THE UNIVERSITY *of* EDINBURGH
informatics

Towards Secure and Resilient IoT Infrastructures

an AI Perspective

Alec F. Diallo

September, 2021

Agenda

1. Boosting the performance of ML classifiers
    - Task: Network Intrusion Detection
    - Constraints: Lightweight / Deployable at the Edge

2. Protecting ML classifiers against adversarial attacks

Boosting the performance of ML classifiers

Network Intrusion Detection Systems

- Two main approaches: *Knowledge oriented* / *Data oriented*

- Shortcomings of existing solutions:

Severe

Threat Level

Low

★ Volume of false alarms too high for practical usage

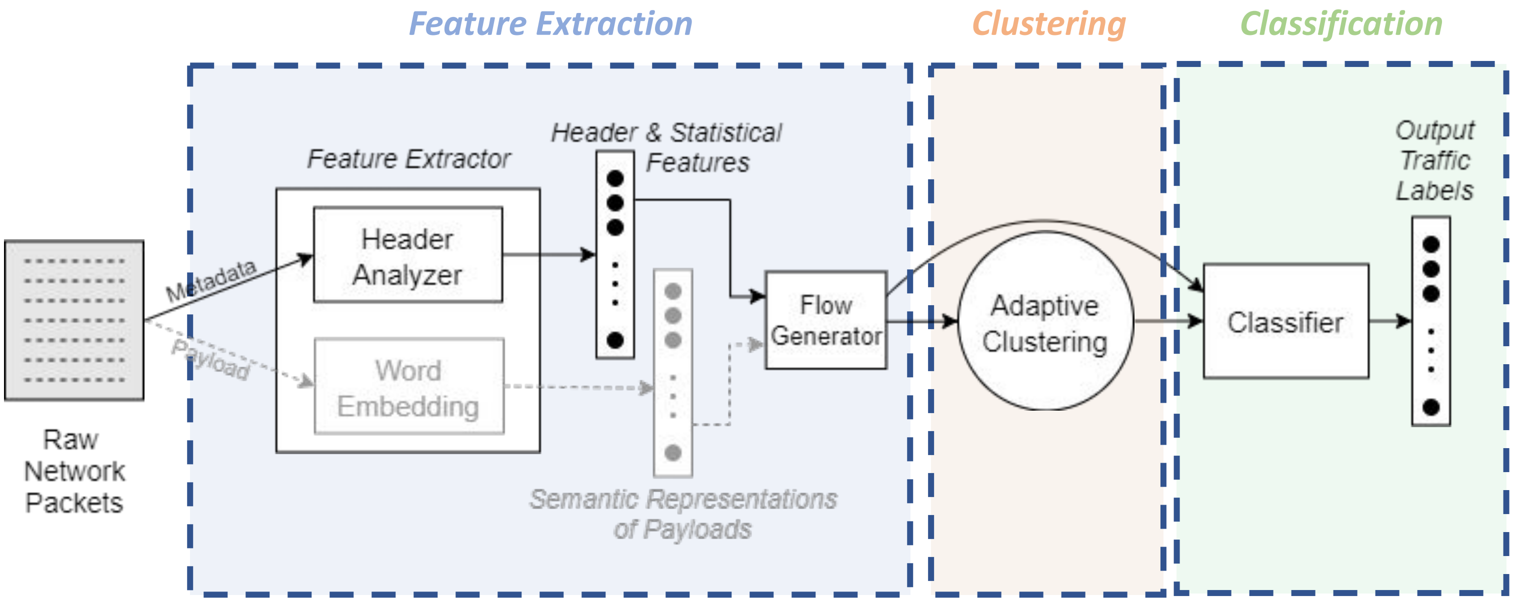
★ Performance degradation with increasing number of attack types

