.

Instead of proposing a new detection model, this work argues that conventional global feature-selection assumptions fundamentally fail under open, imbalanced, and evolving traffic distributions. We demonstrate that feature relevance is inherently contextual and that feature selection in NIDS should be treated as a conditional semantic modeling problem rather than a purely global optimization problem.

The main contributions of this paper are summarized as follows:

- We provide a comparative analysis of feature behavior across multiple NIDS benchmark datasets, highlighting strong intra- and inter-dataset variability.
- We investigate local feature relevance, demonstrating that feature-space geometry is highly class-dependent and that global aggregation attenuates attack-specific discriminative patterns.
- We analyze sparsity and multimodality effects, exposing limitations of conventional global feature-selection metrics in dynamic and imbalanced scenarios.

The remainder of this paper is organized as follows. Section 2 discusses the technical background and related work regarding feature analysis and network intrusion datasets. Section 3 presents the proposed methodology. Section 4 presents a discussion and analysis of the obtained results. Finally, Section 5 concludes the paper and outlines future directions.

2. Technical Background and Related Work

The effectiveness of modern NIDS strongly depends on how network traffic is represented during learning. Existing studies extensively explore feature selection (FS), feature weighting (FW), and dimensionality reduction techniques. However, most approaches remain primarily empirical, validating features through predictive performance rather than investigating their statistical and semantic behavior.

This section reviews benchmark NIDS datasets, summarizes dominant global FS paradigms, and discusses the transition toward local relevance and class-aware weighting. More importantly, it highlights the lack of studies investigating how feature behavior changes across attack classes and datasets.

2.1. Network Intrusion Datasets

Legacy datasets such as KDD'99 and NSL-KDD are now widely criticized for outdated traffic profiles and unrealistic attack distributions. Consequently, recent research has migrated toward more representative datasets [Disha and Waheed 2022, Tripathi et al. 2024]. Among the most widely adopted benchmarks are:

- **UNSW-NB15:** A hybrid dataset composed of legitimate traffic and synthetically generated attacks. Due to its relatively balanced structure and diverse attributes, it became one of the primary benchmarks for evaluating feature selection methods in NIDS [Almomani 2020, Kasongo and Sun 2020].
- **CIC-IDS2017 and CSE-CICIDS2018:** These datasets introduced more realistic background traffic and broader attack diversity, including DDoS, brute force, Heartbleed, botnet activity, infiltration, and web attacks [Göcs and Johanyák 2024, Tripathi et al. 2024].

More recently, standardized NetFlow-based datasets such as NF-UNSW-NB15-V2 and NF-CSE-CICIDS2018-V2 have reduced feature heterogeneity through industry-oriented flow attributes [Sarhan et al. 2021b, Sarhan et al. 2022]. Although this improves interoperability, semantically equivalent features may still exhibit substantially different statistical behavior depending on the dataset and attack context.

2.2. Global Feature Selection Paradigms

Feature selection remains one of the most investigated preprocessing tasks in NIDS, aiming to reduce data dimensionality by eliminating redundant or irrelevant attributes, thus mitigating overfitting and high computational costs [Almomani 2020, Kurniabudi et al. 2020]. Traditional FS methods are commonly divided into:

- **Filter methods:** Statistical and information-theoretic approaches such as Information Gain, Mutual Information, Shannon Entropy, and Chi-square independent of any specific learning algorithm [Kurniabudi et al. 2020, Göcs and Johanyák 2024].
- **Wrapper methods:** Utilize a predictive model as a subroutine to evaluate subsets of features, often using meta-heuristic searches [Almomani 2020, Zhang et al. 2022].
- **Embedded methods:** Integrate FS directly into model training, such as tree-based Gini importance [Disha and Waheed 2022, Tripathi et al. 2024].

A substantial portion of NIDS research has relied on these global FS techniques to optimize detection. For instance, [Almomani 2020] evaluated the impact of various meta-heuristic optimization algorithms to search for optimal feature subsets, including Particle Swarm Optimization (PSO), Grey Wolf Optimizer (GWO), Firefly Algorithm (FFA), and Genetic Algorithms (GA). Their study determined that retaining up to 30 features yielded optimal performance on the UNSW-NB15 dataset. Similarly, [Kasongo and Sun 2020] applied XGBoost to select 19 critical attributes from the same dataset, successfully reducing model complexity without compromising accuracy.

Other work on the CIC-IDS2017 dataset, was utilized Information Gain, one of the most prevalent FS algorithms in intrusion detection, to highlight that Random Forest classifiers could achieve high accuracy with a globally reduced set of 22 features [Kurniabudi et al. 2020]. Other approaches have employed Recursive Feature Elimination (RFE) [Meemongkolkiat and Suttichaya 2021], the Gini index [Moualla et al. 2021], and hybrid ensembles of traditional classifiers (KNN, SVM, Naive Bayes) to establish a global consensus on attribute relevance [Mhawi et al. 2022]. More recently, weighted feature selection frameworks based on optimized Chi-square relevance have also been proposed [Tripathi et al. 2024].

Despite their effectiveness, these approaches share a common limitation: they are fundamentally global. They calculate a single relevance score for each attribute based on the marginal distribution $P(x)$ of the entire dataset, implicitly assuming that a feature's importance remains invariant across all benign and malicious traffic profiles.

