Classify Me Correctly if You Can: Evaluating Adversarial Machine Learning Threats in NIDS

Neea Rusch

Augusta University, United States

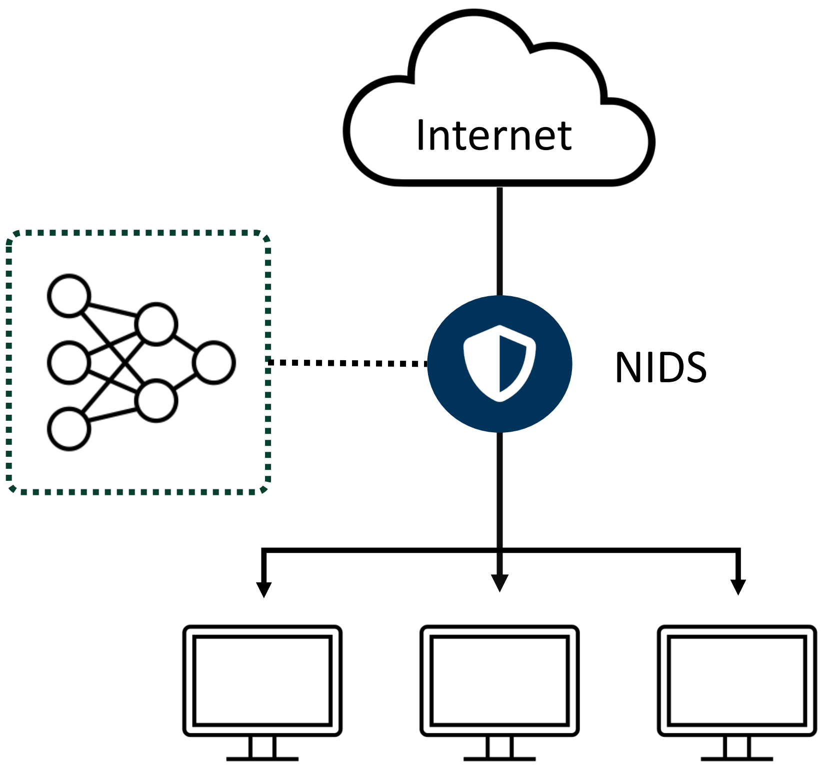
SecureComm 2023 • 20 October 2023

j.w.w. Asma Jodeiri Akbarfam, Hoda Maleki, Gagan Agrawal and Gokila Dorai

Network Intrusion Detection Systems (NIDS) detect and protect against network attacks.

- Defend against different network attacks
- Deployed in various kinds of networks

