The 23rd IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom 2024)

MARS: Robustness Certification for Deep Network Intrusion Detectors via Multi-Order Adaptive Randomized Smoothing

Mengdie Huang[1,2], **Yingjun Lin**[2], Xiaofeng Chen[1]✉, Elisa Bertino[2]

[1] Xidian University

[2] Purdue University

西安電子科技大學
XIDIAN UNIVERSITY

Overview

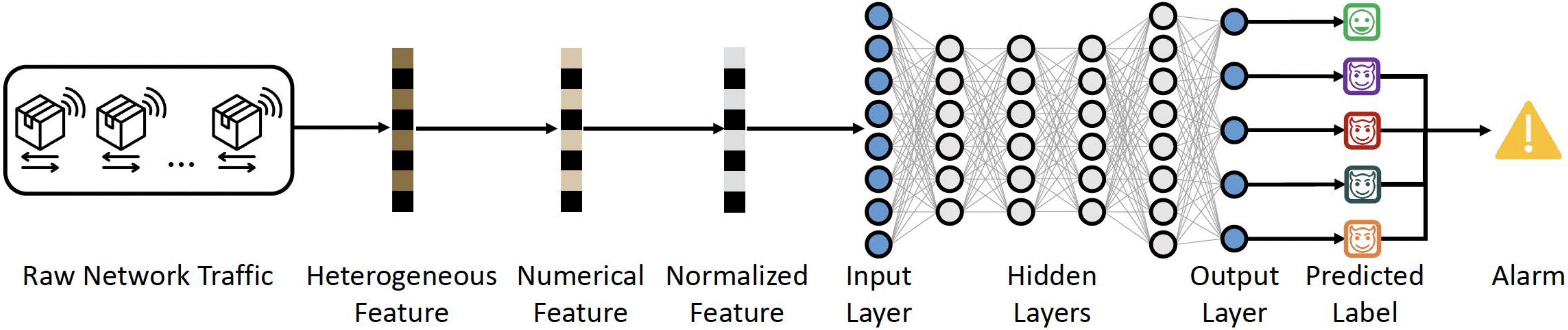
Contents

Keywords

- Deep Neural Network
- Network Intrusion Detection
- Natural Corruption
- Evasion Attack
- Certified Robustness
- Empirical Robustness

Deep Neural Network-based Network Traffic Classifier

- Workflow of the DNN-based Network Intrusion Detector (NID)



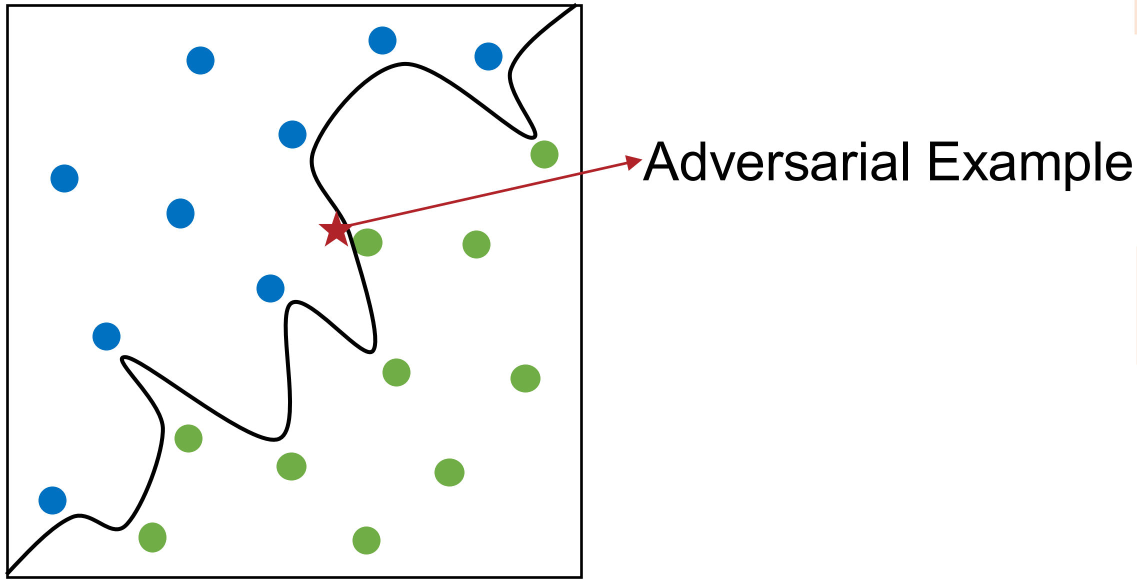
➢ Traffic Data includes both Numeric and Non-numeric Values (e.g. protocol, network service, timestamp, etc.)
  - First, transform the raw network traffic vector x_{raw} into a numerical feature vector x_{num}.
  - Then, normalize it into a feature vector x in a continuous real number range.

Threats to Deep Neural Networks (DNNs)

- Standard Train a Base Classifier F
  - Optimization objective of standard training

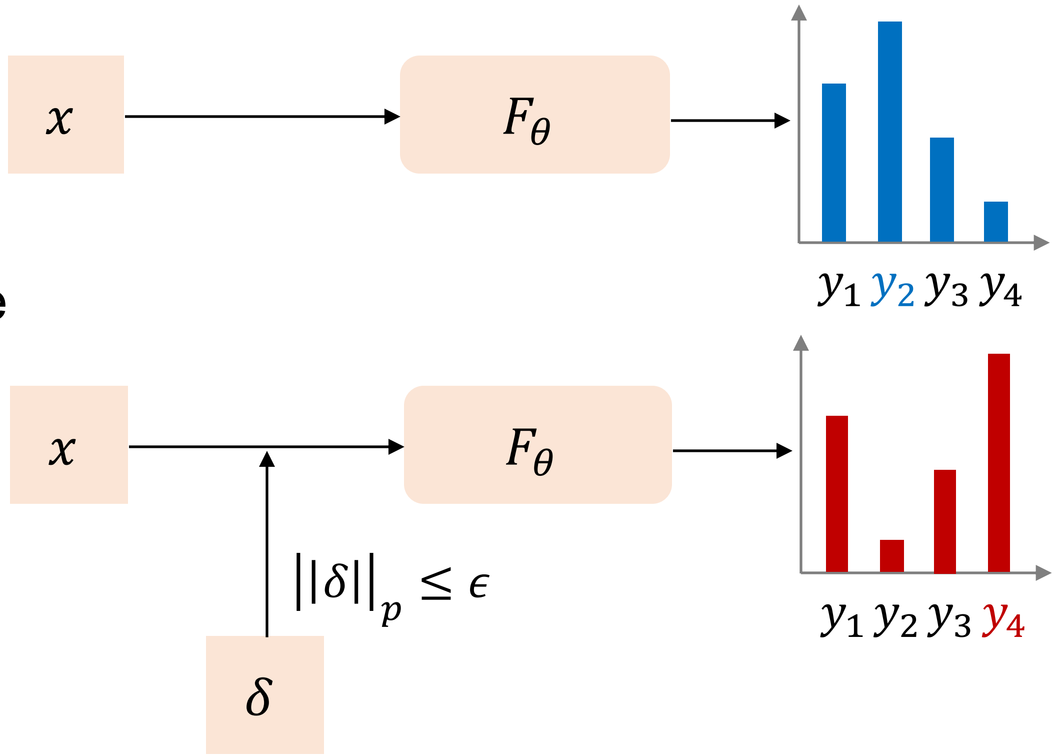
$$\min_{\theta} \mathbb{E}_{(x,y_{true})\sim\mathcal{D}_{train}}[\mathcal{L}(F_\theta(x), y_{true})]$$

Standard Training



- Evasion Attack with Adversarial Example $(x+\delta)$
  - Optimization objective of untargeted attack

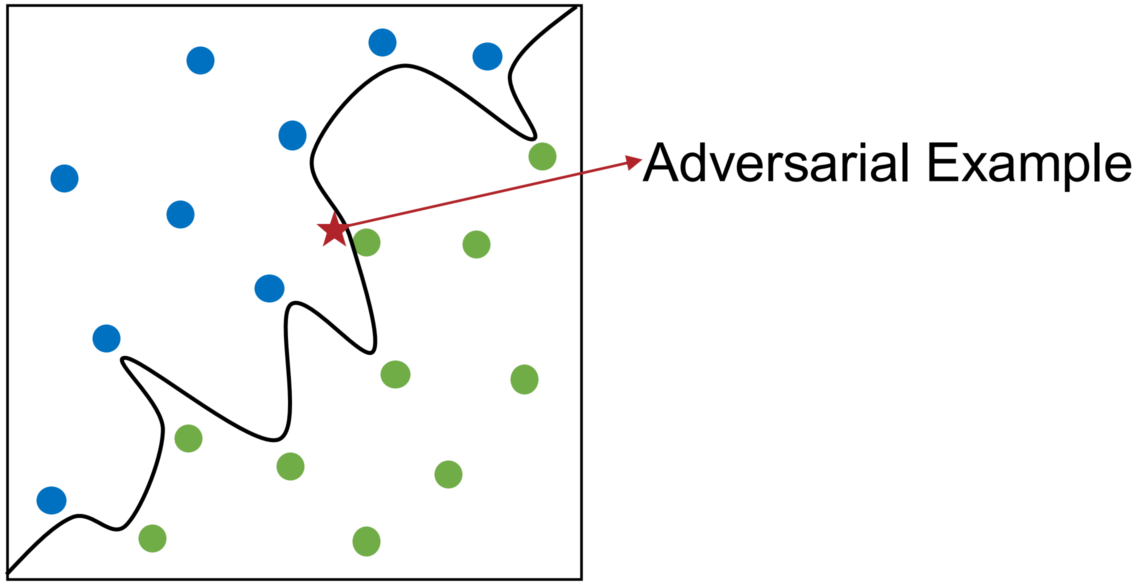
$$\max_{||\delta||_p \leq \epsilon} \mathcal{L}(F_\theta(x+\delta), y_{true})$$