Consequently, global FS often favors dominant traffic patterns while attenuating weak but highly discriminative signals associated with minority attacks. This issue becomes critical in imbalanced, dynamic, and open-set intrusion detection environments.

2.3. Toward Local Relevance and Feature Weighting

Unlike traditional FS methods that perform binary inclusion or exclusion, FW assigns continuous relevance scores to attributes. However, a crucial distinction must be drawn between global weighting and local (class-specific) weighting. While global methods assume homogeneous relevance across the entire feature space, local weighting recognizes that feature importance is non-uniform and varies significantly depending on the specific class or subregion of the instance space [Niño-Adan et al. 2021].

Recent studies have begun exposing the limitations of global representations in NIDS. In [Göcs and Johanyák 2024], it was empirically demonstrated that distinct attack categories in CSE-CICIDS2018 required different optimal feature subsets. Similarly, adaptive frameworks such as *CALMID* [Liu et al. 2021], *MicFoal* [Liu et al. 2023], and *AdaAL-MID* [Han et al. 2024] introduced weighting strategies to prioritize minority and drifted classes during online learning.

Outside the traditional NIDS domain, advances in open-set and active-learning studies also reinforce the importance of localized feature representations. CAAL [Yin et al. 2025] proposed spatially realigned feature spaces through adaptive weighting, while [Ning et al. 2026] explored conflicting evidence derived from class-specific feature contributions to identify informative instances in open-set scenarios. Tail-aware weighting strategies were further explored in [Schmidt et al. 2025] for extreme-event detection.

Although these studies introduce local weighting mechanisms, most remain strongly algorithm-oriented. Their focus is typically predictive optimization rather than understanding the intrinsic statistical behavior of features themselves.

Consequently, an important gap persists in intrusion detection research: the absence of systematic cross-dataset analyses investigating feature geometry, sparsity, multimodality, and semantic consistency across attack classes.

2.4. Class and Sample Weighting

Unlike FW, which operates in the feature space, class and sample weighting operate in the instance space to handle data quality issues [Shu et al. 2022, Bouguelia et al. 2016]:

Class weighting assigns weights to entire classes to mitigate the class imbalance problem. In NIDS, higher weights are typically assigned to rare attack classes to prevent the model from biasing toward benign traffic [Shu et al. 2022, Disha and Waheed 2022]. Sample weighting, in contrast, assigns a weight to each individual sample to prioritize informative instances or reduce the importance of noisy/incorrectly labeled observations [Shu et al. 2022, Bouguelia et al. 2016].

Despite the relevance of these techniques in the instance space, this study focuses exclusively on attribute-space analysis. Accordingly, these strategies are outside the scope of this work and are presented only to delimit the boundaries of our study.

2.5. Research Gap and Motivation

Recent advances in intrusion detection increasingly incorporate local weighting and adaptive learning mechanisms. However, the statistical and semantic behavior of features across attack classes remains insufficiently explored. Most studies focus on predictive architectures optimized for specific datasets while overlooking structural properties such as sparsity, multimodality, and conditional feature relevance.

Consequently, the literature still lacks comprehensive cross-dataset analyses capable of moving beyond superficial statistical observations to investigate how feature distributions, relevance, and discriminative behavior vary according to attack semantics and dataset composition.

To address this limitation, this work adopts a class-aware and cross-dataset analytical perspective. Instead of proposing a new predictive architecture, the study investigates the structural behavior of NIDS features through statistical analysis combined with Gini impurity, Mutual Information, and Shannon Entropy. The proposed framework analyzes conditional distributions $P(x|c)$, sparsity patterns, multimodality, and feature stability across datasets and attack classes.

This issue becomes particularly critical in imbalanced, open-set, and continuously evolving environments, where globally optimized representations frequently fail to capture localized attack-specific patterns.

A comparative summary of the related works, highlighting their feature evaluation paradigms and primary focuses, is presented in Table 1.

Table 1. Comparative Analysis of Related Works in NIDS Feature Analysis

ID	Dataset(s)	Selection/Weighting Technique	Paradigm	Key Observation/Focus
R1	UNSW-NB15	Meta-heuristics (PSO, GWO, FFA, GA)	Global	Evaluated subsets of 13 to 30 features.
R2	UNSW-NB15	XGBoost	Global	Reduced to 19 features maintaining accuracy.
R3	CIC-IDS2017	Information Gain	Global	Focused on time/accuracy optimization via RF.
R4	CIC-IDS2017	RFE	Global	Tree-based models with best performance.
R5	CIC-IDS2017	Classifier Ensemble	Global	Achieved high binary accuracy with 30 features.
R6	CIC-IDS2017	Chi-square Weighting	Global	Proposed a novel weighted global FS method.
R7	CSE-CIC-IDS2018	Information Gain (One-vs-All)	Class-Specific	Identified distinct optimal feature subsets per class.
Ours	**NF-UNSWNB15-V2, CIC-IDS2017, NF-CICIDS2018-V2**	**Multi local FS, and Stability Analysis**	**Local and Cross-dataset**	**Sparsity, geometry, and local relevance analysis.**

References: R1: [Almomani 2020]; R2: [Kasongo and Sun 2020]; R3: [Kurniabudi et al. 2020]; R4: [Meemongkolkiat and Suttichaya 2021]; R5: [Mhawi et al. 2022]; R6: [Tripathi et al. 2024]; R7: [Göcs and Johanyák 2024].