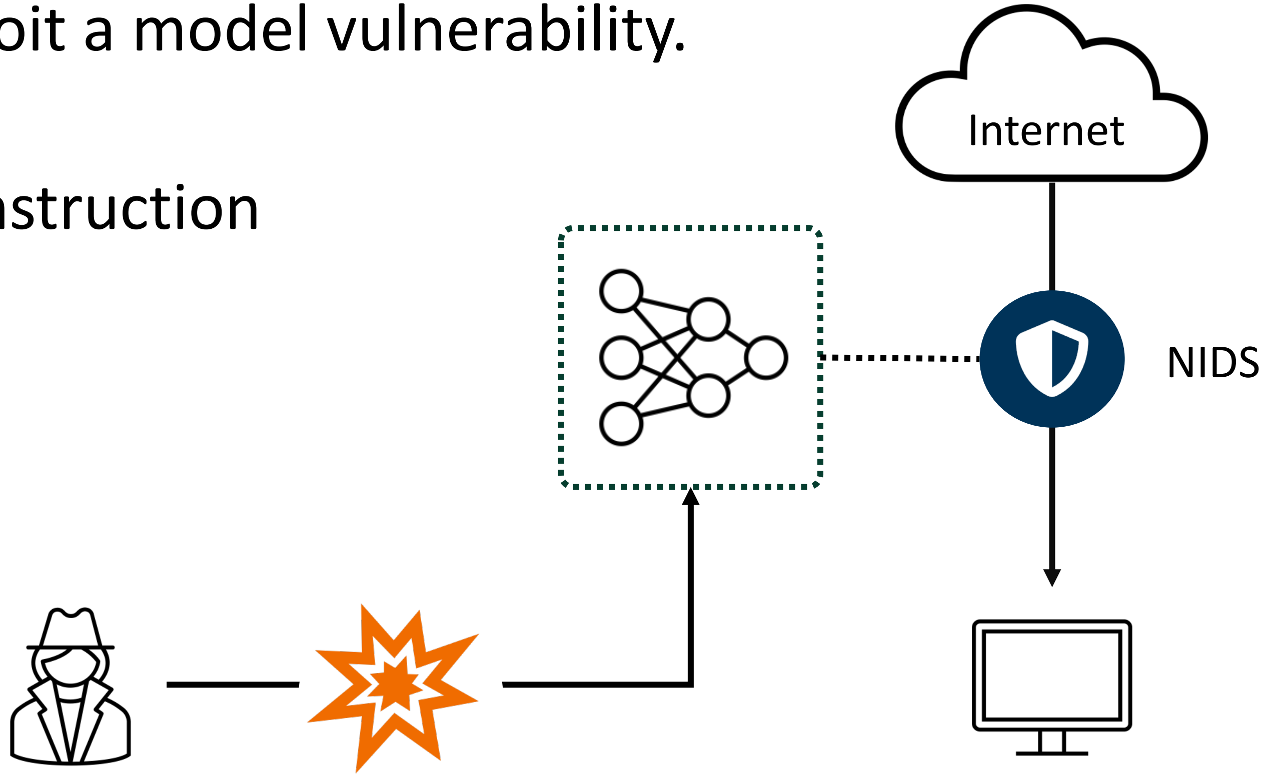
Modern NIDS use **machine learning**.



Problem: machine learning models are susceptible to adversarial attacks.

In **Adversarial Machine Learning** (AML)
adversary attempts to exploit a model vulnerability.

- obtain information of construction
- alter behavior



Adversarial Strategies

Training-phase attacks

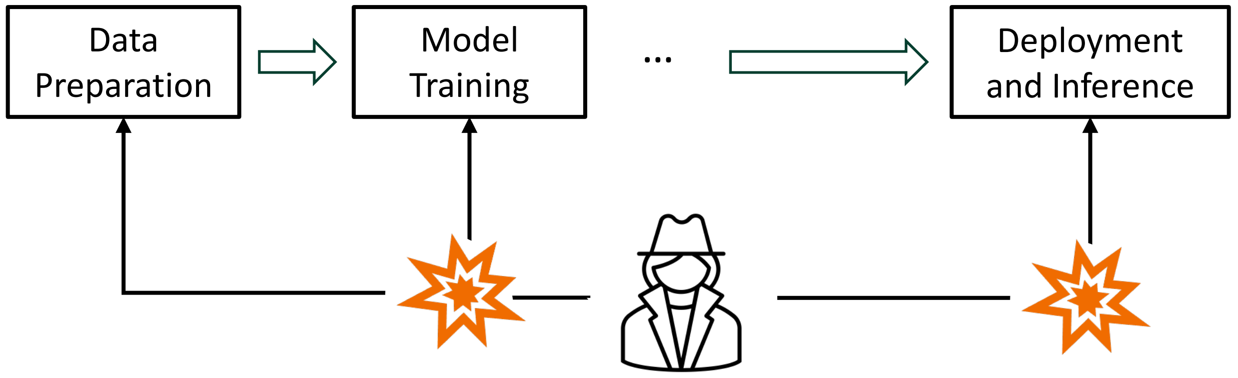
- Contaminate or alter data
- Cause learning bias