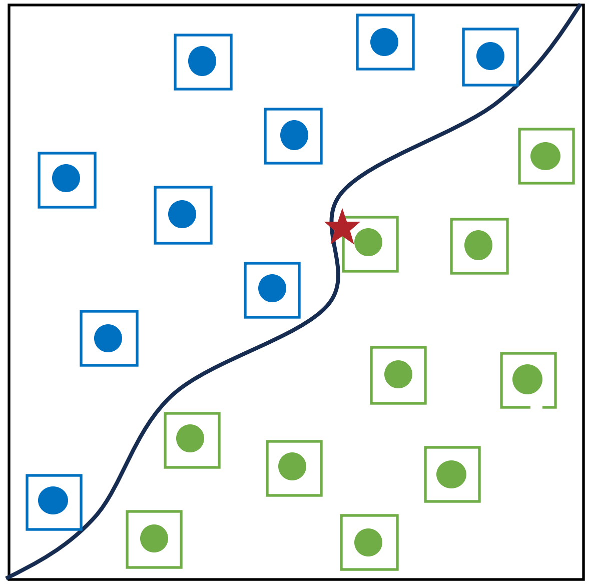
Empirical Defense vs. Certified Defense

- Perspective of Robust Defense for Deep Neural Networks (DNNs)
  - ➢ Empirical Defense
    - Improve the model's prediction accuracy in adversarial attacks through robust training.
  - ➢ Certified Defense
    - Provide the certified robust radius as the robustness certification of the predicted output.

Standard Training

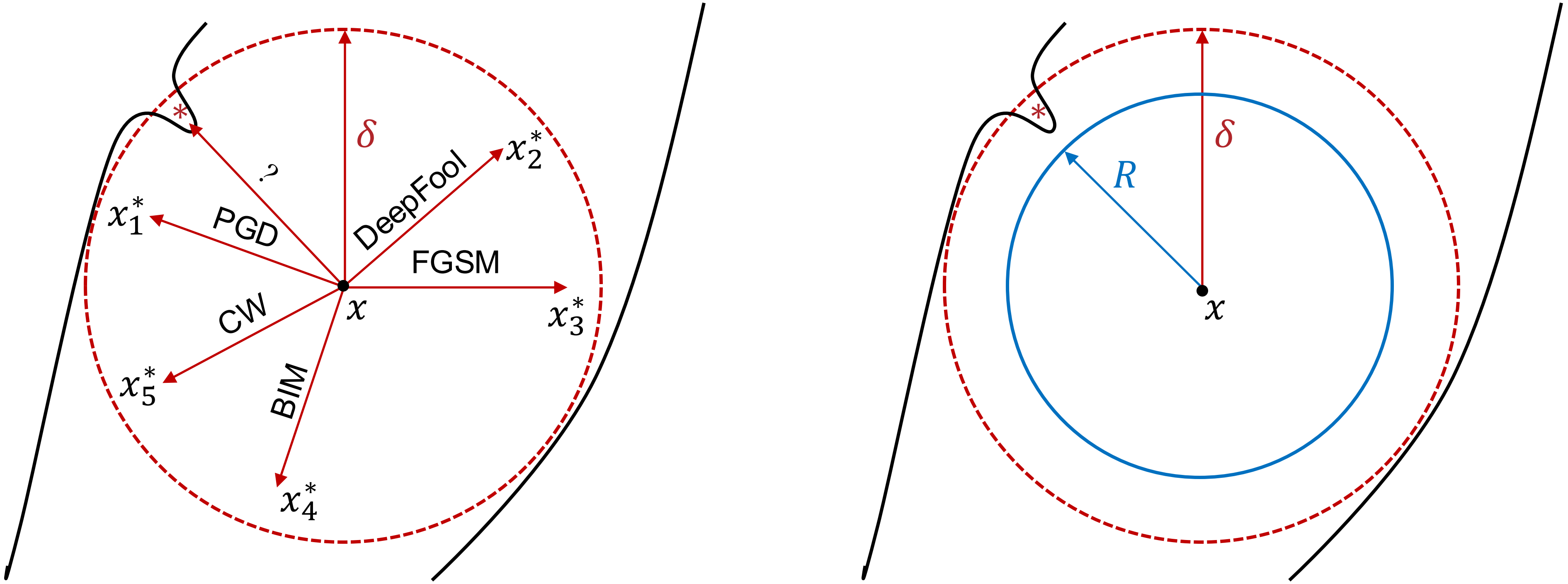


Adversarial Training



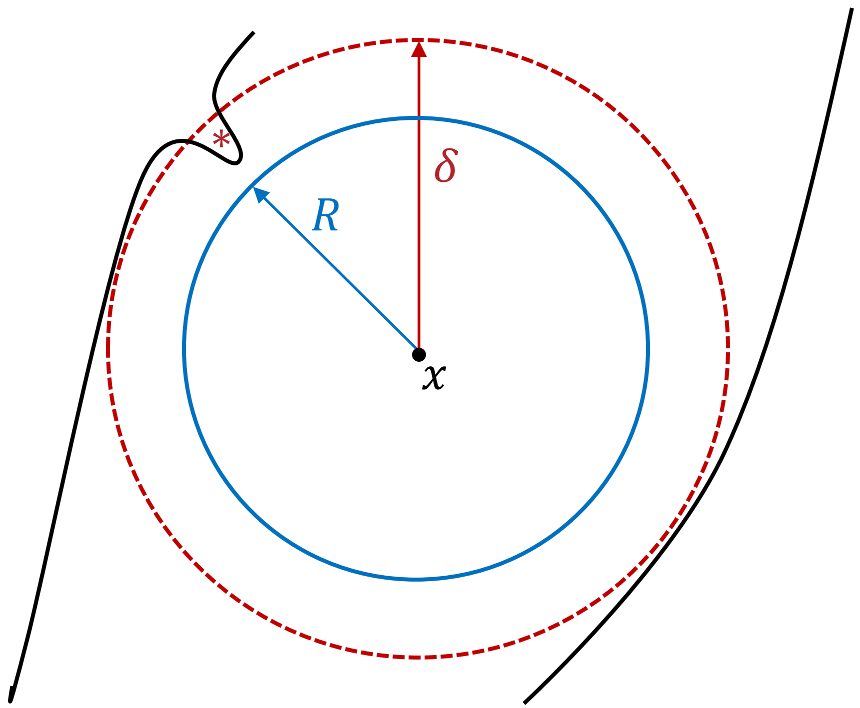
Empirical Defense vs. Certified Defense

- Perspective of Robust Defense for Deep Neural Networks (DNNs)
  - Empirical Defense
    - Improve the model's prediction accuracy in adversarial attacks through robust training.
  - Certified Defense
    - Provide the certified robust radius R as the robustness certification of the predicted output.



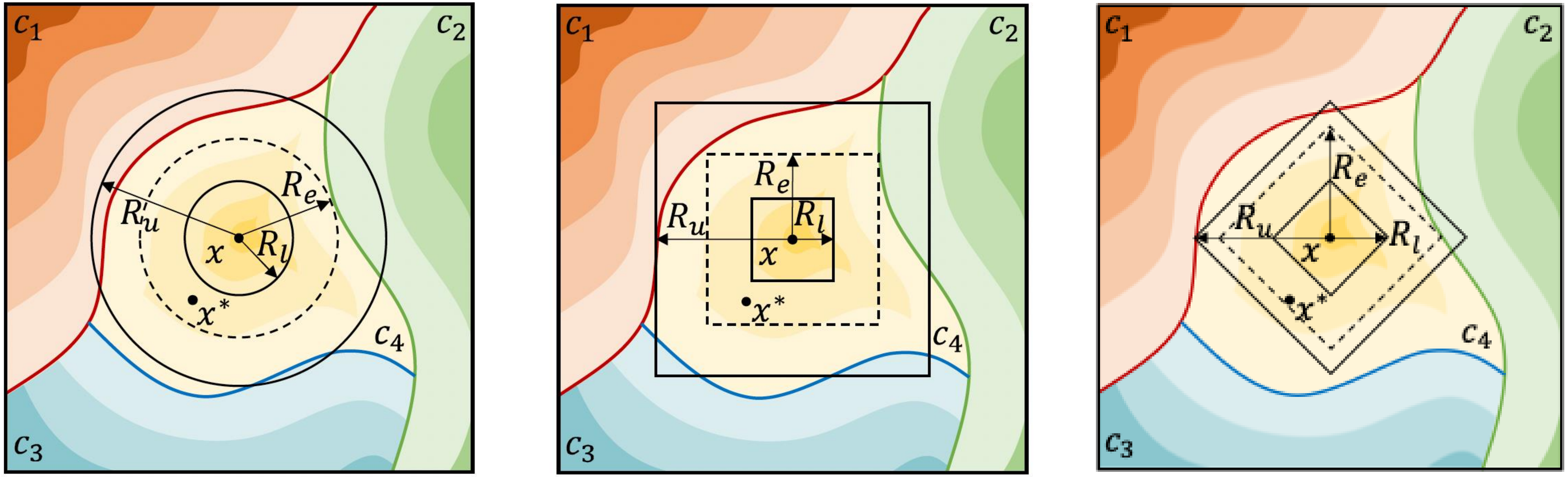
Empirical Defense vs. Certified Defense

- Perspective of Robust Defense for Deep Neural Networks (DNNs)
  - ➢ Empirical Defense
    - Improve the model's prediction accuracy in adversarial attacks through robust training.
  - ➢ Certified Defense
    - Provide the certified robust radius R as the robustness certification of the predicted output.
    - Robustness Guarantee
      - ✓ For input x, predictions of classifier F on perturbed data within an l_p norm-measured radius R around x, are guaranteed to remain consistent.
      - ✓ That is, any small perturbation δ to x within this region, including adversarial attacks, will not change the prediction results.