3. Methodology

This study proposes a diagnostic framework for analyzing the structural behavior of network traffic attributes in multiclass NIDS. Rather than assuming feature importance is globally stable, the proposed methodology views feature selection in NIDS as a conditional semantic modeling problem and investigates how discriminative relevance varies across attack classes, datasets, and traffic regimes.

The framework combines class-wise relevance extraction, stability quantification, statistical distribution analysis, and hierarchical clustering to characterize feature specialization, semantic consistency, sparsity, multimodality, and transferability.

Under this perspective, feature relevance is treated as a conditional property shaped by attack semantics, protocol context, and dataset composition. Consequently, feature quality is evaluated not only through predictive importance, but also through semantic stability and cross-dataset robustness, providing a multidimensional alternative to conventional global feature-importance analysis.

3.1. Datasets and Preprocessing

To enable cross-dataset comparison, the standardized NetFlow-v2 datasets NF-UNSW-NB15-v2 and NF-CSE-CIC-IDS2018-V2 were used, along with the CIC-IDS2017 dataset.

For CIC-IDS2017, flow identification attributes (“id”, “Flow ID”, “Src IP”, “Dst IP”, and “Timestamp”) were removed following [Meemongkolkiat and Suttichaya 2021], since these fields introduce flow-specific bias, resulting in 79 remaining attributes. Port-based attributes were retained for cross-dataset analysis, as they are also present in the NetFlow-v2 datasets.

In the NetFlow-v2 datasets, the "MIN_TTL" and "MAX_TTL" features were excluded due to their unrealistically high predictive performance, observed in NF-UNSW-NB15-V2 both in our experiments and in [Sarhan et al. 2021a]. After preprocessing, 39 attributes remained.

Since this study focuses on class-specific feature analysis rather than predictive modeling, statistical analyses were performed on sampled subsets of the datasets. Uniform random sampling using a fixed random seed of 42 limited each class to a maximum of 5,000 instances. All available samples from smaller classes were retained, whereas the remaining classes were uniformly undersampled. This strategy balances computational efficiency with statistical reliability while preventing majority classes from dominating the statistical analysis.

3.2. Class-Wise Relevance Extraction

Let $X = \{x_1, x_2, \ldots, x_n\}$ denote the feature space and $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ the set of attack classes.

To reduce metric-specific bias, the analysis combined multiple feature-selection and attribution methods, including Mutual Information, Shannon Entropy, Random Forest Gini Importance, and SHAP values.

For each feature x_j, relevance scores were computed both globally, $w_j^{(global)}$, and locally, $w_j^{(c)}$, using a one-vs-all strategy for each class $c \in \mathcal{C}$. All scores were normalized to the interval $[0, 1]$.

To visualize semantic relationships between features and attack classes, hierarchical clustering was applied to the class-wise relevance matrices. The resulting heatmaps and dendrograms enable the identification of feature-specialization regions, attack-similarity groups, contextual dependency structures, and relevance-dilution effects caused by global aggregation.

3.3. Feature Stability and Cross-Dataset Consistency

To quantify semantic consistency across attack classes, we define the Feature Stability Index (FSI):

$$FSI(j) = 1 - \frac{1}{|\mathcal{C}|^2} \sum_{c_1, c_2 \in \mathcal{C}} \left| w_j^{(c_1)} - w_j^{(c_2)} \right| \tag{1}$$

High FSI values indicate generalized discriminative behavior, whereas low values indicate contextual specialization.

To provide a meaningful estimate of transferability across datasets that accounts for varying feature counts and relevance magnitudes, comparisons must be performed only between semantically aligned classes. Accordingly, we define the relative Cross-Dataset Stability (CDS) as:

$$CDS(j) = 1 - \frac{\sum_{c \in \mathcal{C}_{shared}} |w_{j,d_1}^{(c)} - w_{j,d_2}^{(c)}|}{\sum_{c \in \mathcal{C}_{shared}} (w_{j,d_1}^{(c)} + w_{j,d_2}^{(c)})} \tag{2}$$

where $w_{j,d}^{(c)}$ denotes the relevance score of the feature j for class c in the dataset d, and $\mathcal{C}_{shared} = \mathcal{C}_{d_1} \cap \mathcal{C}_{d_2}$ represents the set of semantically aligned attack classes shared between datasets d_1 and d_2.

Semantic alignment of attack classes and features across datasets was established by comparing the official dataset documentation, feature definitions, and attack descriptions, followed by manual curation by the authors. Although carefully curated, this mapping may still be affected by differences in attack implementations and dataset construction.

Higher CDS values indicate greater semantic consistency of feature relevance across corresponding attack classes in different datasets.

Together, FSI and CDS characterize whether a feature is globally robust or strongly dependent on attack context and dataset composition.

3.4. Sparsity and Multimodality Analysis

To characterize statistical heterogeneity, sparsity, and multimodality analyses were performed for each feature-class pair.

Feature sparsity was computed as:

$$S(x_j, c) = \frac{\#(x_j = 0)}{N_c} \tag{3}$$

where N_c is the number of samples in class c.

