Introducing Perturb-ability Score (PS) to Enhance Robustness Against Evasion Adversarial Attacks on ML-NIDS

Ashraf Matrawy
Carleton University
carleton.ca/ngn

A presentation at the
Interdisciplinary Research Center for Intelligent Secure Systems
KFUPM, Dhahran, KSA
November 24, 2024

Acknowledgement

- I would like to acknowledge the contributions of my students to the research presented in these slides. The main contributor in these slides is my PhD student Mohamed Elshehaby.
- Some of the work in this presentation was funded by Natural Sciences and Engineering Research Council of Canada (NSERC).

- We work on ML in network security, security in IoT, 5G and beyond, misinformation, and usable security. Please visit our group page for more information. The Next Generation Networks Group `carleton.ca/ngn`
- Most of the work and figures are taken our draft posted at `https://arxiv.org/abs/2409.07448`.

Carleton University

Outline

- Introduction
- Feature-Space vs Problem-Space Evasion Adversarial Attacks Against ML-NIDS
- Perturb-ability of Features in Problem-Space Against NIDS
- Motivation and Aim
- Evaluating PS
- Enabling a Defense with PS
- Results
- Discussion
- Conclusion

Gap between reality and research: The practicality question?

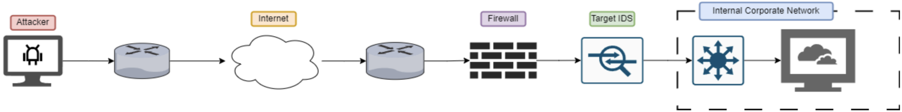


Figure: The Deployment of Network Intrusion Detection System - from our paper in IEEE WF-IoT [1].

Adversarial attacks - Evasion

Figure: Evasion Adversarial Attack [2]

Feature-Space vs Problem-Space Evasion Adversarial Attacks Against ML-NIDS

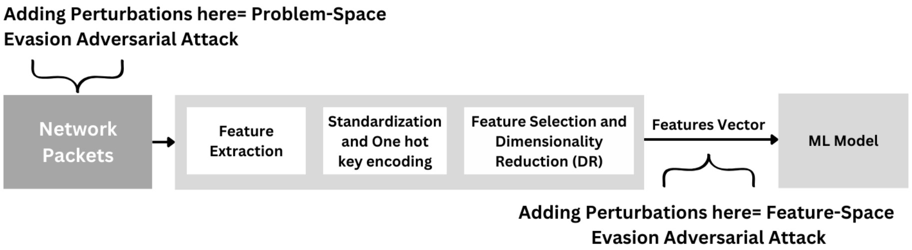


Figure: Evasion Adversarial Attacks in Feature-Space vs Problem-Space Against NIDS

Feature-Space vs Problem-Space Evasion Adversarial Attacks Against ML-NIDS (Cont.)

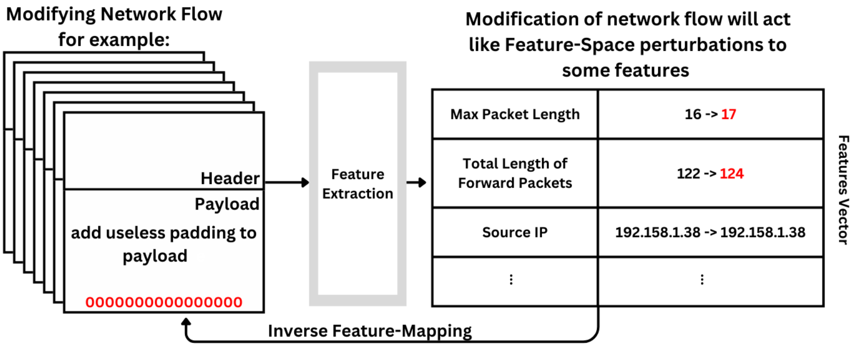
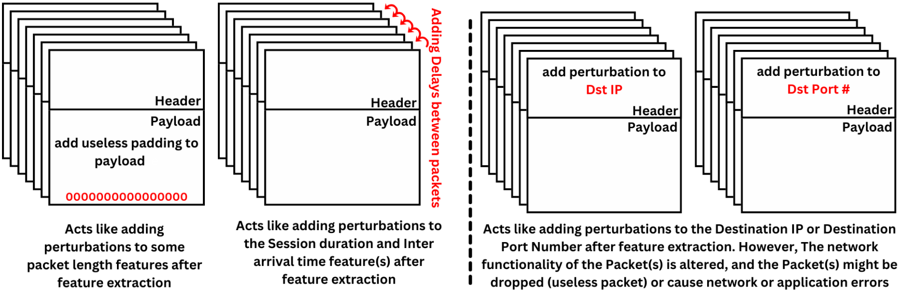