Exploits on trained models

- Alter inputs to avoid detection
- Attempt to recover the model

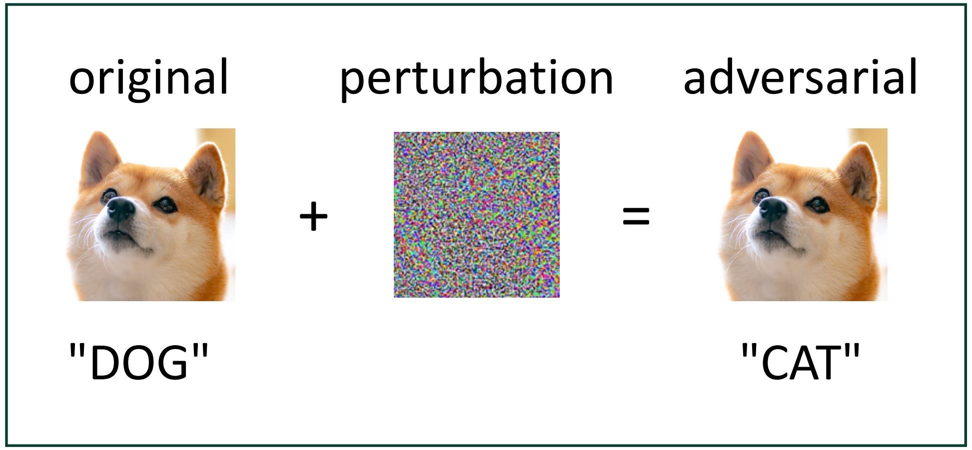
Defenses

- Numerous mechanisms
- Applied at different model deployment stages



Evaluating AML Threats in NIDS

Adversarial machine learning techniques have been studied primarily in **unconstrained** domains.



Network intrusion detection models are trained on network data, with correlation and **constraints** between attributes.

A **constrained domain** adds many new considerations

Acceptable perturbations are restricted.

Traditional evaluation metrics are inapplicable.

Misclassification is class sensitive.

Model invocations must be limited.

$\Sigma f_i = 1$

2^n

≥ 0

TCP	UDP	OTHER	ORIG_PKTS	BYTES	CONN_S0	CONN_SF	RESP_PKTS	DURATION	LABEL
1	0	0	1	32	1	0	0	0.0722	benign
0	1	0	1	128	0	1	1	0.0047	malicious

High-level Motivation

Take the state-of-the-art unconstrained AML attacks and defenses

↓

Adapt to constrained domains

↓

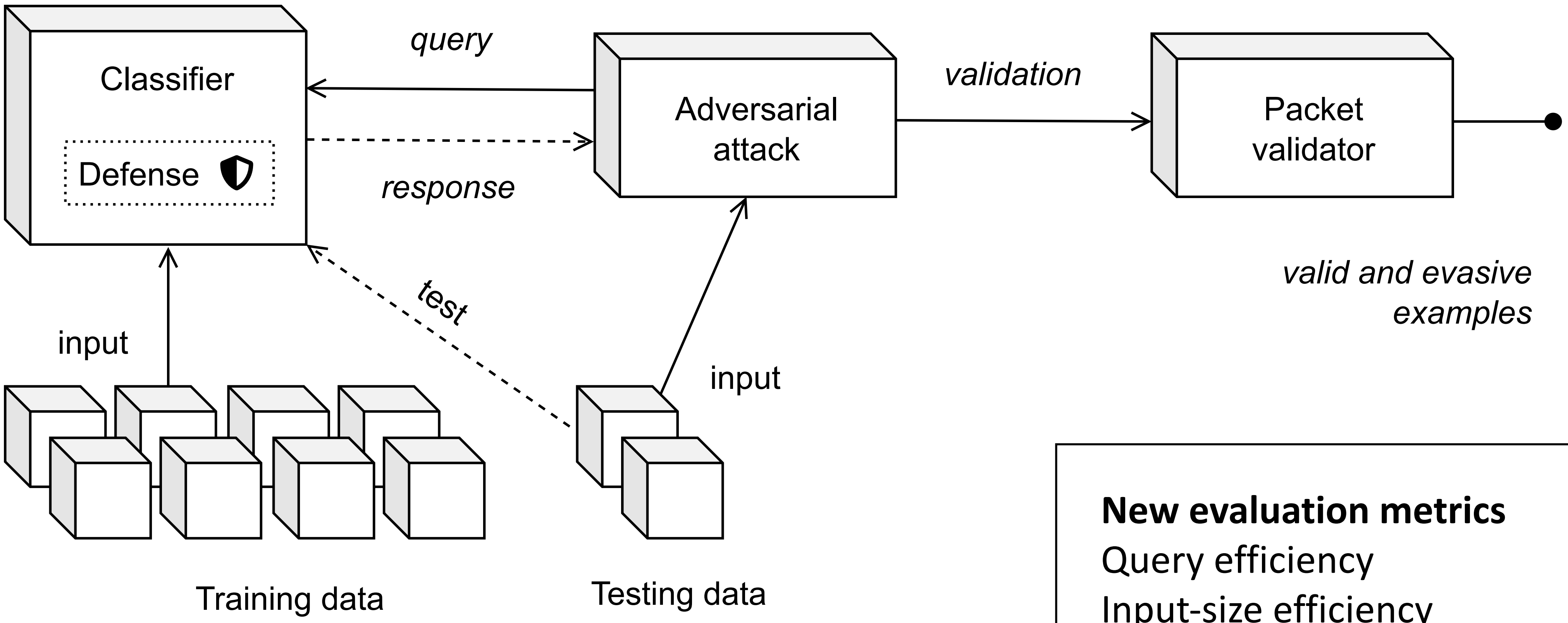
Measure impact of attacks and defenses in NIDS