Certified Defense

- l_p Norm-bounded Certified Radius of DNN-based Multi-class Classifier on the Input x
  - Multiple Norm Types: l_2 norm, l_∞ norm, l_1 norm,
  - Exact Robust Radius: R_e
  - Upper/ Lower Bound of Exact Robust Radius: R_u, R_l



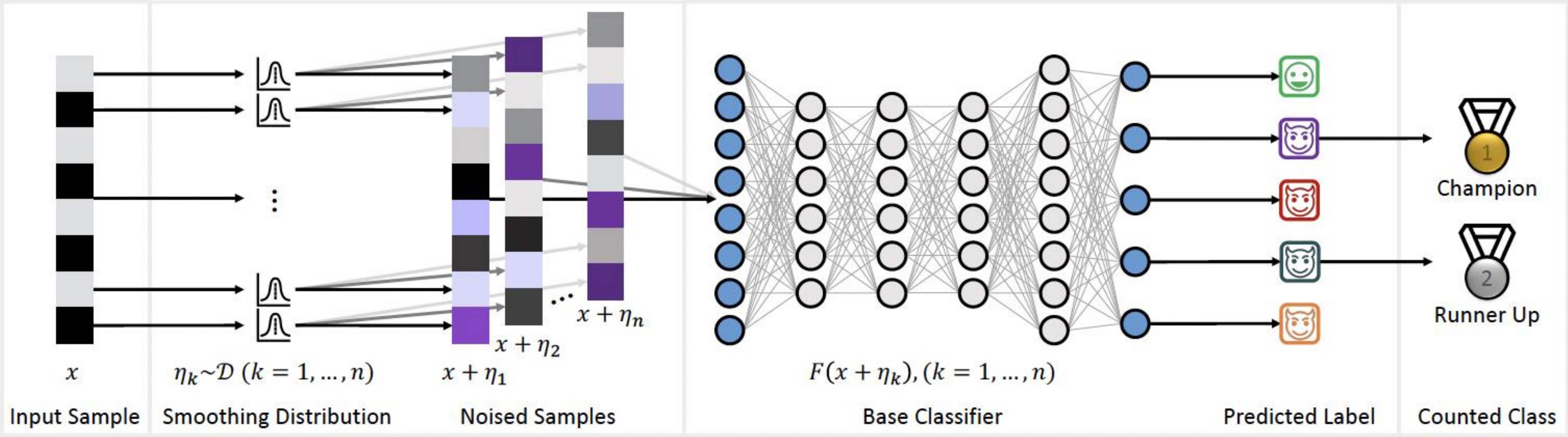
(a) $||\delta||_2 < R$ (b) $||\delta||_\infty < R$ (c) $||\delta||_1 < R$

Certify Robustness of DNN-based Network Traffic Classifiers

- Motivation
  - Certified defense efforts for network intrusion detection have been minimal, only BARS (NDSS'23).
  - The l_2 robustness guarantee is relatively loose and lacks certification for other l_p certified radii.
- Problems to be solved:
  - Pro1: Define a certified radius that can bound heterogeneous network traffic features.
  - Pro2: Expend the certified robust region to tighten the robustness guarantee.
  - Pro3: Provide the multiple l_p norms-bounded robustness guarantees of the model.
- Core Idea:
  - Extend the real-value certified radius R to a vector $(R_1, \ldots, R_d) \in \mathbb{R}^d$, where R_i denotes the dimensional certified radius for the i-th feature x_i of the heterogeneous input x.
  - Introduce the multiple order information of the smoothed classifier to expand the certified region.
  - Align the sampling area of smoothing distribution with the l_p-measured surroundings of the input.

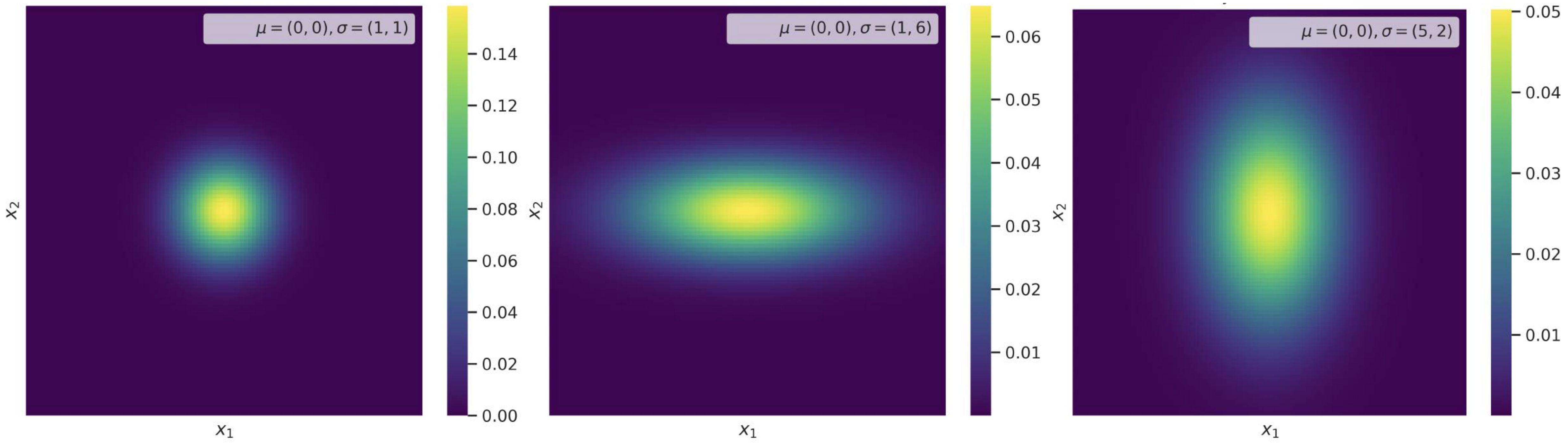
Robustness Certification for DNN-Based Network Traffic Classifiers via MARS

- Framework of Proposed Multi-Order Adaptive Randomized Smoothing (MARS)
  - Prediction Procedure
    - Sampling $n_k = n_{small}$ noise data → Predict the category of the input x.
  - Certification Procedure
    - Sampling $n_k = n_{large}$ noise data → Calculate the robust radius R of the model on the input x.



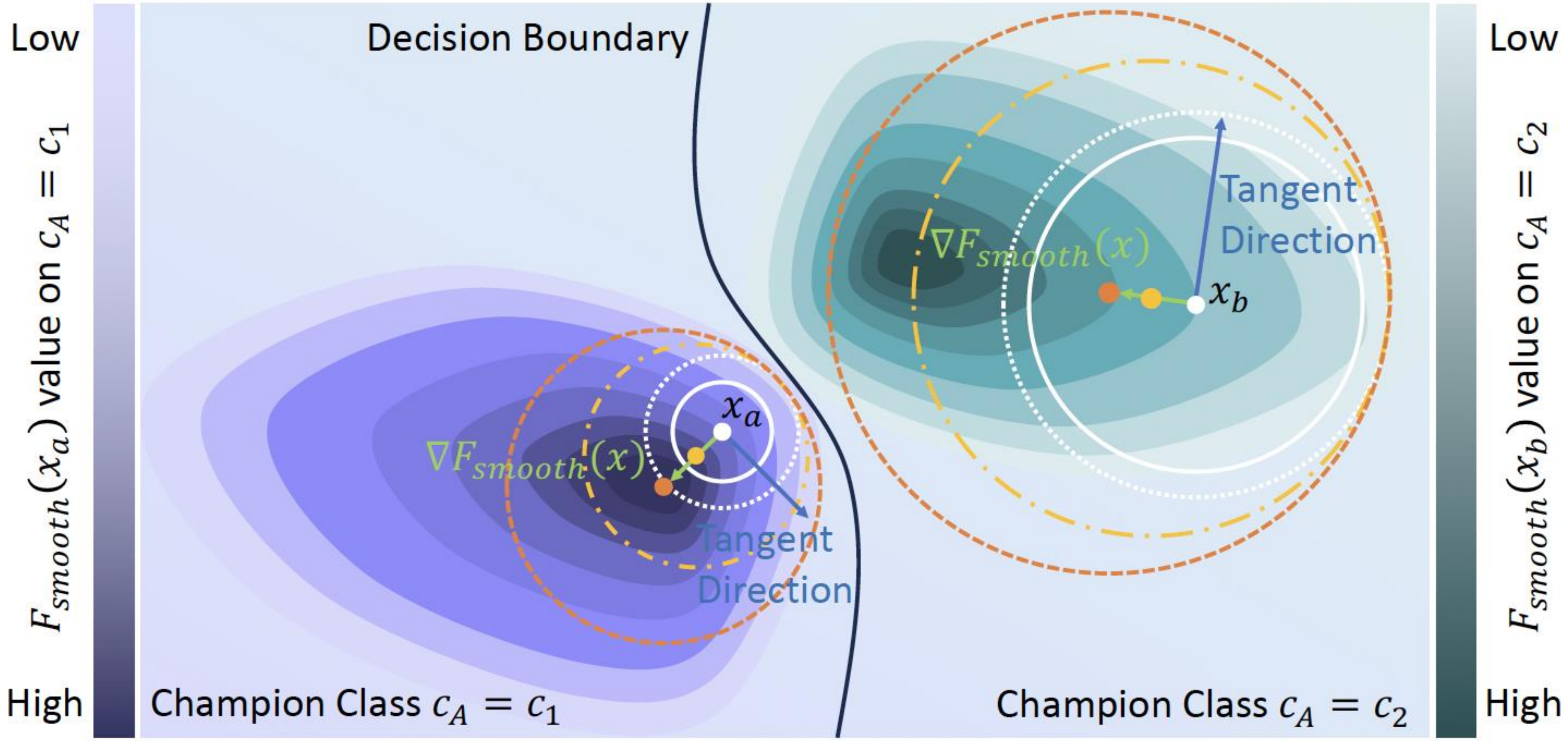
Robustness Certification for DNN-Based Network Traffic Classifiers via MARS