★ Unable to distinguish similar but different attacks

★ Trade-off between speed and accuracy

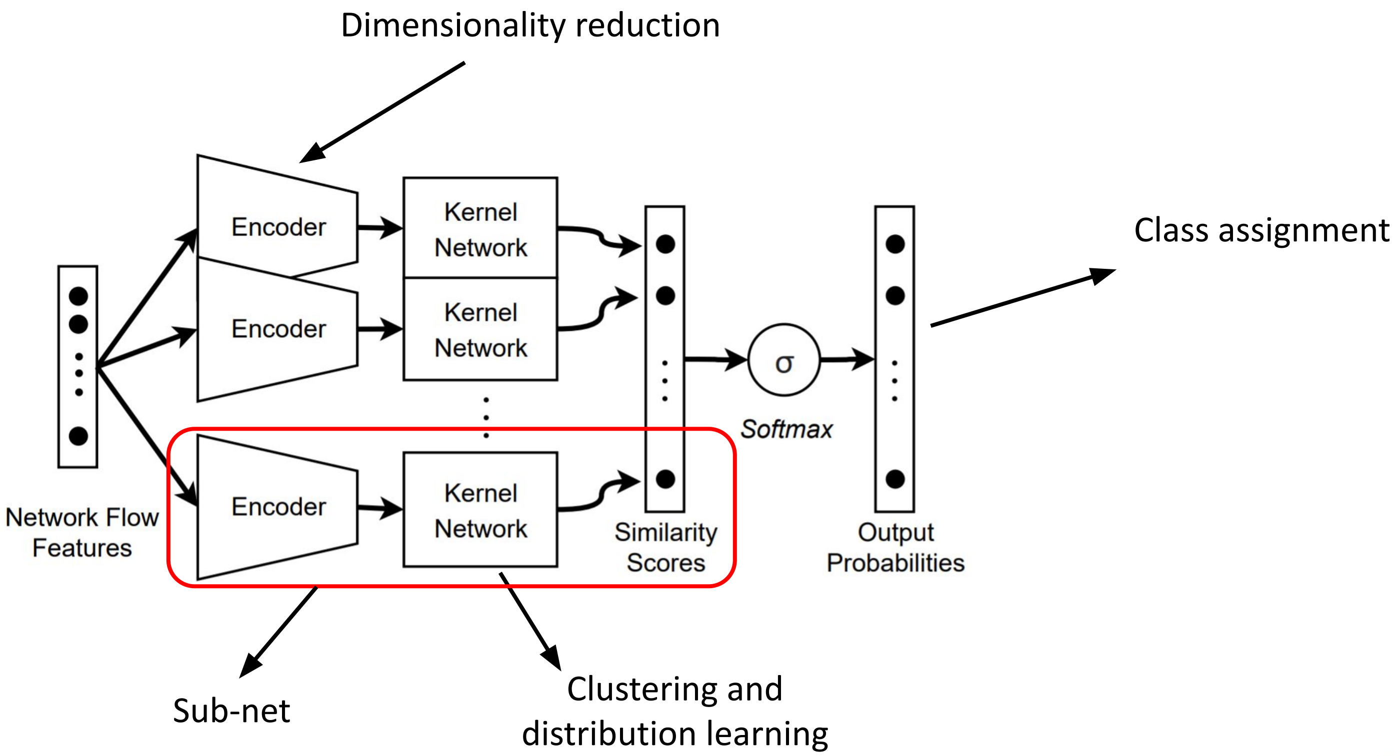
- Threat models: *Attacker inside/outside the LAN*

5 Proposed Solution

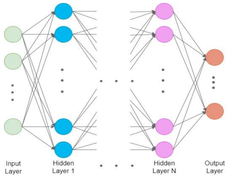


ACID Architecture: Adaptive Clustering-based Intrusion Detection

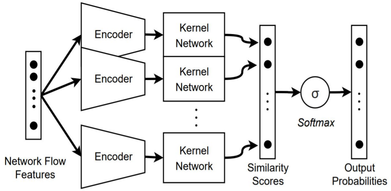
Solution | Adaptive Clustering network (AC-Net)



Solution | Adaptive Clustering network (AC-Net)