Concrete approach

1) Design an **evaluation system** —
   includes choice input data, classifier, defense, and attack.

2) Capture domain constraints as **rules** —
   adversarially generated record must satisfy all applicable rules.

3) Add to the evaluation system a post-hoc packet **validator** —
   identifies adversarial examples that satisfy domain constraints.



New evaluation metrics

Query efficiency

Input-size efficiency

Class-label distribution

Evasion success rate

Experimental evaluation

The implementation enabled to evaluate classifiers, attacks, and defenses.
By varying different parameters, we can study their impact on NIDS security.

Data sets	2×	IoT-23, UNSW-NB15
Classifiers	2×	XGBoost, Deep Neural Network
Defenses	2×	Robust Trees, Adversarial Training
Attacks	2×	HopSkipJump Attack, Zeroth Order Optimization
Validator	1×	Validates TCP, UDP and other traffic flows

github.com/aucad/aml-networks

Limited model queries

Adversarial attack success rate for 48 attack configurations, as *fractions*.

"Valid" represents the fraction of evasive records that also pass validation.

Model/	**HopSkipJumpAttack**						**Zeroth Order Optimization**					
	Evasions			Valid			Evasions			Valid		
Iterations	2	5	10	2	5	10	2	5	80	2	5	80
IoT-23												
DNN	.34	.27	.31	0	0	.01	0	0	0	0	0	0
DNN-🛡	0	0	0	0	0	0	0	0	0	0	0	0
XGB	.43	.39	.41	.06	.07	.18	.47	.49	.49	.05	.05	.04
XGB-🛡	.38	.38	.38	.01	.01	.03	.03	.07	.07	.03	.06	.07
UNSW-NB15												
DNN	.79	.68	.81	.41	.39	.42	.28	.36	.29	.25	.30	.24
DNN-🛡	.02	.11	.07	.02	.11	.07	0	0	0	0	0	0
XGB	.93	.92	.91	.47	.46	.47	.50	.69	.78	.49	.65	.69
XGB-🛡	.64	.65	.65	.38	.38	.38	.09	.31	.32	.09	.30	.31

Limited model queries

Adversarial success rate by transmission protocol on UNSW-NB15 data.

Benign—Malicious column shows class-label distribution of evasive and valid records.

Model/ Protocol	**Evasions** TCP	UDP	other	**Valid** TCP	UDP	other	**Benign– Malicious**
HopSkipJumpAttack							
DNN	.79	.85	.81	.78	.02	.03	27–73
DNN-🛡	.14	0	0	.14	0	0	0–100
XGB	.91	.94	.88	.89	.02	.01	30–70
XGB-🛡	.75	.43	.78	.73	0	0	17–83
Zeroth Order Optimization							
DNN	.35	.23	.22	.34	.13	.14	52–48
DNN-🛡	0	0	0	0	0	0	–
XGB	.89	.70	.55	.88	.50	.43	34–66
XGB-🛡	.54	.11	.01	.53	.11	.01	24–76

Summary

An evaluation system with a post-hoc constraint validator — added constrains to unconstrained state-of-the-art attacks.

Experimentally measured attacks and defenses — despite constraints, AML attacks pose challenges to NIDS.

Many possible future directions —
e.g., performing validation during an adversarial search and using the validator feedback to improve attack success.