- Phase 1: Smoothing Distribution Parameters Optimization
  - Distribution Shape Optimization.
    - Encourage noised samples to be near the decision boundary of the classifier for x.
  - Distribution Scale Optimization.
    - Expand the noise sampling area by adjusting the smoothing distribution's scalar parameter.



Robustness Certification for DNN-Based Network Traffic Classifiers via MARS

- Phase 2: Multi-order Information-based Certified Robust Radius Calculation
  - Zero-order Output Probability Information-based Certified Radius Calculation
  - First-order Gradient Information-based Certified Radius Extension

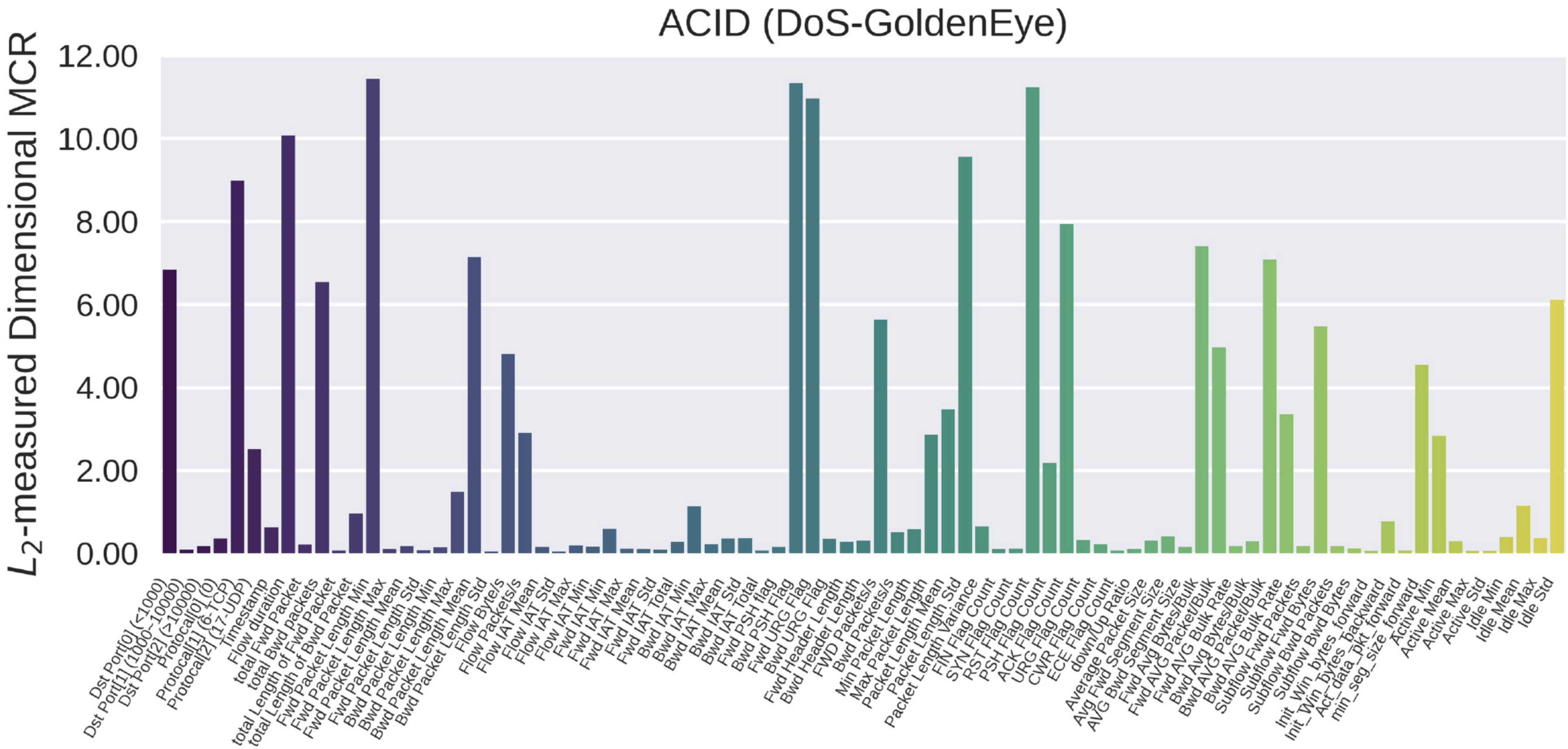


Robustness Certification for DNN-Based Network Traffic Classifiers via MARS

- Phase 3: Dimensional Robust Radius Weight Calculation Calculation
  - ➢ Dimensional Feature Sensitivity Analysis
  - ➢ Dimensional Radius Contribution Quantification

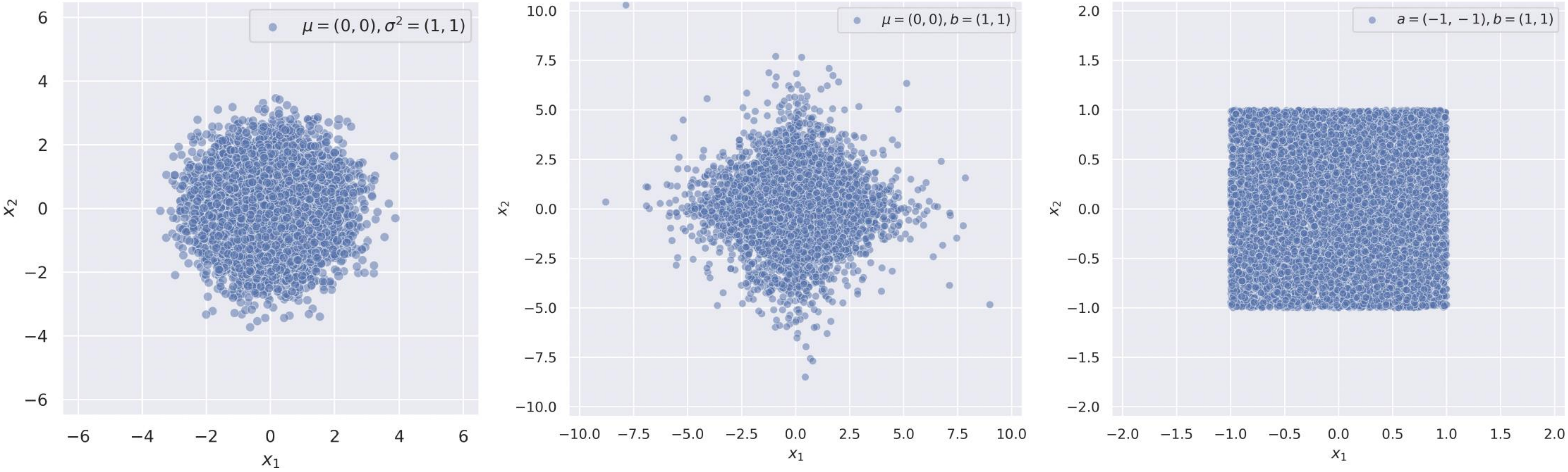
$$s_i = d(f_\theta^c(x)) / d(x_i) \qquad s = (s_1, \ldots, s_d)$$

$$R_i = w_i \times R, w_i = \frac{R_i}{R} = \frac{1/d}{\tilde{s}_i} = \frac{1}{d\tilde{s}_i}$$



Robustness Certification for DNN-Based Network Traffic Classifiers via MARS

- Smoothing Distribution Diversity
  - Gaussian Distribution aligns with l_2 norm-bounded certified region
  - Laplacian Distribution aligns with l_1 norm-bounded certified region
  - Uniform Distribution aligns with l_∞ norm-bounded certified region