Figure: Example of Evasion Adversarial Attacks Problem-Space Perturbations Against NIDS

Carleton University

Perturb-ability of Features in Problem-Space Against NIDS



(a) Two Examples of Perturb-able Features in Problem-Space Against NIDS

(b) Two Examples of Non-Perturb-able Features in Problem-Space Against NIDS

Figure: Examples of Perturb-able vs Non-Perturb-able Features in Network Traffic

Motivation and Aim

- Intuitive assumption that attackers can only access the problem-space rather than the feature-space. This perspective aligns with the reality of most network environments, where attackers can manipulate packet contents but do not have direct control over the feature extraction process for more details on our threat model.
- In response to this, our aim is to introduce the novel notion of the **Perturb-ability Score (PS)** metric, which is designed to enhance the robustness of ML-based NIDS.
- **The PS metric helps to identify features in the problem-space that are susceptible to manipulation by attackers, without compromising the malicious functionality of network traffic.**

Motivation and Aim (Cont.)

By quantifying the perturb-ability of each feature within NIDS domain constraints, PS facilitates the selection of features that are inherently more resistant to adversarial attacks. Our aimed classification is shown in Fig. 24.

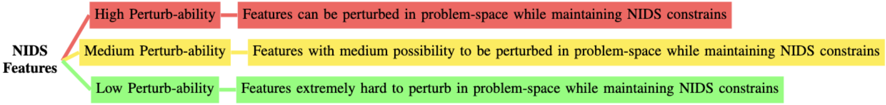


Figure: Classification of NIDS Features based on our proposed PS, where green represents a feature in the Low Perturb-ability class, yellow represents a feature in the Medium Perturb-ability class, and red represents a feature in the High Perturb-ability class

Carleton University

Evaluating PS

PS aims to measure how easily each feature in a dataset can be perturbed while maintaining NIDS problem-space constraints. The score ranges from 0, indicating features that are very difficult to perturb, to 1, indicating features that are easy to perturb. The total PS for each feature is the geometric mean of five different criteria, each of which has a value from 0 to 1.

- **PS_1: Strict Header/Network or Malicious Functionality**
- **PS_2: Feature Value Range**
- **PS_3: Correlated Features**
- **PS_4: Unaccessible Features**
- **PS_5: Features Correlated with numerous flow Packets**

For more info: `https://arxiv.org/abs/2409.07448`

Carleton University

Evaluating PS (Cont.)

PS_1: Strict Header/Network or Malicious Functionality

This PS field focuses on strict Header features and network/malicious functionality of network flows after adding perturbations in the problem-space. $PS_1[f_i]$ will be 0 if any of the following conditions are true (which will make $PS_{Total}[f_i]$ equals 0);

C1: the feature f_i is a strict header feature (IP addresses in TCP flows, destination port number or protocol)

C2: adding perturbation to feature f_i will affect the network or malicious functionality of the flow.

$PS_1[f_i]$ can be described with the following equation:

$$PS_1[f_i] = \begin{cases} 0, & \text{if (C1 or C2)} \\ 1, & \text{otherwise} \end{cases}$$

Evaluating PS (Cont.)