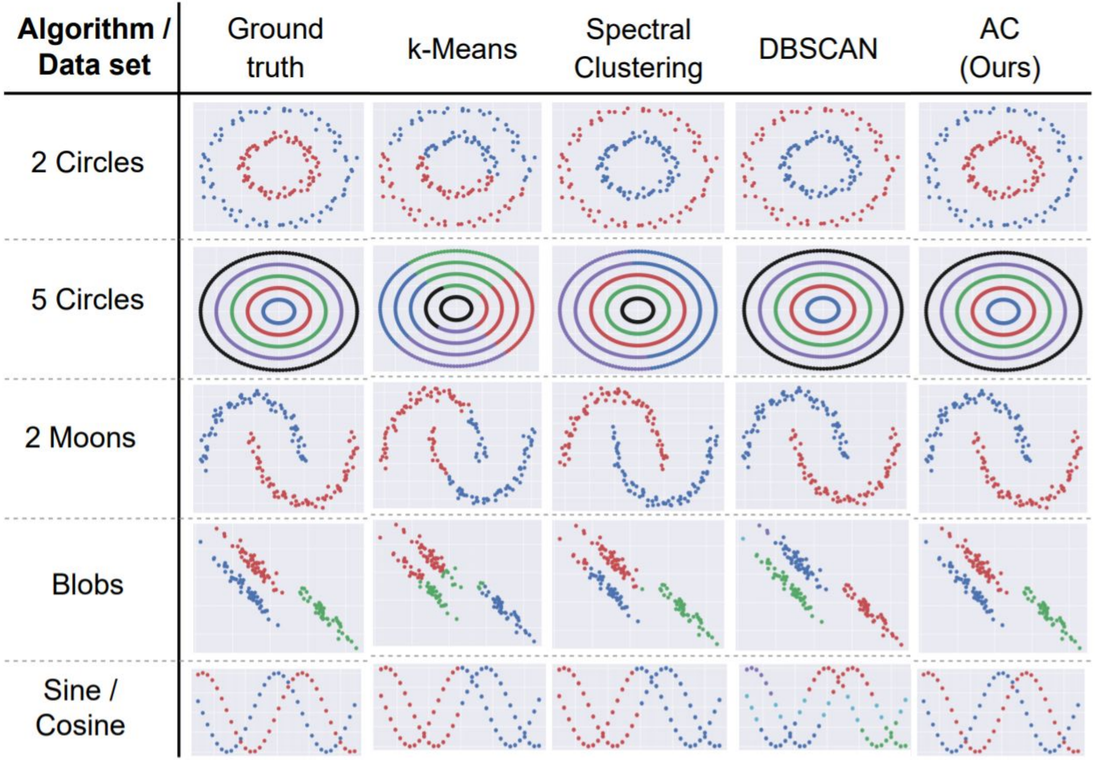


VS



	Classical Neural Networks	***Adaptive Clustering Networks***
Scalability	Difficult	Easy
Parallelization	Data	Data + Sub-nets
Model Complexity	High (1 network = all tasks)	Low (1 sub-net = 1 task)
Architecture	Fixed (high risk of network saturation, conflicts in learned parameters)	Flexible (no network saturation, no conflict in learned parameters)
Sensitivity	Extreme (input features, unbalanced datasets, ...)	Marginal
Advantages	None	- Optimal class separation - Intrinsic support for continual learning - Built-in clustering mechanism