Experimental Setup

- Dataset
  - ➢ Three datasets created from CIC-IDS-2018

Dataset	DoS-Hulk-Drift Dataset		Infiltration-Drift Dataset		Diverse-Intrusions Dataset	
	Class	Number	Class	Number	Class	Number
Training	Benign	52996	Benign	52996	Benign	52996
	SSH-Bruteforce	9385	SSH-Bruteforce	9385	FTP-Bruteforce	12590
	Infiltration	7390	DoS-Hulk	34789	DDoS-HOIC	53476
	-	-	-	-	Bot	22584
Test	Benign	13249	Benign	13249	Benign	13249
	SSH-Bruteforce	2346	SSH-Bruteforce	2346	FTP-Bruteforce	3148
	Infiltration	1894	DoS-Hulk	8697	DDoS-HOIC	13369
	DoS-Hulk	43486	Infiltration	9327	Bot	5646

- Model
  - ➢ CADE

    Contrastive Autoencoder for Drifting detection and Explanation

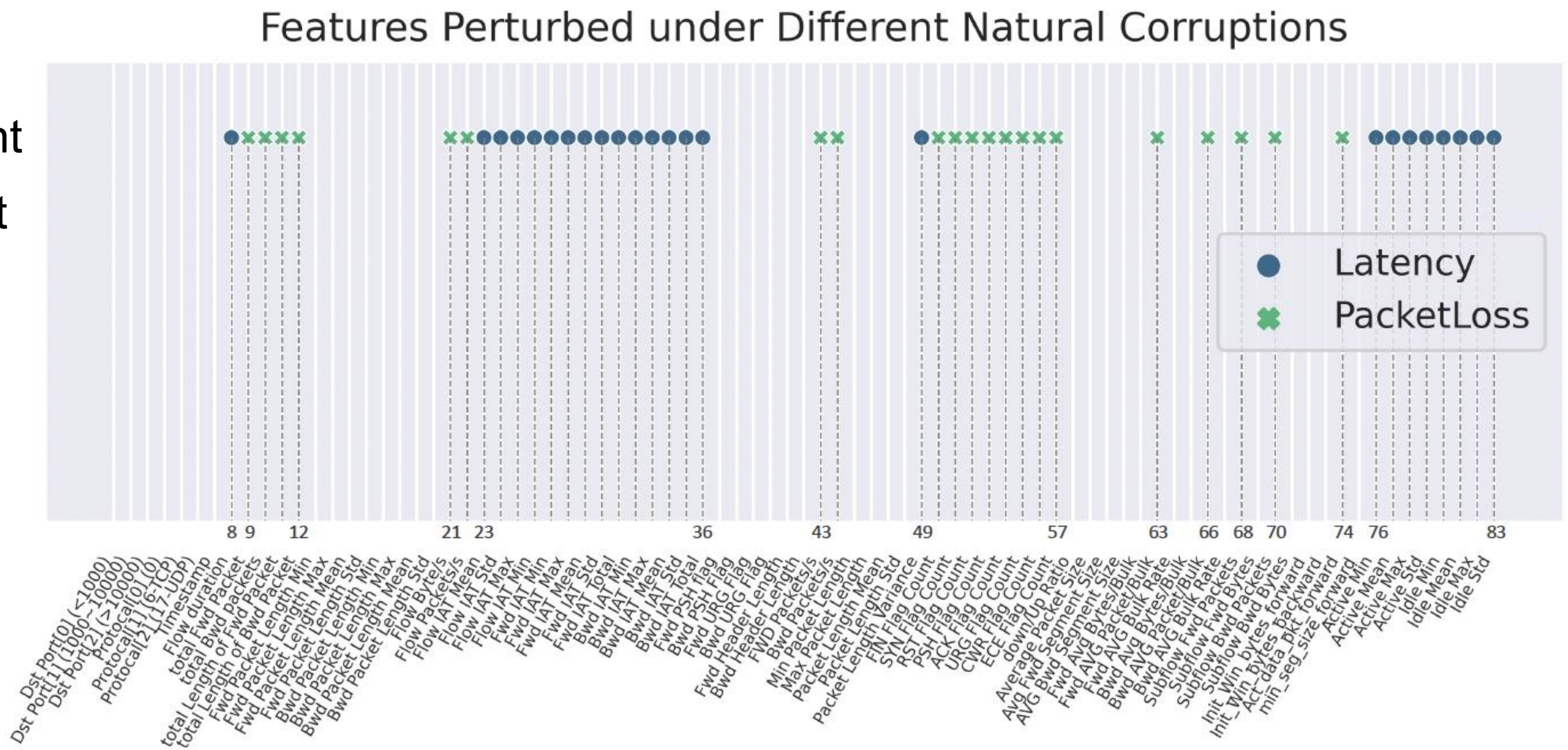
    (USENIX 2021)
  - ➢ ACID

    Adaptive Clustering-based Intrusion Detection

    (INFOCOM 2021)

Experimental Setup

- Attack Configuration
  - Evasion Attack
    - PGD: Projected Gradient Descent
    - EAD: Elastic-Net Attack to DNN
  - Natural Corruption
    - Latency
    - Packet Loss



Experimental Setup

- Attack Configuration
  - Evasion Attack
    - PGD: Projected Gradient Descent
    - EAD: Elastic-Net Attack to DNN
  - Natural Corruption
    - Latency
    - Packet Loss

Perturbed Featured under Latency

No	Feature Name	No	Feature Name
8	*Flow_Duration*	34	*Bwd_IAT_Mean*
23	*Flow_IAT_Mean*	35	*Bwd_IAT_Std*
24	*Flow_IAT_Std*	36	*Bwd_IAT_Total*
25	*Flow_IAT_Max*	50	*Packet_Length_Variance*
26	*Flow_IAT_Min*	76	*Active_Min*
27	*Fwd_IAT_Min*	77	*Active_Mean*
28	*Fwd_IAT_Max*	78	*Active_Max*
29	*Fwd_IAT_Mean*	79	*Active_Std*
30	*Fwd_IAT_Std*	80	*Idle_Min*
31	*Fwd_IAT_Total*	81	*Idle_Mean*
32	*Bwd_IAT_Min*	82	*Idle_Max*
33	*Bwd_IAT_Max*	83	*Idle_Std*

Experimental Setup

- Attack Configuration
  - Evasion Attack
    - PGD: Projected Gradient Descent
    - EAD: Elastic-Net Attack to DNN
  - Natural Corruption
    - Latency
    - Packet Loss

Perturbed Featured under Packet Loss

No	Feature Name	No	Feature Name
9	*Total_Fwd_Packet*	53	*PSH_Flag_Count*
10	*Total_Bwd_packets*	54	*ACK_Flag_Count*
11	*Total_Length_of_Fwd_Packet*	55	*URG_Flag_Count*
12	*Total_Length_of_Bwd_Packet*	56	*CWR_Flag_Count*
21	*Flow_Byte/s*	57	*ECE_Flag_Count*
22	*Flow_Packets/s*	63	*Fwd_AVG_Packet/Bulk*
43	*FWD_Packets/s*	66	*Bwd_AVG_Packet/Bulk*
44	*Bwd_Packets/s*	68	*Subflow_Fwd_Packets*
50	*FIN_Flag_Count*	70	*Subflow_Bwd_Packets*
51	*SYN_Flag_Count*	74	*Act_data_pkt_forward*
52	*RST_Flag_Count*	-	-

Experimental Setup

- Comparison of Certified Defense Methods
  - VRS: Vanilla Randomized Smoothing (ICML 2019) → designed for Image
  - FRS: First Order-based Randomized Smoothing (NeurIPS 2020) → designed for Image
  - BARS: Boundary-Adaptive Randomized Smoothing (NDSS 2023) → designed for Traffic

Method	Heterogeneity	Universality	Robustness Guarantee Diversity			Adversarial Attacks			Natural Corruptions	
			l_2 Radius	l_1 Radius	l_∞ Radius	l_2 Attack	l_1 Attack	l_∞ Attack	Latency	Loss
VRS [17]	○	●	●	○	○	○	○	○	○	○
FRS [35]	○	●	●	●	●	○	○	○	○	○
BARS [18]	●	●	●	○	○	○	○	●	○	○
MARS	●	●	●	●	●	●	●	●	●	●

[17] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In International Conference on Machine Learning (ICML). 1310–1320.

[35] Jeet Mohapatra, Ching-Yun Ko, Tsui-WeiWeng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. 2020. Higher-order certification for randomized smoothing. In Advances in Neural Information Processing Systems (NeurIPS). 4501–4511.

[18] Kai Wang, Zhiliang Wang, Dongqi Han, Wenqi Chen, Jiahai Yang, Xingang Shi, and Xia Yin. 2023. BARS: Local Robustness Certification for Deep Learning based Traffic Analysis Systems. In Network and Distributed Systems Security (NDSS) Symposium.