PS_2: Feature Value Range $PS_2[f_i]$ will be 1 if f_i's number of Possible Values (PV) is greater than 255 (this feature will be similar to computer vision's pixel, and it will be flexible to perturb). On the other hand, if f_i's PV ($PV[f_i]$) is less than or equal to 255, $PS_2[f_i]$ will be equal to a linear function where its output is 1 if f_i's PV is 255, and 0.5 if f_i's PV is 2 (binary). If $PV[f_i]$ is less than 2 (equals 1), it indicates that f_i is non-perturb-able, in which case $PS_2[f_i]$ will be set to 0. However, in this case (where $PV[f_i]$ equals 1), we recommend dropping that feature, as it does not contribute meaningful information to the ML model.
$PS_2[f_i]$ can be described with the following equation:

$$PS_2[f_i] = \begin{cases} 1 & \text{if } \mathrm{PV}[f_i] > 255 \\ 0 & \text{if } \mathrm{PV}[f_i] < 2 \\ 0.5 + (0.5 \times \frac{(\mathrm{PV}[f_i]-2)}{(255-2)}) & \text{otherwise} \end{cases}$$

Evaluating PS (Cont.)

PS_3: Correlated Features This PS field considers the correlation between a NIDS feature and other features. Due to network constraints within NIDS, many features exhibit problem-space correlations. For instance, the flow duration feature is typically correlated with the total forward and backward inter-arrival times. Such correlated features limit the attacker's flexibility. The gradients of the targeted model might recommend a specific perturbation to one feature and a different perturbation to another. As the number of correlated features associated with a single feature increases, it becomes more difficult to perturb that feature in the problem-space. $PS_3[f_i]$ will follow a linear function, where its output is 0.5 if the number of Correlated Features (CF) of f_i is equal to or greater than a threshold (the maximum number observed in our experiments was 10, which we chose as the threshold), and 1 if f_i's CF ($CF[f_i]$) is 0.
$PS_3[f_i]$ can be described with the following equation:

$$PS_3[f_i] = 1 - 0.05 \times \min(CF[f_i], 10)$$

Carleton University

Evaluating PS (Cont.)

PS_4: Unaccessible Features This PS field focuses on features that attackers cannot access. Examples of such features include backward features (e.g., Minimum Backward Packet Length) and interflow features (e.g., number of flows that have a command in an FTP session (ct_ftp_cmd)).

$PS_4[f_i]$'s value will depend on the following conditions;

C3: the feature f_i is not a backward or interflow feature. In other words, attackers can access f_i. **C4:** the feature f_i is a backward or interflow feature; however, it is highly correlated with a forward feature. In other words, attackers can modify f_i in an indirect way. **C5:** the feature f_i is a backward or interflow feature; however, it is correlated with multiple forward features. In other words, attackers can modify f_i indirectly, but it will be challenging for them as it is correlated with multiple features.

Otherwise (if none of C3, C4, or C5 apply): the feature f_i is a backward or interflow feature and it is not correlated with any forward feature. In other words, attackers cannot access f_i.

$$PS_4[f_i] = \begin{cases} 1, & \text{if (C3 or C4)} \\ 0.5, & \text{if (C5)} \\ 0, & \text{otherwise} \end{cases}$$

Carleton University

Evaluating PS (Cont.)

PS_5: Features Correlated with numerous flow Packets This PS field considers features that are correlated with numerous flow packets.
$PS_5[f_i]$'s value will depend on the following condition;
C6: f_i is a feature that requires modifying the entire flow of packets (forward, backward, or both), such as mean or standard deviation features.

$$PS_5[f_i] = \begin{cases} 0.5, & \text{if (C6)} \\ 1, & \text{otherwise} \end{cases}$$

Carleton University

Evaluating PS (Cont.)

$\mathrm{PS_{Total}}[f_i]$ The overall Perturb-ability Score ($\mathrm{PS_{Total}}[f_i]$) for each feature f_i is calculated as the geometric mean of the five individual PS fields we defined. $\mathrm{PS_{Total}}[f_i]$ can be described with the following equation:

$$\mathrm{PS_{Total}}[f_i] = \sqrt[5]{\prod_{j=1}^{5} \mathrm{PS_j}[f_i]}$$

The $\mathrm{PS_{Total}}$ will be calculated for all features f_i in the dataset, from $i = 1$ to n, where n is the number of features in the dataset.