Multimodality was evaluated using peak detection and the Bimodality Coefficient (BC) derived from skewness and kurtosis. Values above 0.55 indicate non-unimodal behavior, indicating that the feature distribution departs from the expected single-peaked Gaussian or normal distribution and instead exhibits multiple behavioral regimes or separated density concentrations.

These analyses enable the identification of features that show contextual inactivity, fragmented behavioral regimes, and hidden class-specific structures masked by global aggregation. However, caution is required when interpreting sparse attributes, since zero-valued observations may also carry semantic significance depending on the attribute and the underlying network behavior being analyzed.

4. Results and Discussion

This section presents evidence of the divergence between global and class-specific feature relevance. Although feature rankings varied across datasets and FS methods, the statistical behavior remained consistent. Due to space limitations and the widespread adoption of entropy-based Information Gain in NIDS research, we present this metric as a representative case study.

4.1. Class-wise Feature Relevance Variability

To quantify class-wise discrepancy, we computed the standard deviation and min-max relevance range of each feature across attack classes. High variance indicates that a feature is highly discriminative for some attacks while nearly irrelevant for others.

Since relevance scores are normalized to sum to 1.0, the values must be interpreted within their respective feature spaces. The average feature weight is approximately 0.0127 in CIC-IDS2017 and 0.0256 in the NetFlow-v2 datasets. Consequently, a relevance shift of 0.10 represents a substantial discriminative variation, indicating that seemingly small numerical differences may correspond to transitions from negligible importance to dominant predictors depending on the attack semantics.

Table 2. Highest class-wise relevance discrepancies in NF-UNSW-NB15-V2

Feature	Std. Dev.	Range	Strongest Class	Score	Ignored By	Score
L4_DST_PORT	0.1090	0.3672	Shellcode	0.3764	Benign	0.0092
MIN_IP_PKT_LEN	0.0958	0.3094	Benign	0.3172	Analysis	0.0078
SERVER_TCP_FLAGS	0.0631	0.2104	Benign	0.2109	Shellcode	0.0005
PROTOCOL	0.0347	0.0944	Analysis	0.0961	Benign	0.0018
L4_SRC_PORT	0.0331	0.1019	Backdoor	0.1051	Benign	0.0032
TCP_WIN_SIZE	0.0318	0.0936	Exploits	0.0941	Worms	0.0005
CLIENT_TCP_FLAGS	0.0297	0.0885	Generic	0.0891	Shellcode	0.0006
IN_BYTES	0.0281	0.0817	DoS	0.0824	Analysis	0.0007
OUT_BYTES	0.0264	0.0749	Generic	0.0755	Backdoor	0.0006
FLOW_DURATION	0.0248	0.0712	Reconnaissance	0.0719	Shellcode	0.0007

Table 3. Highest class-wise relevance discrepancies in CIC-IDS2017

Feature	Std. Dev.	Range	Strongest Class	Score	Ignored By	Score
Packet Length Std	0.0507	0.1442	DoS/DDoS	0.1492	Brute Force	0.0051
Dst Port	0.0490	0.1285	Brute Force	0.1321	DoS/DDoS	0.0036
Bwd Packet Length Std	0.0443	0.1233	DoS/DDoS	0.1234	PortScan	0.0001
Flow IAT Mean	0.0426	0.1188	Infiltration	0.1199	Brute Force	0.0011
PSH Flag Count	0.0425	0.1186	Brute Force	0.1187	PortScan	0.0000
Packet Length Variance	0.0418	0.1169	DoS/DDoS	0.1175	Web Attack	0.0006
Fwd Packet Length Max	0.0399	0.1113	PortScan	0.1121	Bot	0.0008
Average Packet Size	0.0384	0.1091	DoS/DDoS	0.1097	Brute Force	0.0006
Flow Duration	0.0375	0.1036	Infiltration	0.1041	PortScan	0.0005
Subflow Fwd Bytes	0.0369	0.1018	DoS/DDoS	0.1024	Bot	0.0006

Table 4. Highest class-wise relevance discrepancies in NF-CSE-CIC-IDS2018-V2

Feature	Std. Dev.	Range	Strongest Class	Score	Ignored By	Score
L4_SRC_PORT	0.0676	0.1973	Brute Force-XSS	0.1974	Bot	0.0001
NUM_PKTS_128_TO_256_BYTES	0.0615	0.2253	SSH-Bruteforce	0.2253	SlowHTTPTest	0.0000
L4_DST_PORT	0.0610	0.2199	Bot	0.2200	LOIC-UDP	0.0001
NUM_PKTS_UP_TO_128_BYTES	0.0472	0.1961	LOIC-UDP	0.1996	Bot	0.0035
L7_PROTO	0.0467	0.1666	Bot	0.1669	LOIC-UDP	0.0003
TCP_FLAGS	0.0432	0.1511	DoS-Hulk	0.1518	FTP-Bruteforce	0.0007
IN_BYTES	0.0416	0.1427	Bot	0.1430	Slowloris	0.0003
OUT_BYTES	0.0399	0.1364	DDoS-HOIC	0.1371	Bot	0.0007
FLOW_DURATION	0.0387	0.1298	SlowHTTPTest	0.1305	Bot	0.0007
MAX_IP_PKT_LEN	0.0375	0.1246	SSH-Bruteforce	0.1252	Slowloris	0.0006

Tables 2, 3, and 4 summarize the features exhibiting the highest class-wise relevance instability in the datasets. Several important observations emerge from these analyses.