Experimental Setup

- Evaluation Metrics
  - Certified Robustness
    - Mean Certified Radius
    - Certified Accuracy

$$Mean\ Certified\ Radius\ (MCR) = \frac{1}{N}\sum_{i=1}^{N} R_i$$

$$Certified\ Accuracy\ (CerAcc) = \frac{N_{(F_{smooth}(x)=y_{true})\&(R\geq R_{given})}}{N}$$

  - Empirical Robustness
    - Robust Accuracy on Adversarial (Malicious) Examples

$$Recall = \frac{TP}{TP+FN}$$

    - Robust Accuracy on Corrupted (Malicious & Benign) Examples

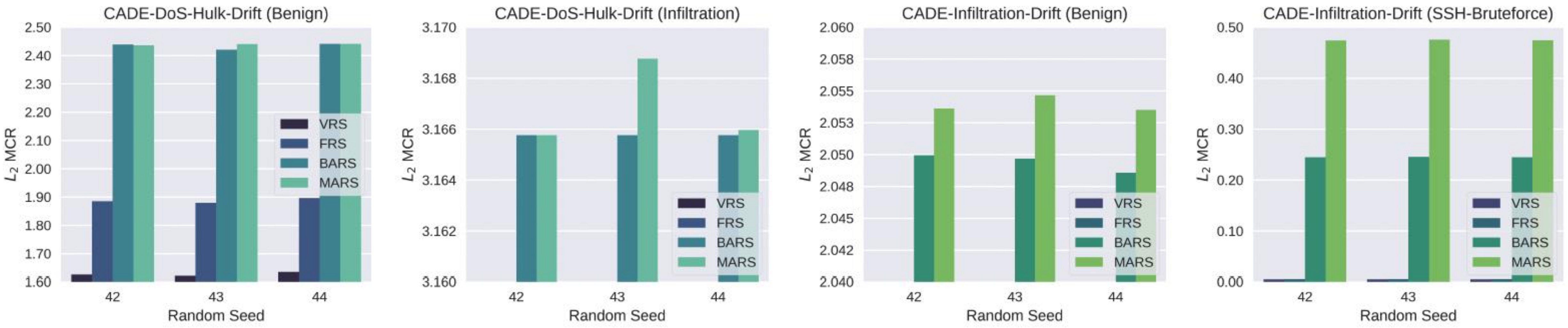
$$Robust\ Accuracy\ (RobAcc) = \frac{N_{(F_{smooth}(x^*)=y_{true})}}{N} = \frac{TP+TN}{TP+TN+FP+FN}$$

  - Regular Predictive Performance
    - Clean Accuracy

$$Clean\ Accuracy\ (CleAcc) = \frac{N_{(F_{smooth}(x)=y_{true})}}{N}$$

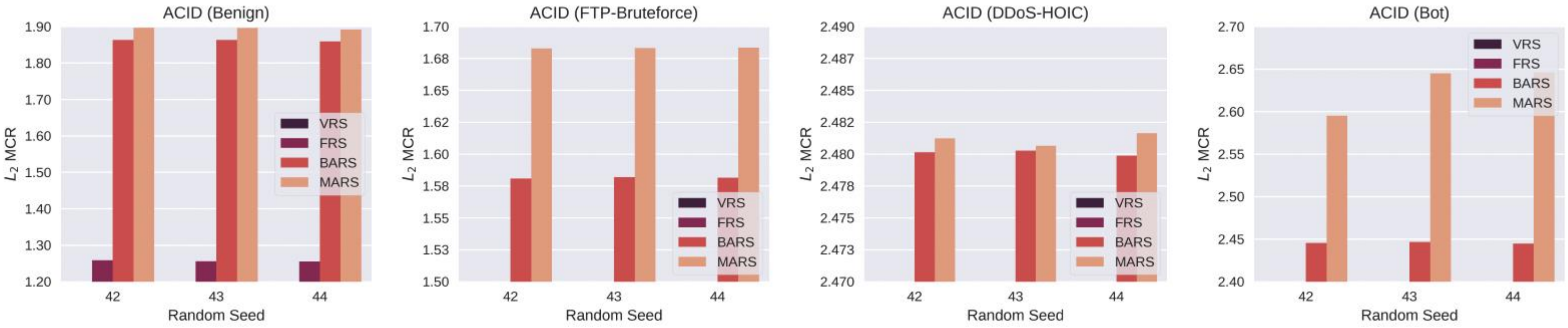
Evaluation Results and Analysis

- Exp 1: Comparison of l_2-bounded Certified Robustness with SOTA Method
  - ➢ Exp Setup: n_{small} =100, n_{large} =10,000. Compare the l_2 overall MCR R of the model by category.
  - ➢ Observation: ***MARS always outperforms certified defense baselines VRS, FRS, and BARS***.
    - For CADE trained on DoSHulk-Drift dataset, MARS shows a 0.23% and 0.03% higher MCR in Benign and Infiltration classes, respectively, than SOTA BARS.
    - For CADE trained on Infiltration-Drift dataset, MARS exhibits a 0.22%, 93.66%, and 0.2% MCR increase in Benign, SSH-Bruteforce, and DoS-HULK categories compared to BARS.



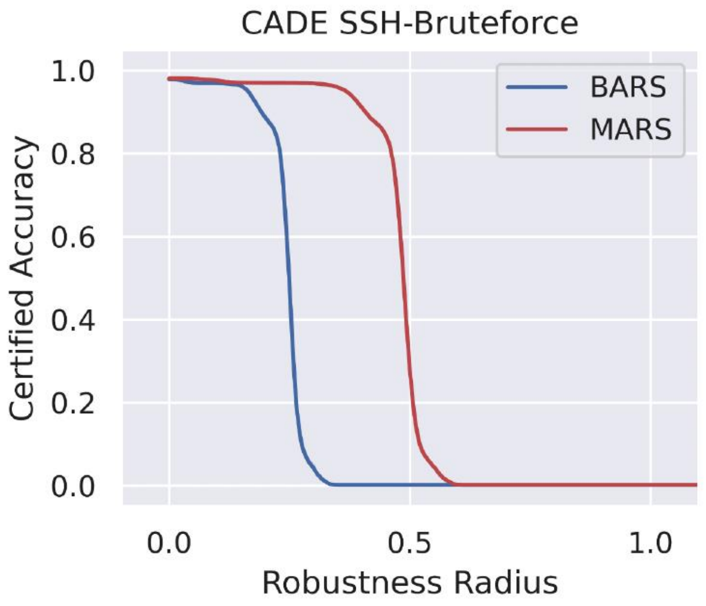
Evaluation Results and Analysis

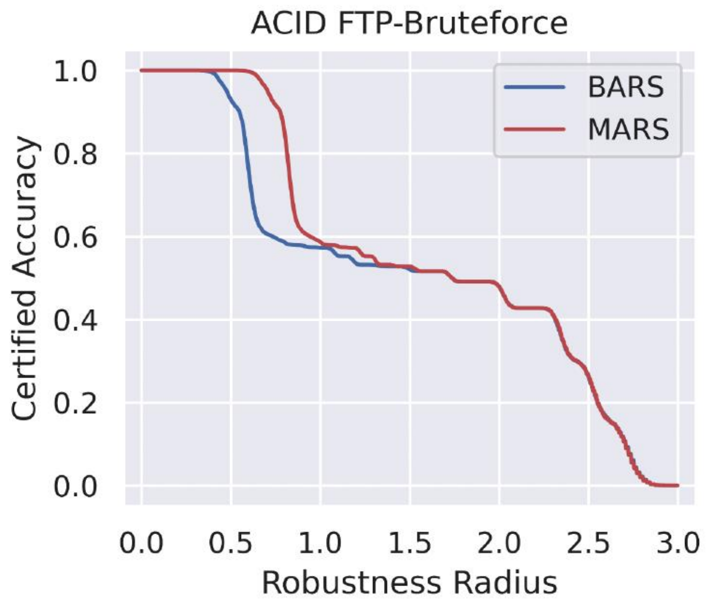
- Exp 1: Comparison of l_2-bounded Certified Robustness with SOTA Method
  - Exp Setup: n_{small} =100, n_{large} =10,000. Compare the l_2 overall MCR R of the model by category.
  - Observation: ***MARS always outperforms certified defense baselines VRS, FRS, and BARS***.
    - For ACID trained on Diverse Intrusion dataset, MARS exhibits a 1.75%, 6.44%, 0.04%, and 7.49% MCR increase in Benign, FTP-Bruteforce, DDoS-HOIC, and Bot categories compared to SOTA Certified Defense BARS.



Evaluation Results and Analysis

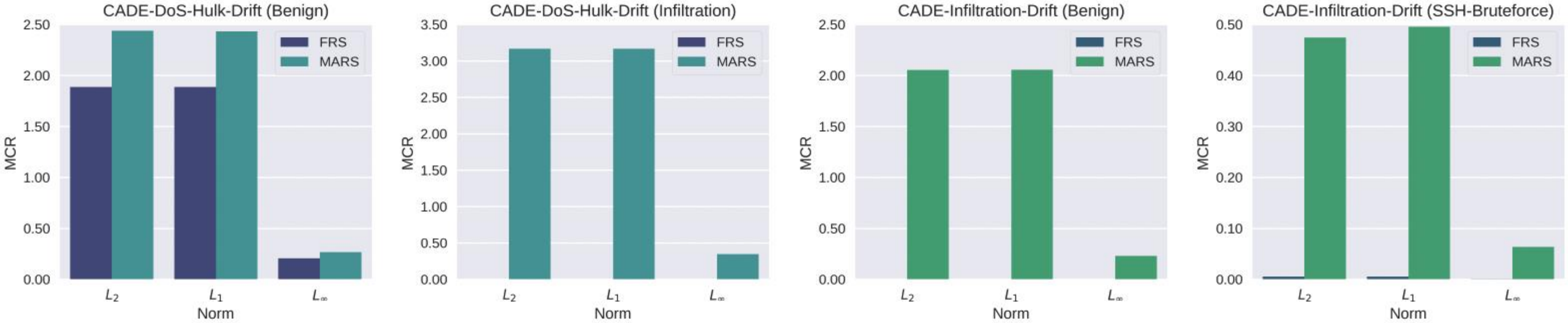
- Exp 1: Comparison of l_2-bounded Certified Robustness with SOTA Method
  - Exp Setup: Compare the Certified Accuracy of the model w.r.t the l_2-bounded certified radius.
  - Observation: ***MARS demonstrated the certified robustness of the model in a larger region***.
    - For CADE, MARS maintains 100% accuracy until the MCR threshold reaches 0.4, while the that of the SOTA methods begins to drop sharply when the threshold just exceeds 0.15.
    - For ACID, MARS shows significant advantages over SOTA until the MCR reaches 1.5.