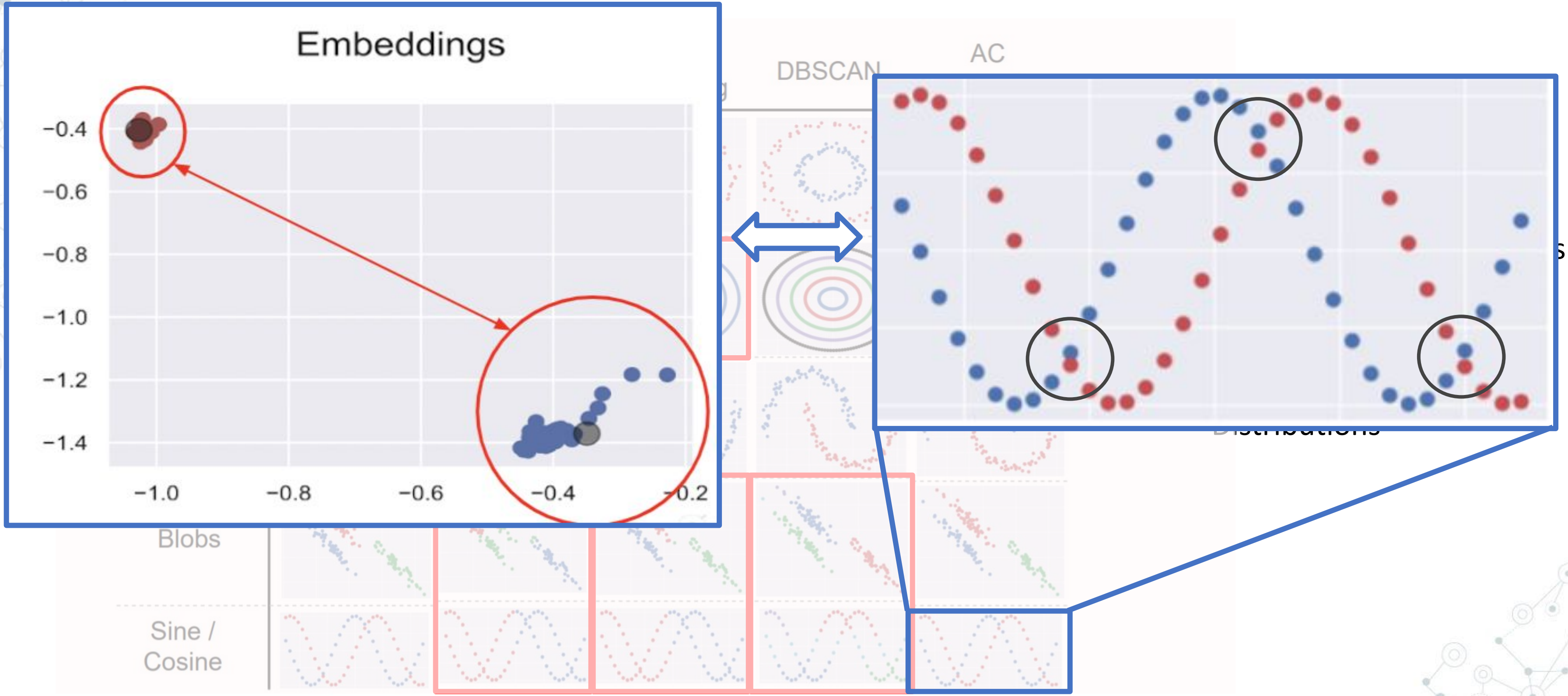
Results | Clustering with AC-Net



Scenarios:

- Number of clusters/groups
- Shape
- Ambiguity
- Distributions

Results | Clustering with AC-Net



Results | Intrusion Detection

Binary classification
(Benchmark: ISCX-IDS 2012)

- **FAR:** False Alarm Rate

- *Classifier:* Random Forests
- Encoding dimension: 10
- Payload features: 50

Approach	Payload-based Features	Accuracy (%)	FAR (%)	F_1 Score (%)
DAGMM	No	62.91	30.65	53.07
N-BaIoT	No	89.19	10.80	89.19
Deep NN	No	88.14	7.41	70.35
TR-IDS	Yes	98.88	1.12	98.87
ACID (ours)	No	99.78	0.23	99.44
ACID (ours)	**Yes**	**100.0**	**0.00**	**100.0**

Comparison of ACID with existing methods

10 Results | Intrusion Detection

Multi-label classification (ACID)

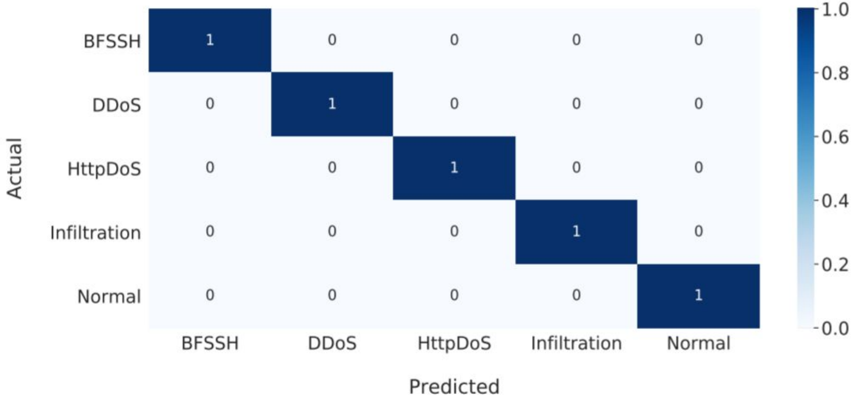
Metric Dataset	Accuracy (%)	FAR (%)	F_1 (%)	Classes	Samples
KDD CUP'99	100.0	0.00	100.0	23	43,510
ISCX-IDS 2012	100.0	0.00	100.0	5	10,547
CSE-CIC-IDS 2018	100.0	0.00	100.0	15	144,772