Calibration of values and thresholds is important.

Enabling a Defense with PS

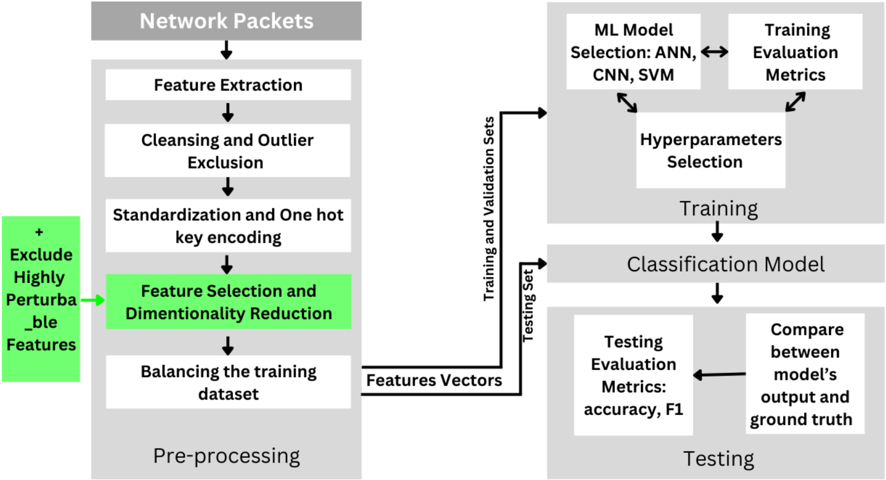


Figure: Using PS as a Potential Defense against Practical Problem-Space Evasion Adversarial Attacks

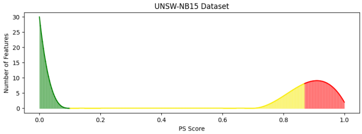
Results

Table: The number and percentage of features in every perturb-ability class, based on our proposed PS, where green indicates low perturb-ability features class, yellow indicates medium perturb-ability features class, and red indicates high perturb-ability features class

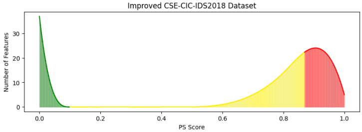
Pert. Class / Dataset	# and % of Low Pert. Features	# and % of Med. Pert. Features	# and % of High Pert. Features	Total
UNSW-NB15 [3]	26 (55.3%)	4 (8.5%)	17 (36.1%)	47
CSE-CIC-IDS2018* [4]	38 (43.2%)	19 (21.6%)	31 (35.2%)	88

* Improved CSE-CIC-IDS2018 Dataset by Liu et al. [4]

Results (Cont.)



(a) UNSW-NB15 Dataset



(b) Improved CSE-CIC-IDS2018 Dataset

Figure: The histogram of PS values for each dataset where green indicates low perturb-ability features class, yellow indicates medium perturb-ability features class, and red indicates high perturb-ability features class

Results (Cont.)

srcip	sport	dstip	dsport	proto	state	dur
sbytes	dbytes	sttl	dttl	sloss	dloss	service
Sload	Dload	Spkts	Dpkts	swin	dwin	stcpb
dtcpb	smeansz	dmeansz	trans_depth	res_bdy_len	Sjit	Djit
Stime	Ltime	Sintpkt	Dintpkt	tcprtt	synack	ackdat
is_sm_ips_ports	ct_state_ttl	ct_flw_http_mthd	is_ftp_login	ct_ftp_cmd	ct_srv_src	ct_srv_dst
ct_dst_ltm	ct_src_ ltm	ct_src_dport_ltm	ct_dst_sport_ltm	ct_dst_src_ltm		

Figure: UNSW-NB15 Dataset's features classified based on our proposed PS, where green indicates a feature with low perturb-ability, yellow indicates a feature with medium perturb-ability, and red indicates a feature with high perturb-ability.

Results (Cont.)