Table 2 shows that `MIN_IP_PKT_LEN` exhibits an extreme relevance discrepancy exceeding $40\times$ between Analysis and Benign traffic, reinforcing the contextual nature

of feature relevance. Similarly, in Table 3, packet-length statistics emerge as dominant discriminative attributes for DoS/DDoS traffic while becoming nearly irrelevant for Brute Force and Bot attacks. In Table 4, several features exhibit near-zero relevance for certain attacks while becoming dominant predictors for others. For example, `NUM_PKTS_128_TO_256_BYTES` strongly represents SSH-Bruteforce traffic but becomes negligible for DoS attacks-SlowHTTPTest. Overall, these results illustrate only a subset of the possible findings supported by the class-wise relevance analysis.

4.2. Visualization of Feature Specialization

Figure 1 summarizes the feature-specialization behavior previously identified in NF-UNSW-NB15-V2. Columns represent attack classes, while rows group features exhibiting similar discriminative behavior.

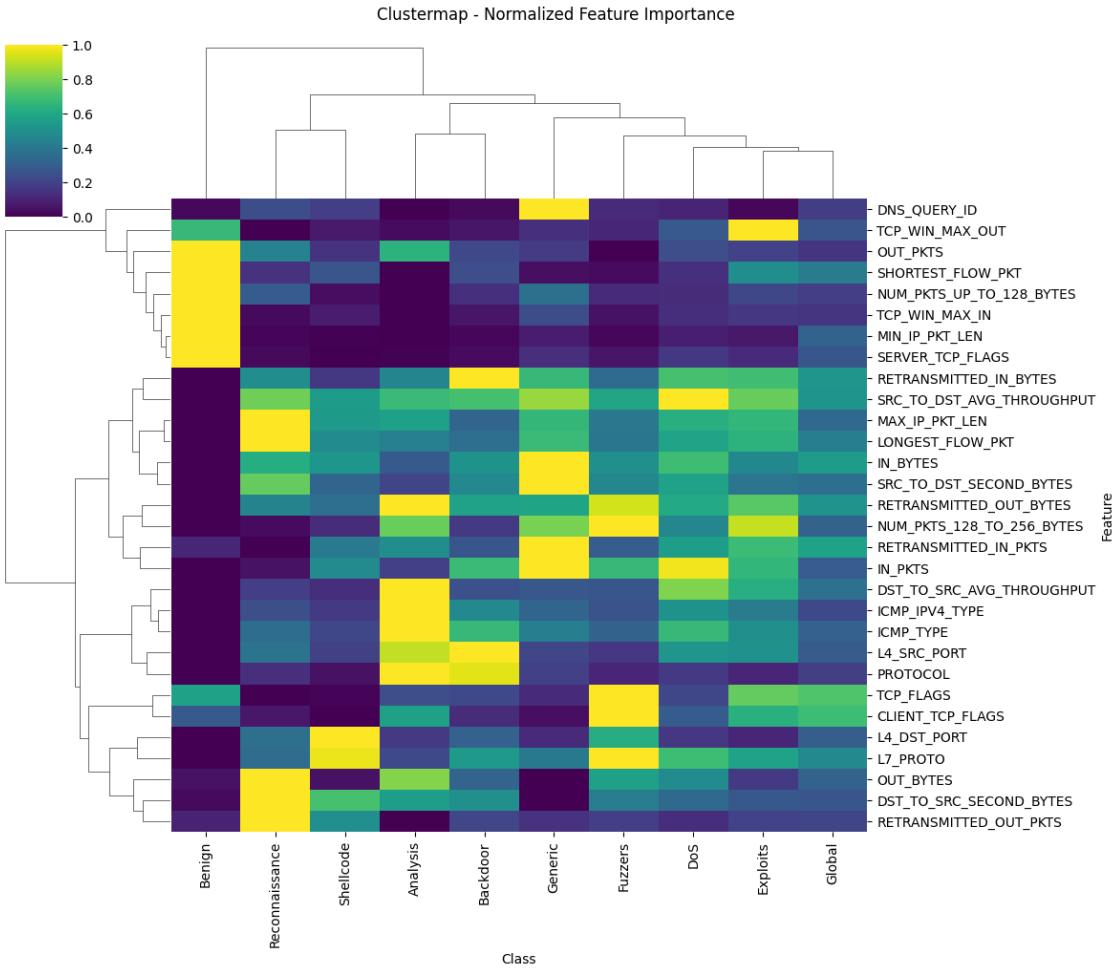

Figure 1. Clustermap of normalized feature relevance in NF-UNSW-NB15-V2. The rightmost column represents global relevance.

The heatmap provides strong visual evidence of attack-specific specialization. Features such as `L4_SRC_PORT`, `L4_DST_PORT`, and `PROTOCOL` exhibit strong activation for Analysis, Backdoor, and Shellcode attacks, while throughput-related attributes become dominant for DoS and Exploits traffic.

The Benign column reveals that only a small subset of characteristics accounts for the majority of importance scores. More importantly, the Global column highlights the statistical dilution effect introduced by global aggregation. Several localized discriminative regions become substantially attenuated after averaging.