Properties

Datasets:

- Time span: *20 years*
- Number of attack types: *40*
- Raw network traffic traces
- Train/Test split: 70/30
- Payload features: Yes
- Test set ≅ 0.2 Billion packets

- *Classifier:* Random Forests
- Encoding dimension: 10
- Payload features: 50



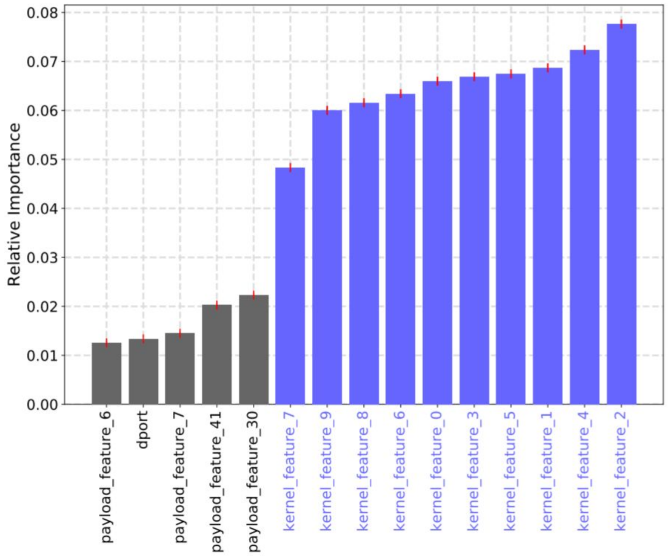
Normalized confusion matrix for multi-label classification using ACID on the ISCX-IDS 2012 dataset.

11 Impact factors | ISCX-IDS 2012

- *Classifier:* Random Forests
- Encoding dimension: 10
- Payload features: 50

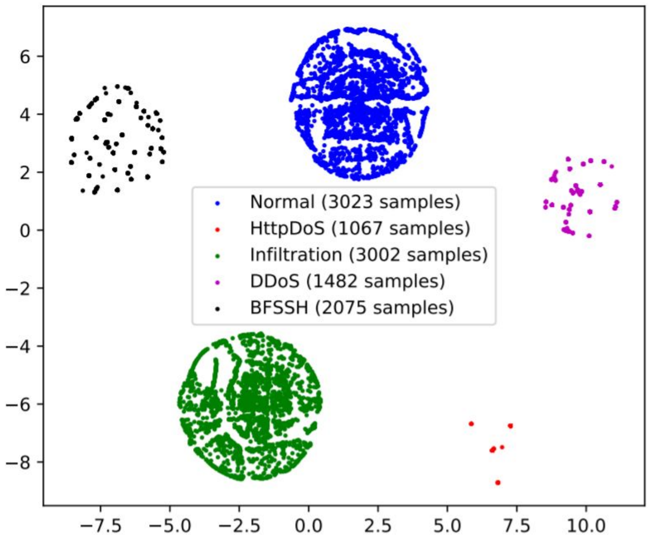
Feature ranking:

15 most important features in the classification process



t-SNE (from AC-Net's Embeddings)

t-SNE: A tool used to simplify the visual exploration of high-dimensional data points



Complexity Analysis

Speed Analysis (per packets)

Payload features?	*Duration*
No	0.78 us
Yes	145 us

Environmental setup

- 1 Virtual Machine
- 4 CPU cores @ 1.1GHz
- 4 GB RAM
- 50 GB Storage

✔ ***Deployable on constrained devices***

> 100x speed up

Payload Features	Number of Parameters	Batch size	Model Complexity (MFLOP)	Execution Time (seconds)
No	789,855	1	1.49	0.08 ± 0.01
		128	191.68	0.10 ± 0.02
Yes	942,460	1	25.71	0.19 ± 0.04
		128	3291.43	18.59 ± 0.74