Flow ID	Src IP	Src Port	Dst IP	Dst Port
Protocol	Timestamp	Flow Duration	Total Fwd Packet	Total Bwd packets
Total Len of Fwd Pack	Total Len of Bwd Pack	Fwd Packet Length Max	Fwd Packet Length Min	Fwd Packet Length Mean
Fwd Packet Length Std	Bwd Packet Length Max	Bwd Packet Length Min	Bwd Packet Length Mean	Bwd Packet Length Std
Flow Bytes/s	Flow Packets/s	Flow IAT Mean	Flow IAT Std	Flow IAT Max
Flow IAT Min	Fwd IAT Total	Fwd IAT Mean	Fwd IAT Std	Fwd IAT Max
Fwd IAT Min	Bwd IAT Total	Bwd IAT Mean	Bwd IAT Std	Bwd IAT Max
Bwd IAT Min	Fwd PSH Flags	Bwd PSH Flags	Fwd URG Flags	Bwd URG Flags
Fwd RST Flags	Bwd RST Flags	Fwd Header Length	Bwd Header Length	Fwd Packets/s
Bwd Packets/s	Packet Length Min	Packet Length Max	Packet Length Mean	Packet Length Std
Packet Len Variance	FIN Flag Count	SYN Flag Count	RST Flag Count	PSH Flag Count
ACK Flag Count	URG Flag Count	CWR Flag Count	ECE Flag Count	Down/Up Ratio
Average Packet Size	Fwd Segment Size Avg	Bwd Segment Size Avg	Fwd Bytes/Bulk Avg	Fwd Packet/Bulk Avg
Fwd Bulk Rate Avg	Bwd Bytes/Bulk Avg	Bwd Packet/Bulk Avg	Bwd Bulk Rate Avg	Subflow Fwd Packets
Subflow Fwd Bytes	Subflow Bwd Packets	Subflow Bwd Bytes	FWD Init Win Bytes	Bwd Init Win Bytes
Fwd Act Data Pkts	Fwd Seg Size Min	Active Mean	Active Std	Active Max
Active Min	Idle Mean	Idle Std	Idle Max	Idle Min
ICMP Code	ICMP Type	Total TCP Flow Time		

Figure: Improved CSE-CIC-IDS2018 Dataset's features classified based on our proposed PS, where green indicates a feature with low perturb-ability, yellow indicates a feature with medium perturb-ability, and red indicates a feature with high perturb-ability.

Carleton University

Results (Cont.)

Table: The performance of an ANN/Random Forest (RF)/SVM/CNN-based NIDS

	Dataset →	UNSW-NB15				Improved CSE-CIC-IDS2018			
	Model ↓	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
	ANN	0.9879	0.9129	0.9998	0.9544	1.0000	0.9998	0.9998	0.9998
	SVM	0.9879	0.9129	0.9996	0.9543	0.9999	0.9984	0.9994	0.9989
	RF	0.9891	0.9216	0.9986	0.9585	1.0000	0.9997	1.0000	0.9998
	CNN	0.9879	0.9126	0.9999	0.9543	1.0000	0.9993	0.9999	0.9996
	ANN	0.9879	0.9127	1.0000	0.9543	0.9998	0.9965	1.0000	0.9983
	SVM	0.9879	0.9129	0.9997	0.9543	0.9999	0.9982	0.9998	0.9990
	RF	0.9892	0.9220	0.9987	0.9588	1.0000	0.9998	1.0000	0.9999
	CNN	0.9879	0.9128	1.0000	0.9544	1.0000	0.9996	1.0000	0.9998
	ANN	0.9880	0.9130	0.9999	0.9545	1.0000	0.9996	0.9998	0.9997
	SVM	0.9879	0.9129	0.9997	0.9543	0.9999	0.9983	1.0000	0.9991
	RF	0.9897	0.9251	0.9993	0.9607	1.0000	0.9998	1.0000	0.9999
	CNN	0.9882	0.9145	0.9997	0.9552	1.0000	0.9994	1.0000	0.9997

Results (Cont.)