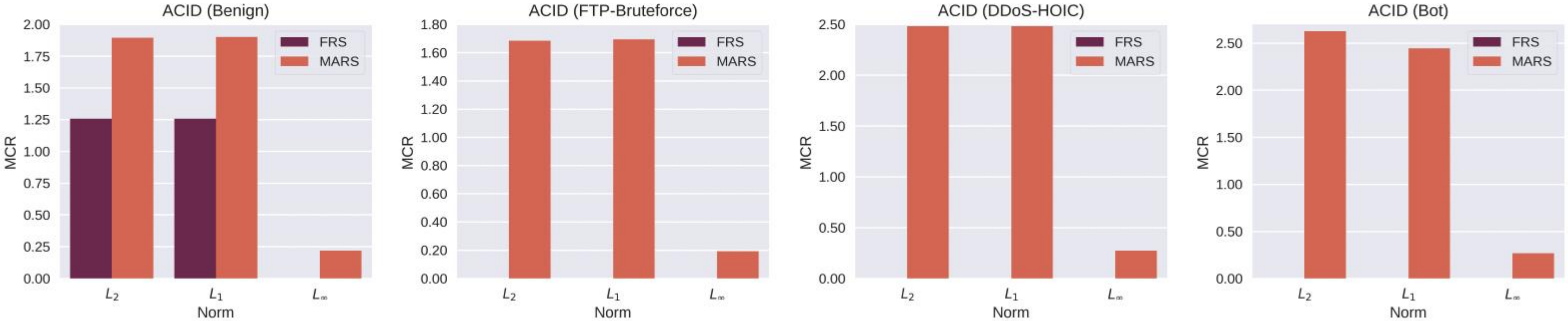
Evaluation Results and Analysis

- Exp 2: Comparison of Various l_p-bounded Certified Robustness with SOTA Method
  - Exp Setup: n_{small} =100, n_{large} =10,000. Compare the l_1, l_∞ MCR of the model by category with FRS, since neither VRS nor BARS supports l_1-bounded and l_∞-bounded robustness certification.
  - Observation: ***MARS consistently provides larger l_p-bounded radius compared to FRS***.
    - FRS fails certification on many classes (MCR=0) due to indiscriminate smoothing of network traffic features, MARS produces non-trivial l_2, l_1, and l_∞ radii.
    - For CADE trained on DoSHulk-Drift dataset, MARS outperforms FRS by 29.25%, 28.95%, and 28.72% in l_2, l_1, and l_∞ radii on Benign, respectively.



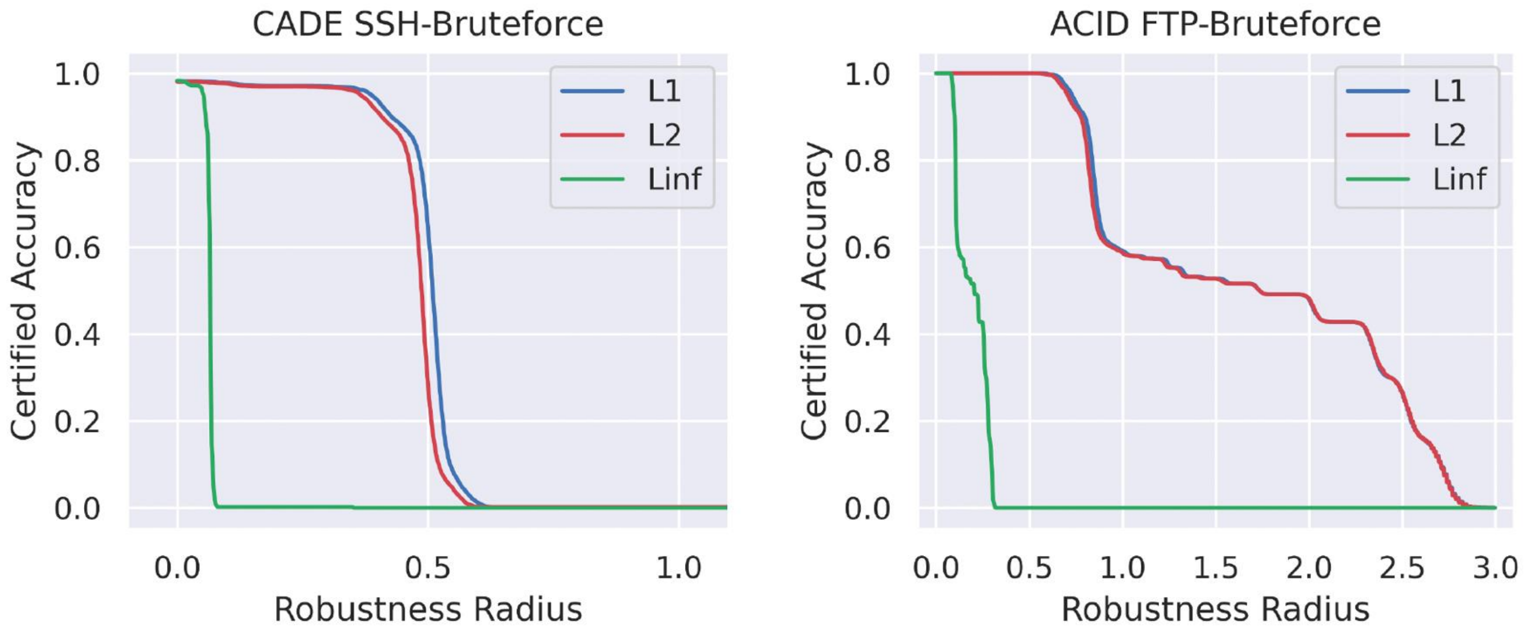
Evaluation Results and Analysis

- Exp 2: Comparison of Various l_p-bounded Certified Robustness with SOTA Method
  - Exp Setup: n_{small} =100, n_{large} =10,000. Compare the l_1, l_∞ MCR of the model by category with FRS, since neither VRS nor BARS supports l_1-bounded and l_∞-bounded robustness certification.
  - Observation: ***MARS consistently provides larger l_p-bounded radius compared to FRS***.
    - FRS fails certification on many classes (MCR=0) due to indiscriminate smoothing of network traffic features, MARS produces non-trivial l_2, l_1, and l_∞ radii.
    - For ACID trained on Diverse Intrusion dataset, MARS outperforms FRS by 50.78% and 51.32% in l_2 and l_1 radii on Benign, respectively.



Evaluation Results and Analysis

- Exp 2: Comparison of Various l_p-bounded Certified Robustness with SOTA Method
  - Exp Setup: Compare the Certified Accuracy of the model w.r.t the l_p-bounded certified radius.
  - Observation: ***l_2 radius is usually smaller than the l_1 radius and larger than the l_∞ radius.***
    - At the same radius, the area bounded by l_1 norm should be the smallest, and the area defined by l_∞ should be the largest.
    - Different norm-bounded radii calculated experimentally are consistent with theoretical results.



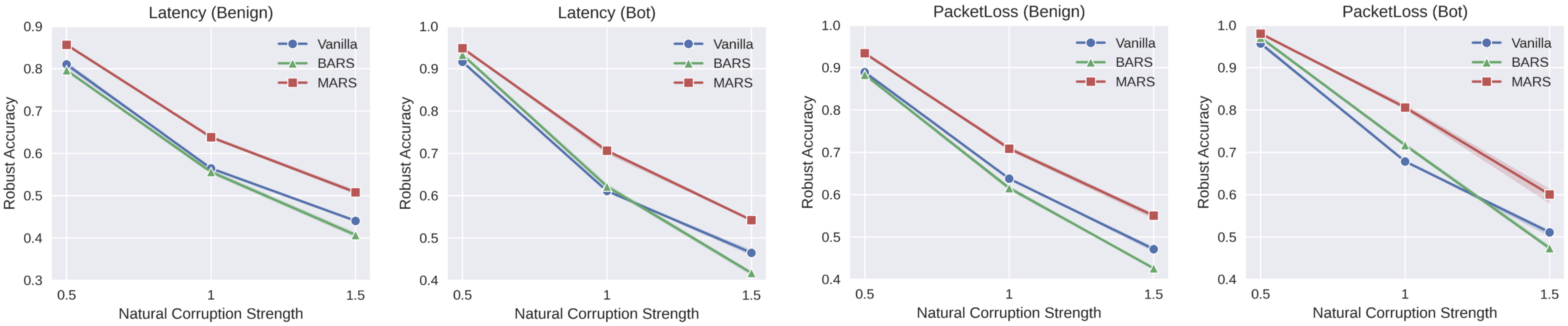
Evaluation Results and Analysis

- Exp 3: Comparison of Empirical Robustness against Evasion Attacks with SOTA Method
  - Exp Setup: Attack ACID with PGD and EAD adversarial Bot. Iteration is 20. For l_2-PGD and l_1-EAD, perturbation limit ϵ is 1.0, with per-step budget ϵ_s of 0.75. For l_∞-PGD, ϵ is 0.2 and ϵ_s is 0.1.
  - Observation: ***MARS surpasses SOTA defense in robustness against evasion attacks***.
    - MARS improves robust accuracy over the Vanilla detector (base model without defense) by 13.79% for l_2-PGD, 33.94% for l_∞-PGD, and 10.01% for l_1-EAD.
    - MARS also outperforms SOTA BARS, boosting robust accuracy by 1.7% for l_2-PGD, 7.17% for l_∞-PGD, and 10.11% for l_1-EAD.
    - MARS well retain the clean accuracy of the ACID on clean Bot samples, reaching 100%.