Protecting ML classifiers against adversarial attacks

Adversarial Attacks

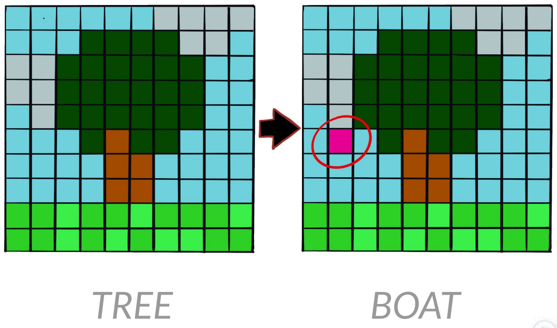
- *Definition:* Way of applying *subtle* perturbations to the inputs of a machine learning model, causing it to malfunction or produce a deceitful output.

- *Adversarial Sample:* $f_\theta : \mathbf{X} \to \mathbf{Y},$

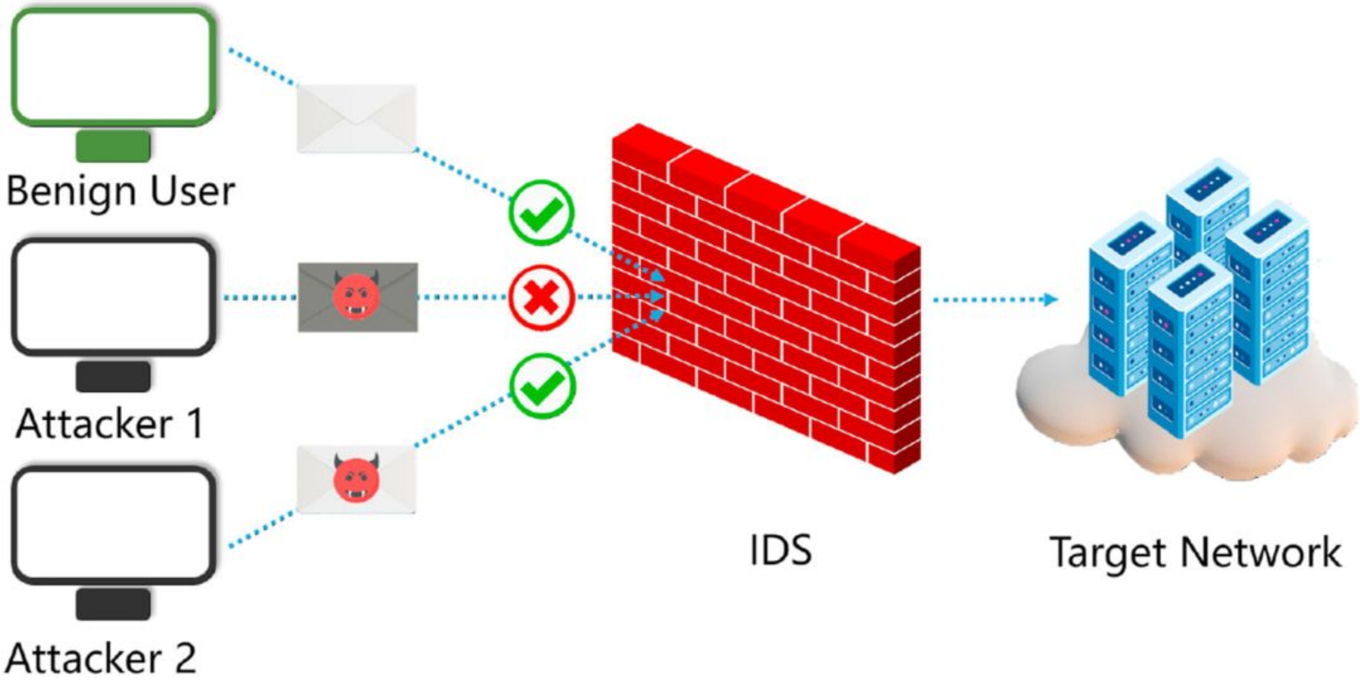
$$\mathbf{x}' = \mathbf{x} + \eta, f(\mathbf{x}) = \mathbf{y}, \mathbf{x} \in \mathbf{X},$$