TABLE 5. MAPPING PROBLEM-SPACE EVASION ADVERSARIAL ATTACKS' TRAFFIC MORPHING TECHNIQUES TO FEATURES, THE FEATURES ARE COLORED BASED ON OUR PS CLASSIFICATION.*

Problem-space Attack and its Problem-space Morphing Techniques	Potentially Perturb-ed Features in Feature-space in UNSW-NB15	Potentially Perturb-ed Features in Feature-space in improved CSE-CIC-IDS2018
Han et al. [11] modify the interarrival times of packets in the original traffic, change values to the Time to Live (TTL) field, request to establish connections that are already established (or in the process of being established), and add padding to payloads. [11]	sttl, dur, Sjit, Sintpkt, Sload, Stime, Ltime, tcprtt, synack, ackdat . sbytes, smeansz, Sload, dbytes, Dload . Spkts, Dpkts.	Flow Duration, Timestamp, Flow Bytes/s, Flow Packets/s, Fwd IAT Total, Fwd IAT Mean, Fwd IAT Std, Fwd IAT Max, Fwd IAT Min, Fwd Packets/s, .Total Length of Fwd Packet, Fwd Packet Length Max, Min, Fwd Packet Length Mean, Fwd Packet Length Std, Fwd Bulk Rate Avg, Fwd Bytes/Bulk Avg, Fwd Segment Size Avg, Subflow Fwd Bytes , Fwd Act Data Pkts. Total Fwd Packets, Subflow Fwd Packets, Total Bwd Packets, Subflow Bwd Packets . Fwd PSH Flag, Bwd PSH Flags, Fwd URG Flags, Fwd RST Flags, FIN Flag Count, SYN Flag Count, RST Flag Count, PSH Flag Count, ACK Flag Count, URG Flag Coun, CWR Flag Count, ECE Flag Count
Hashemi et al. [12] split the original packet payload into multiple packets, modify the timing between packets by either increasing or decreasing the intervals, and inject dummy packets with random lengths, transmission times, and flag settings. [12]	dur, Sjit, Sintpkt, Sload, Stime, Ltime, tcprtt, synack, ackdat . sbytes, smeansz, Sload, dbytes, Dload Spkts, Dpkts.	Flow Duration, Timestamp, Flow Bytes/s, Flow Packets/s, Fwd IAT Total, Fwd IAT Mean,Fwd IAT Std, Fwd IAT Max, Fwd IAT Min, Fwd Packets/s, .Total Length of Fwd Packet, Fwd Packet Length Max, Fwd Packet Length Min, Fwd Packet Length Mean, Fwd Packet Length Std, Fwd Bulk Rate Avg, Fwd Bytes/Bulk Avg, Fwd Segment Size Avg, Subflow Fwd Bytes , Fwd Act Data Pkts. Total Fwd Packets, Subflow Fwd Packets, Total Bwd Packets, Subflow Bwd Packets . Fwd PSH Flag, Bwd PSH Flags, Fwd URG Flags, Fwd RST Flags, FIN Flag Count, SYN Flag Count, RST Flag Count, PSH Flag Count, ACK Flag Count, URG Flag Coun, CWR Flag Count, ECE Flag Count
Vitorino et al. [13] [14] [15] modify various flow attributes such as flow duration, average interarrival time between packets, packet rate (packets per second), average forward packet length, smallest forward segment size, minimum interarrival time between packets, and maximum interarrival time. [13] [14] [15]	dur, Sjit, Sload, sbytes, Spkts, Sintpkt smeansz	Flow Duration, Fwd IAT Total, Fwd IAT Mean, Fwd IAT Std, Fwd IAT Max,Fwd Packet Length Mean, Fwd IAT Min, Fwd IAT Max, Flow Bytes/s, Flow Packets/s
Yan et al. [16] modify length-related features by padding packets with irrelevant characters, increase the packet count by duplicating the request multiple	dur, Sjit, Sintpkt, Sload, Stime, Ltime, tcprtt, synack, ackdat . sbytes, smeansz, Sload, dbytes, Dload Spkts, Dpkts.	Flow Duration, Timestamp, Flow Bytes/s, Flow Packets/s, Fwd IAT Total, Fwd IAT Mean, Fwd IAT Std, Fwd IAT Max, Fwd IAT Min, Fwd Packets/s, .Total Length of Fwd Packet,