Method	CleanAcc/Recall on Clean Bot (%)	RobustAcc/Recall on Adversarial Bot (%)		
		l_2-PGD	l_∞-PGD	l_1-EAD
Vanilla	100.00±00.00	83.95±00.00	55.02±00.01	00.27±00.00
BARS [18]	100.00±00.00	96.04±00.05	81.78±00.20	00.16±00.01
MARS	100.00±00.00	**97.74±00.13**	**88.95±00.31**	**10.28±00.06**

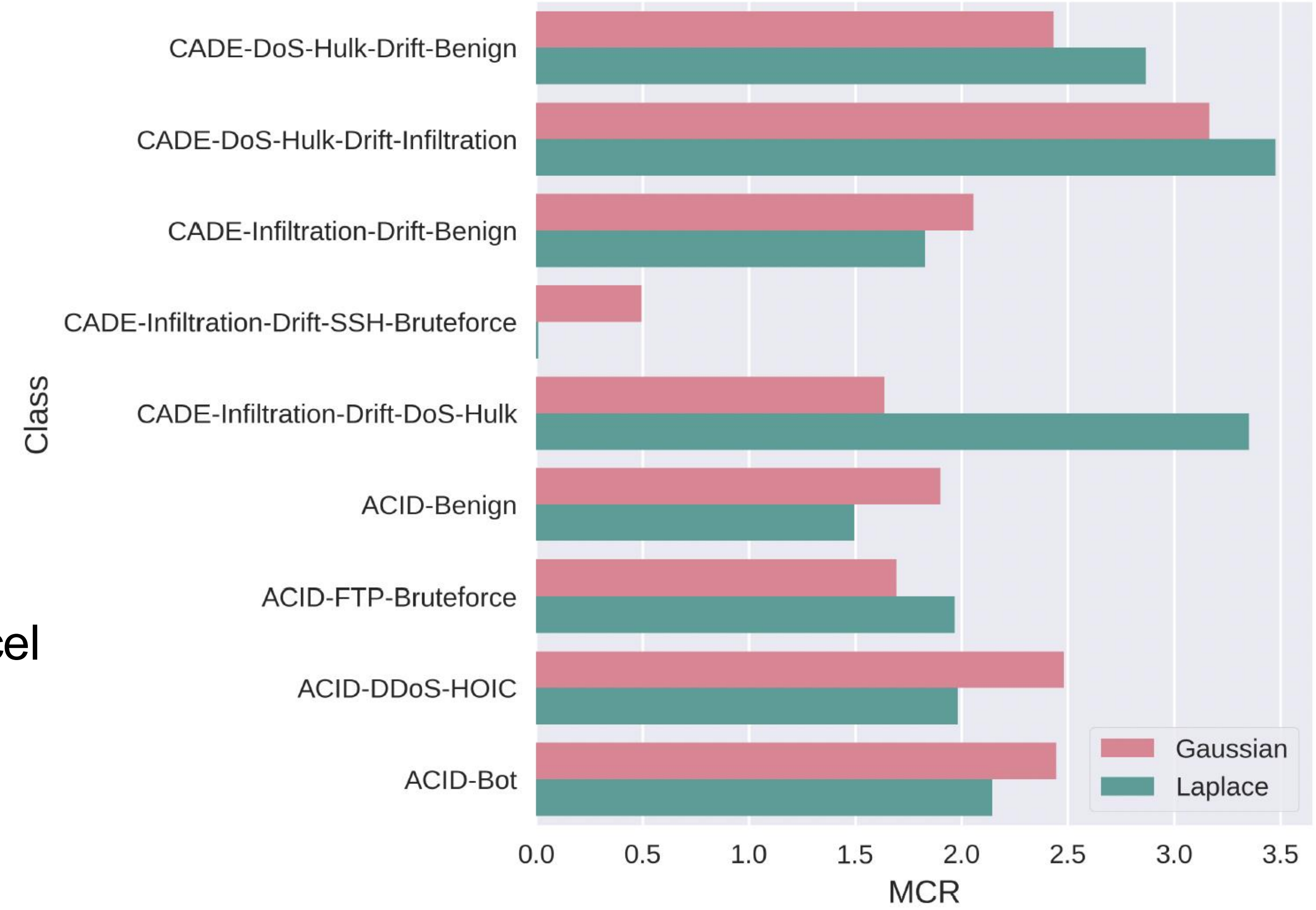
Evaluation Results and Analysis

- Exp 4: Comparison of Empirical Robustness against Natural Corruptions with SOTA Method
  - ➢ Exp Setup: Generate natural corrupted samples from clean benign/malicious samples using Latency and PacketLoss. Use random noise following a Gaussian distribution with mean 0. Adjust the standard deviation σ in {0.5, 1.0, 1.5} to mimic the different corruption strengths.
  - ➢ Observation: ***MARS surpasses SOTA in robustness against various corruption intensities***.
    - MARS outperforms SOTA BARS in robust accuracy, exceeding it by 8.53% on corrupted Benign and 7.5% on corrupted Bot.



Evaluation Results and Analysis

- Exp 5: l_p Certified Robustness with Different Smoothing Distributions
  - Exp Setup:
    - All baselines use Gaussian as the smoothing distribution.
    - MARS considers distribution diversity and sequentially uses Gaussian, Laplacian, and Uniform distributions.
  - Observation:
    - Different distributions each excel in different classes.
    - Using a single distribution may miss a tighter certified radius.



Evaluation Results and Analysis

- Exp 6: Dimension-Wise Certified Robustness
  - Exp Setup: MARS's Top-5 and bottom-5 dimension-wise radius of the ACID.
  - Observation:
    - The model demonstrates greater sensitivity to *inter arrival time (IAT)-related* features while showing greater robustness to *forward packet length-related* features.
    - This finding is consistent with the previous observation that the vanilla ACID model exhibited significantly reduced robust accuracy on corrupted samples using Latency.

No	Radius	FeatureName	Description
24	0.0426	Flow_IAT_Std	Standard deviation time two flows.
20	0.0433	Bwd_Packet_Length_Std	Standard deviation size of packet in backward direction.
79	0.0488	Active_Std	Standard deviation time a flow was active before becoming idle.
72	0.0569	Init_Win_bytes_forward	Number of bytes sent in initial window in the forward direction.
78	0.0576	Active_Max	Maximum time a flow was active before becoming idle.
8	10.0741	Flow_Duration	Flow duration.
39	10.9644	Fwd_URG_Flag	Number of times URG flag was set in packets travelling in the forward direction (0 for UDP).
52	11.2367	RST_Flag_Count	Number of packets with RST.
38	11.3300	Bwd_PSH_Flag	Number of times PSH flag was set in packets travelling in the backward direction (0 for UDP).
13	11.4358	Fwd_Packet_Length_Min	Minimum size of packet in forward direction.
All	2.2305	MCR	Mean certified radius per class.

Summary

- Contribution
  - Robustness Certification Framework
    - Proposed MARS, a novel certification framework to calculate the robust radius of DNN-based network intrusion detectors that requires no modification to model structure.
  - Multi-Order Information Utilization
    - Introduced a method to expand certified regions by leveraging multi-order information of the classifier beyond zero-order techniques.
  - Dimensional-Wise Robust Radius
    - Designed a dimensional robust radius calculation approach for inputs with heterogeneous features, like network traffic.
  - New Threat Model
    - Extended empirical robustness evaluation of traffic classifier to account for natural corruption (e.g., Latency and Packet Loss) in addition to evasion attacks using adversarial examples.

Future Work

- Target issues
  - Non-l_p Robustness Certification against Structural Perturbations
    - Different from the l_p-norm bounded changes of input features, for structural perturbations that change the overall structure or composition of the input (such as adding, deleting, or reordering nodes/edges in a graph), special non-l_p robustness certification is needed to evaluate and guide the model's robustness improvement.
  - Robustness Certification for Multi-modal Models
    - Current certified defense techniques often face challenges in evaluating robustness across multiple data modalities. Designing a framework that can certify robustness by considering the interactions between heterogeneous and homogeneous data inputs simultaneously will be interesting.

Thank You!

Mengdie Huang[1,2], **Yingjun Lin**[2], Xiaofeng Chen[1], Elisa Bertino[2]

[1] Xidian University

[2] Purdue University

Q&A

Yingjun Lin (Link)
lin1368@purdue.edu