Equivalent conclusions were observed across the remaining datasets, although the underlying relevance distributions and specialization patterns differ.

Beyond intra-dataset variability, substantial inter-dataset instability was also observed. For example, the destination-port attribute dominates several UNSW-NB15 classes but show weaker and more unstable relevance in CIC-IDS2017. Similar instability was observed for packet-length statistics, TCP flags, and flow-duration descriptors.

These results suggest that feature relevance is strongly influenced by dataset generation methodology, traffic composition, and attack implementation details.

4.3. Sparsity and Multimodality Analysis

Beyond relevance variability, network attributes exhibit distinct sparsity and multimodality patterns that challenge global selection assumptions.

A core finding of this analysis is that some feature relevance is right-linked to statistical heterogeneity. Attributes that exhibit extreme sparsity across the dataset often operate as specialist features. This relationship suggests that high sparsity is not merely a lack of information, but a discriminative state that characterizes protocol-specific and behavioral anomalies.

Table 5 illustrates the 10 non-port features in the NF-UNSW-NB15-V2 dataset exhibiting the highest statistical heterogeneity, combining peak density and sparsity metrics.

Table 5. Highest multimodality and peak density (NF-UNSW-NB15-V2)

Class	Feature	Peaks Found	Bimodality Coeff.	Sparsity (%)	Type
Shellcode	MAX_IP_PKT_LEN	95	0.4676	0.0	Multimodal
Reconnaissance	DNS_QUERY_ID	0	0.9231	99.64	Unimodal
Exploits	L7_PROTO	3	0.8532	30.9	Multimodal
Reconnaissance	IN_BYTES	1	0.8573	0.0	Multimodal
Generic	L7_PROTO	1	0.8443	64.9	Multimodal
Reconnaissance	PROTOCOL	2	0.8143	0.0	Multimodal
Fuzzers	PROTOCOL	2	0.8099	0.01	Multimodal
DoS	SHORTEST_FLOW_PKT	6	0.7938	0.0	Multimodal
Reconnaissance	LONGEST_FLOW_PKT	2	0.7290	0.0	Multimodal
Fuzzers	SHORTEST_FLOW_PKT	1	0.7128	0.0	Multimodal

The analysis shows that FTP- and DNS-related attributes behave mainly as binary protocol markers. For example, DNS_QUERY_ID is highly relevant for *Infiltration* (Rank 2) and *Generic* traffic (Rank 8), see also Figure 1, yet it exhibits nearly 100% sparsity across all NF-CSE-CIC-IDS2018 classes except *Benign* and *Infiltration*. This indicates that the model exploits the absence of DNS activity as a negative discriminative signal used to exclude unrelated classes. Moreover, because DNS_QUERY_ID is an identifier-oriented attribute, its relevance reflects dataset composition more than attack semantics, making it a strong candidate for preprocessing exclusion.

Similarly, ICMP_IPV4_TYPE remains inactive for most classes but becomes

highly discriminative for *Analysis* (Rank 3) and some *DoS* variants (Rank 10), acting as a specialist indicator for ICMP-oriented reconnaissance traffic.

Other structural attributes exhibit similar specialization patterns. `NUM_PKTS_128_TO_256_BYTES` is the Top-1 feature for *SSH-Bruteforce*, capturing the characteristic packet-size geometry of automated brute-force traffic. Likewise, `RETRANSMITTED_IN_PKTS` becomes the primary indicator for *DoS attacks-Slowloris* due to the retransmission behavior induced by intentional connection delays. In contrast, volumetric descriptors such as `OUT_BYTES` become highly discriminative for *DoS attacks-Hulk* and *Fuzzers*.

The coexistence of high sparsity and high class-specific relevance confirms the dilution effect introduced by global aggregation. Features critical for detecting protocol-specific or low-volume attacks are frequently suppressed by global feature-selection methods because they lack universal discriminative behavior.

4.4. Feature Stability and Cross-Dataset Consistency

This section quantifies feature semantic consistency using FSI to measure intra-dataset specialization and CDS to evaluate cross-environment transferability, defined respectively in Equations 1 and 2.

The FSI analysis reveals a dichotomy between specialist features and generalized structural attributes. A low FSI indicates a specialist feature dependent on specific attack mechanics, while a high FSI suggests a generalized discriminative attribute. Table 6 summarizes these profiles across the analyzed datasets.

Table 6. Feature Stability Index (FSI) and Specialization Profiles

Feature	Dataset	FSI Score	Behavioral Profile
`Bwd Packet Len Min`	CIC-IDS2017	0.12	**Ultra-Specialist**: Dominant for Benign traffic (Rank 1) but irrelevant for DoS (Rank 71).
`L4_DST_PORT`	UNSW-NB15	0.28	**Contextual**: Primary signal for Shellcode (0.376) but weak for Benign (0.009).
`NUM_PKTS_128_TO`	CSE-CIC18	0.31	**Attack-Specific**: Dominant for SSH-Bruteforce but negligible for others.
`MAX_IP_PKT_LEN`	CSE-CIC18	0.74	**Generalized**: Maintains Top-10 ranking across nearly all attack classes.
`SERVER_TCP_FLAGS`	UNSW-NB15	0.68	**Structural**: High stability; captures communication at the protocol level.