Discussion: The Usual Suspects

- From our research, we identify recurring traffic morphing techniques such as; Forward IAT, Forward Packet Length, and Forward Payload Size features.
- We refer to these features as "usual suspects"
- We recommend that NIDS researchers focus on these features, as modifying them does not compromise network functionality or the malicious intent of adversarial flows.

Discussion: Five Features

- Sheatsley et al. [5] demonstrated that adversarial attacks could succeed by modifying just five random features, achieving a 50% success rate.
- We highlight that our PS-enabled defense does not aim to reduce the number of features but focuses on selecting features that are non-perturb-able in the problem-space.
- We argue that the real-world complexity of problem-space constraints makes adversarial attacks less feasible compared to feature-space experiments made in Sheatsley et al.'s research.

Discussion: Problem-space Evasion Adversarial Attacks are Already Extremely Hard for an Attacker

- There are numerous challenges for attackers, including limited access to feature vectors, correlations between NIDS features, and the difficulty of translating feature-space manipulations into problem-space modifications.
- Attackers often rely on trial-and-error techniques, which lack theoretical guidance, and struggle to maintain the malicious functionality of the flow after problem-space modifications.
- Thus, introducing a simple addition in the feature-selecting phase in the architecture of ML-NIDS through the usage of the PS scoring mechanism to eliminate easily perturb-able features could be the last nail in the coffin for these already highly impractical and complex problem-space evasion adversarial attacks against NIDS.

Carleton University

Discussion: NIDS Datasets

- We acknowledge the limitations of current NIDS datasets, including poor diversity, feature dependence, and unclear ground truth.
- We addressed these issues by using an improved version of the CSE-CIC-IDS2018 dataset and employing thorough data pre-processing techniques.
- Our focus was on comparing models with access to all features versus models limited to non-perturb-able features, rather than directly evaluating the datasets.

Carleton University

Discussion: We are losing information! Are we?

- Some might argue that dropping features using PS could lead to information loss, emphasizing the importance of domain expertise in its application.
- Our results suggest that current NIDS literature may rely on more features than necessary, as we achieved promising results with fewer, non-perturb-able features.
- We also question the reliance on features that attackers can easily perturb.

Discussion: Adversarial Attacks on ML-NIDS Research Direction

- Many studies overestimate attacker capabilities by assuming access to information rarely available in real-world scenarios.
- Problem-space adversarial attacks are more practical than feature-space attacks but remain constrained by **collateral damage**.
- We stress the need to address significant issues in NIDS datasets and the unrealistic assumptions in current research.

References I

[1] J. Kadri, A. Lott, M. Brendan, K. Morozov, S. Virr, M. Elshehaby, and A. Matrawy, "Work in progress: Evasion adversarial attacks perturbations in network security," in *2024 IEEE 10th World Forum on Internet of Thing*. IEEE, 2024.

[2] O. Ibitoye, R. Abou-Khamis, A. Matrawy, and M. O. Shafiq, "The threat of adversarial attacks on machine learning in network security–a survey," *arXiv preprint arXiv:1911.02621*, 2019.

[3] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 military communications and information systems conference (MilCIS)*. IEEE, 2015, pp. 1–6.

Carleton University

References II

[4] L. Liu, G. Engelen, T. Lynar, D. Essam, and W. Joosen, "Error prevalence in nids datasets: A case study on cic-ids-2017 and cse-cic-ids-2018," in *2022 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2022, pp. 254–262.

[5] R. Sheatsley, N. Papernot, M. J. Weisman, G. Verma, and P. McDaniel, "Adversarial examples for network intrusion detection systems," *Journal of Computer Security*, vol. 30, no. 5, pp. 727–752, 2022.

Carleton University

Questions?

carleton.ca/ngn