$$f(\mathbf{x}') \neq \mathbf{y},$$

$$\text{or } f(\mathbf{x}') = \mathbf{y}', \mathbf{y}' \neq \mathbf{y},$$



15 Adversarial Attacks

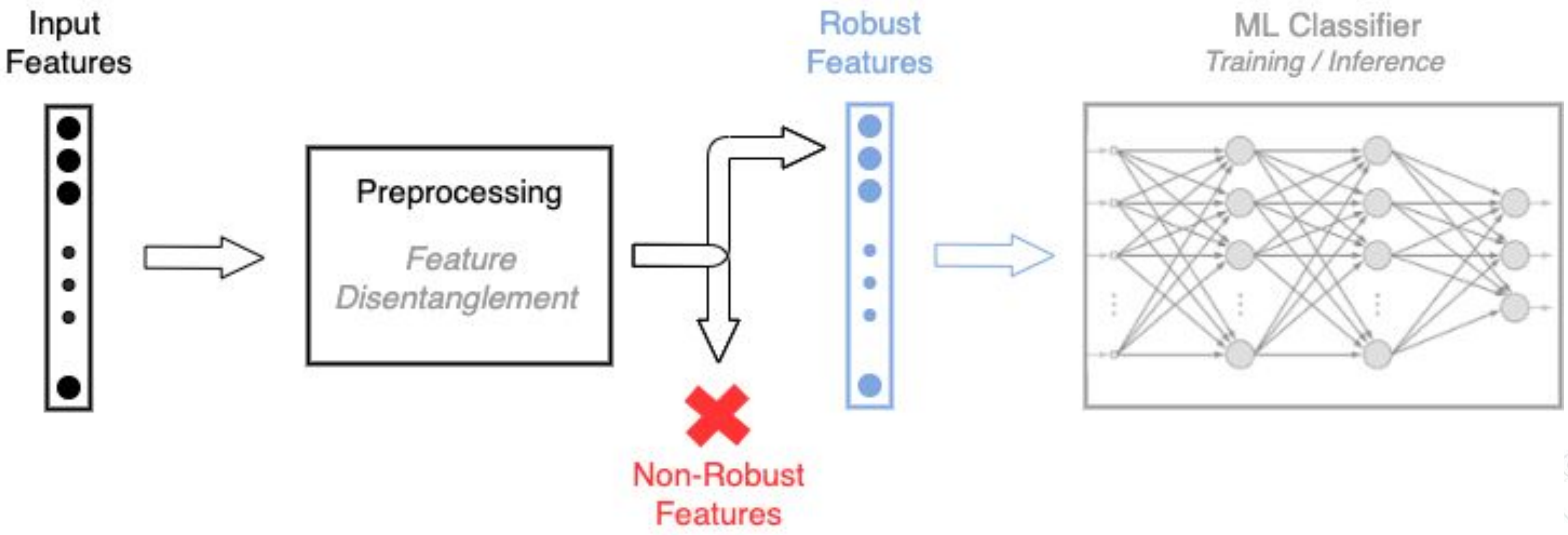


Desired Properties

- Retention of performance on non-adversarial samples
- Similar performance on adversarial samples
- Generalization: *known* + *unknown* attacks (Attack-agnostic)
- Preprocessing-based approach
  i.e., Can be combined with other defenses
- Low computational overhead

17 Methodology

<u>Goal:</u> Improve robustness by only relying *"robust/useful"* features from inputs.



Preprocessing

Robust Features Extraction

Preliminary Results

CIFAR-10: Top-1 accuracy under different attacks (perturbation bound ε= 8/255)

Attacks / Defenses	No Attack	FGSM	PGD 20 steps	PGD 100 steps
No Defense	**94.21**	14.64	0.00	0.00
AT (Madry et al., 2018)	87.30	56.1	45.80	45.05
TRADES (Zhang et al., 2019)	84.92	60.87	55.38	55.13
ME-NET (Yang et al., 2019) 0.4-0.6	84.00	71.39	57.50	53.50
Ours	92.61	**94.93**	**94.92**	**94.92**

Thank you!

Read more ?

Alec F. Diallo, Paul Patras. *"Adaptive Clustering-based Malicious Traffic Classification at the Network Edge"* **- IEEE INFOCOM 2021.**

- PhD supported by arm

Source code available at:
github.com/Mobile-Intelligence-Lab/ACID

Contact: alec.frenn@ed.ac.uk