The CDS analysis exposes significant variability in feature relevance when transitioning between datasets. Table 7 summarizes the global relevance rankings and relative cross-dataset stability of representative features across the analyzed environments.

Global feature importance can be misleading because it aggregates heterogeneous attack behaviors and may favor dataset-specific artifacts instead of intrinsic attack seman-

tics. In this context, FSI and CDS help expose whether a feature preserves consistent discriminative behavior across classes and datasets, respectively. Consequently, features exhibiting both low FSI and low CDS tend to be highly contextual, unstable, dependent on dataset composition, and poorly transferable across environments.

A primary limitation of FSI and CDS lies in distinguishing intrinsic semantic specialization from environmental noise. Although low stability often indicates valuable class-specific discriminative behavior, it may also arise from dataset artifacts, traffic generation biases, or class imbalance. Consequently, a low stability score does not inherently guarantee predictive utility. These metrics must be complemented by qualitative protocol analysis to confirm that the observed specialization captures genuine attack mechanics rather than spurious dataset correlations.

Table 7. Relative Cross-Dataset Stability (CDS) and Relevance Rankings

Feature	CIC17 Rank	CIC18 Rank	UNSW Rank	CDS	Stability Interpretation
`MAX_IP_PKT_LEN`	4*	1	5	**0.94**	**Structural**: Stable payload boundary across datasets.
`L4_DST_PORT`	1	5	4	**0.91**	**Service-dependent**: Service-driven variability.
`OUT_BYTES`	11†	4	6	**0.88**	**Class-stable pattern**: Consistent volumetric stability.
`IN_BYTES`	13‡	10	14	**0.85**	**Class-stable pattern**: Inbound traffic intensity pattern.
`L7_PROTO`	—	3	16	**0.82**	**Collector-sensitive**: Affected by dataset composition.
`TCP_FLAGS`	18	12	23	**0.65**	**Contextual**: Dependent on session representation.
`MIN_IP_PKT_LEN`	54	15	25	**0.37**	**Unstable**: Sensitive to dataset construction.
`FLOW_DURATION`	21	27	36	**0.35**	**Artifact-prone**: Poor cross-dataset generalization.

Mapped from: *`Packet Length Max`, †`Total Length of Bwd Packet`, and ‡`Total Length of Fwd Packet`.

4.5. Critical Discussion on Port-Based Features

Port-related attributes consistently achieved high discriminative relevance. However, these features often encode service-level behavior rather than intrinsic attack semantics. Their importance primarily reflects associations between attacks and services operating on characteristic ports, causing models to learn application fingerprints instead of underlying attack patterns.

This behavior may reduce generalization when services migrate to different ports and may also inflate performance due to dataset-specific artifacts.

5. Conclusion

This work challenged the conventional assumption that globally optimized feature-selection strategies are sufficient for modern NIDS. Through comprehensive class-wise

and cross-dataset analyses, the results demonstrated that feature relevance is inherently contextual, varying significantly according to attack semantics, protocol behavior, and dataset construction methodology. The experiments revealed strong intra- and inter-dataset instability, showing that many attributes traditionally ranked as highly relevant under global selection exhibit low semantic consistency, high sparsity, and poor transferability.

The proposed class-specific analysis exposed a statistical dilution phenomenon in which global aggregation attenuates localized discriminative signatures, masking informative attack-specific patterns. In contrast, local relevance analysis isolates critical behavioral regions and preserves the discriminative geometry associated with specific attack classes. A central contribution of this study is the introduction of the Feature Stability Index (FSI) and Cross-Dataset Stability (CDS), which provide quantitative evidence that feature quality depends not only on predictive importance, but also on stability and semantic consistency. Notably, in open-set scenarios, features considered irrelevant under global metrics may become highly discriminative for novel attacks, since global representations frequently suppress subtle yet critical indicators of emerging threats.

The results also suggest that conventional FS methods may unintentionally prioritize dataset-specific artifacts over generalized behavioral descriptors. By identifying contextually meaningful attributes for each attack scenario, the proposed framework strengthens explainable AI studies in cybersecurity, enabling more interpretable decisions, clearer attack characterization, and improved understanding of the behavioral role of individual features.

Overall, this work establishes a new perspective for explainable and adaptive NIDS by demonstrating that preserving local discriminative structures is essential for achieving robustness, interpretability, and transferability under open, imbalanced, and continuously evolving traffic distributions. More broadly, the results suggest that conventional global feature-selection assumptions are insufficient for modern intrusion-detection environments, where discriminative relevance emerges from localized behavioral structures rather than global statistical dominance. Consequently, feature selection should be treated as a conditional semantic modeling problem, where contextual and class-specific analysis becomes necessary to preserve attack semantics and mitigate the cross-dataset generalization limitations observed in current approaches.

As future work, we plan to develop a functional NIDS architecture to conduct a comparative experiment between conventional global feature-selection strategies and the proposed class-specific approach. This will allow us to empirically demonstrate performance gains, particularly regarding minority attack recall. Additionally, we will explore integrating this class-aware strategy into a stream-based NIDS for Edge and IoT environments, ensuring continuous adaptation to evolving network behaviors and improving detection performance over time.

Acknowledgments

The authors used *ChatGPT* (OpenAI) solely for language refinement, including grammatical correction and improvements in clarity and readability. All content was critically reviewed and edited by the authors, who take full responsibility for the final manuscript.

Referências

Almomani, O. (2020). A feature selection model for network intrusion detection system based on PSO, GWO, FFA and GA algorithms. *Symmetry*, 12(6):1046.

Awad, M., Fraihat, S., Salameh, K., and Redhaei, A. A. (2022). Examining the suitability of NetFlow features in detecting IoT network intrusions. *Sensors*, 22(16):6164.

Bendale, A. and Boult, T. E. (2016). Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1563–1572.

Bouguelia, M.-R., Belaïd, Y., and Belaïd, A. (2016). An adaptive streaming active learning strategy based on instance weighting. *Pattern Recognition Letters*, 70:38–44.

Disha, R. A. and Waheed, S. (2022). Performance analysis of machine learning models for intrusion detection system using gini impurity-based weighted random forest (GIWRF) feature selection technique. *Cybersecurity*, 5(1).

Göcs, L. and Johanyák, Z. C. (2024). Identifying relevant features of CSE-CIC-IDS2018 dataset for the development of an intrusion detection system. *Intelligent Data Analysis*, 28(6):1527–1553.

Han, M., Li, C., Meng, F., He, F., and Zhang, R. (2024). An adaptive active learning method for multiclass imbalanced data streams with concept drift. *Applied Sciences*, 14(16):7176.

Kasongo, S. M. and Sun, Y. (2020). Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset. *Journal of Big Data*, 7(1).

Kurniabudi, Stiawan, D., Darmawijoyo, Idris, M. Y. B., Bamhdi, A. M., and Budiarto, R. (2020). CICIDS-2017 dataset feature analysis with information gain for anomaly detection. *IEEE Access*, 8:132911–132921.

Li, C., Zhang, E., Geng, C., and Chen, S. (2024). All beings are equal in open set recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12):13446–13454.

Liu, W., Zhang, H., Ding, Z., Liu, Q., and Zhu, C. (2021). A comprehensive active learning method for multiclass imbalanced data streams with concept drift. *Knowledge-Based Systems*, 215:106778.

Liu, W., Zhu, C., Ding, Z., Zhang, H., and Liu, Q. (2023). Multiclass imbalanced and concept drift network traffic classification framework based on online active learning. *Engineering Applications of Artificial Intelligence*, 117:105607.

Meemongkolkiat, N. and Suttichaya, V. (2021). Analysis on network traffic features for designing machine learning based IDS. *Journal of Physics: Conference Series*, 1993(1):012029.

Mhawi, D. N., Aldallal, A., and Hassan, S. (2022). Advanced feature-selection-based hybrid ensemble learning algorithms for network intrusion detection systems. *Symmetry*, 14(7):1461.

Moualla, S., Khorzom, K., and Jafar, A. (2021). Improving the performance of machine learning-based network intrusion detection systems on the UNSW-NB15 dataset. *Computational Intelligence and Neuroscience*, 2021:1–13.

Ning, K.-P., Ke, H.-J., Yao, J.-Y., Liu, Y.-Y., Tian, Y.-H., and Yuan, L. (2026). Evidence conflict sampling for open-set active learning. *International Journal of Computer Vision*, 134(1).

Niño-Adan, I., Manjarres, D., Landa-Torres, I., and Portillo, E. (2021). Feature weighting methods: A review. *Expert Systems with Applications*, 184:115424.

Sarhan, M., Layeghy, S., Moustafa, N., and Portmann, M. (2021a). NetFlow datasets for machine learning-based network intrusion detection systems. In *Big Data Technologies and Applications*, pages 117–135. Springer International Publishing.

Sarhan, M., Layeghy, S., and Portmann, M. (2021b). Towards a standard feature set for network intrusion detection system datasets. *Mobile Networks and Applications*, 27(1):357–370.

Sarhan, M., Layeghy, S., and Portmann, M. (2022). Evaluating standard feature sets towards increased generalisability and explainability of ML-based network intrusion detection. *Big Data Research*, 30:100359.

Scheirer, W. J., de Rezende Rocha, A., Sapkota, A., and Boult, T. E. (2013). Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772.

Schmidt, S., Schenk, L., Schwinn, L., and Günnemann, S. (2025). Joint out-of-distribution filtering and data discovery active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25677–25687.

Shu, J., Yuan, X., Meng, D., and Xu, Z. (2022). CMW-Net: Learning a class-aware sample weighting mapping for robust deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:11521–11539.

Tripathi, G., Singh, V. K., Sharma, V., and Vinodbhai, M. V. (2024). Weighted feature selection for machine learning based accurate intrusion detection in communication networks. *IEEE Access*, 12:20973–20982.

Yin, T., Liu, N., Sun, H., and Xia, S. (2025). Boosting active learning via re-aligned feature space. *Knowledge-Based Systems*, 311:113085.

Zhang, Y., Zhang, H., and Zhang, B. (2022). An effective ensemble automatic feature selection method for network intrusion detection. *Information*, 13(7):314.

Zhao, R., Mu, Y., Zou, L., and Wen, X. (2022). A hybrid intrusion detection system based on feature selection and weighted stacking classifier. *IEEE Access*, 10:71414–71426.