IEEE 18° International Symposium on Network Computing and Applications
September 27, 2019
Cambridge, MA, USA

Evaluating the Effectiveness of Adversarial Attacks against Botnet Detectors

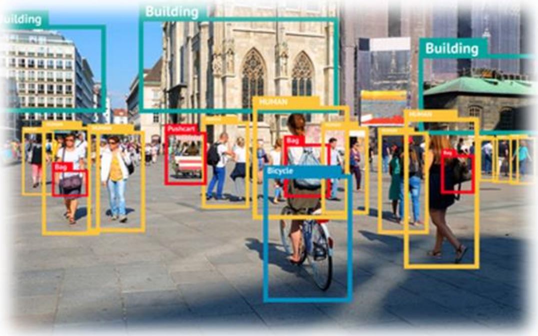
Giovanni Apruzzese
PhD Candidate in Information and Communication Technologies
University of Modena and Reggio Emilia
giovanni.apruzzese@unimore.it

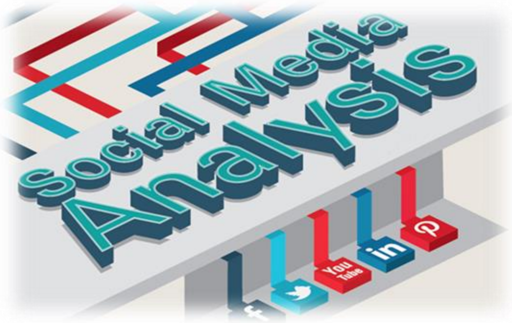
Scuola di Ingegneria
Dipartimento di Ingegneria "Enzo Ferrari"

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Machine Learning in the Real World

The popularity of Machine Learning is skyrocketing.





Machine Learning algorithms are effective, but what about **CyberSecurity**?

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Machine Learning & CyberSecurity at a glance…

FORTINET

FortiGuard Artificial Intelligence (AI) Delivers Proactive Threat Detection at Machine Speed and Scale

JUNIPER NETWORKS

Machine learning moves to the front lines of defense against an expanding threat surface.

Machine Learning: New Frontiers in Advanced Threat Detection

Symantec

SOPHOS

Sophos Adds Advanced Machine Learning to Its Next-Generation Endpoint Protection Portfolio

MACHINE LEARNING HELPS US FIND NEW ATTACKS

F-Secure

KASPERSKY lab

Machine learning in Kaspersky Endpoint Security 10 for Windows

TREND MICRO

The truth is Trend Micro has been using machine learning since 2005.

MACHINE LEARNING PREVENTS PRIVILEGE ATTACKS AT THE ENDPOINT

CYBERARK

FireEye

McAfee is evolving its machine learning cybersecurity technology

McAfee

Rapid7 Attacker Behavior Analytics Brings Together Machine Learning and Human Security Expertise

RAPID7

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

…but all that shines is not gold!

Main issues of ML for CyberSecurity:

Model training & selection

- Where and how to find high quality and labeled training dataset?
- How to compare different ML approaches

Evolution over time (concept drift)

- How frequently should the model be re-trained?

False positives and false negatives

- 1% false positive rate in large organization = **thousands** of daily false alarms

Vulnerability to Adversarial Attacks

- How effective are adversarial attacks against Cyber Detectors based on machine learning?

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Adversarial Attacks against Machine Learning

Adversarial Attacks involve the creation of specific samples with the goal of thwarting the Machine Learning algorithm.

Even **tiny perturbations** can **greatly affect** the prediction performance

- Rich research area within the image processing field…
- …but comprehensive analyses from a **CyberSecurity** perspective are scarce (especially in the context of *Network Intrusion Detection*)

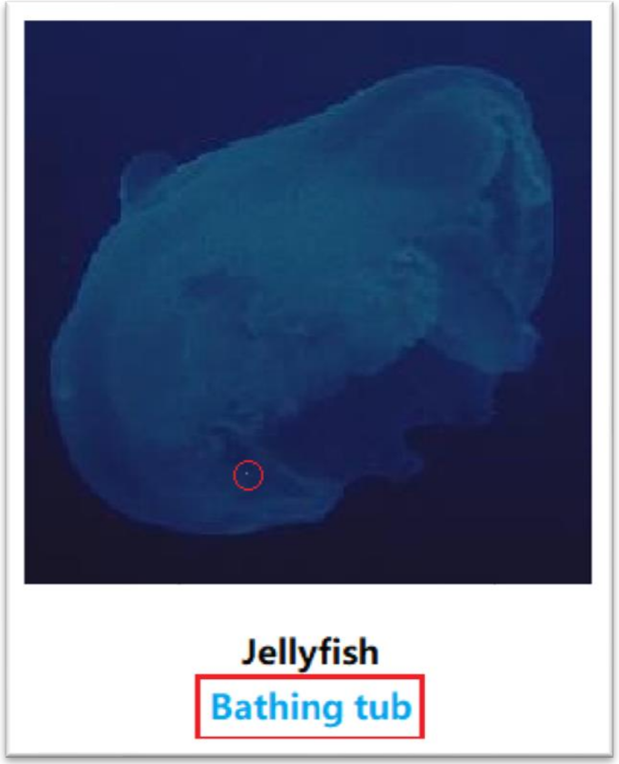


Image Reference: Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "*One pixel attack for fooling deep neural networks*." IEEE Transactions on Evolutionary Computation (2019).

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Focus, Motivation and Contribution

- Past literature has shown that Botnet Detectors can be easily ($Recall < 10\%$) evaded by slightly altered (adversarial) malicious samples.

- We expand these research efforts with an **extensive experimental campaign** providing the following three-fold contribution:

More Algorithms (12)	• Past work has only focused on <u>small subsets</u> of ML algorithms
More Datasets (4)	• Past work is based on just <u>one dataset</u>
Defence Evaluation (*feature removal*)	• <u>Lack of evaluations</u> of defensive approaches

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Datasets and Algorithms

We consider 4 public datasets of labelled network flows containing botnet-specific traffic

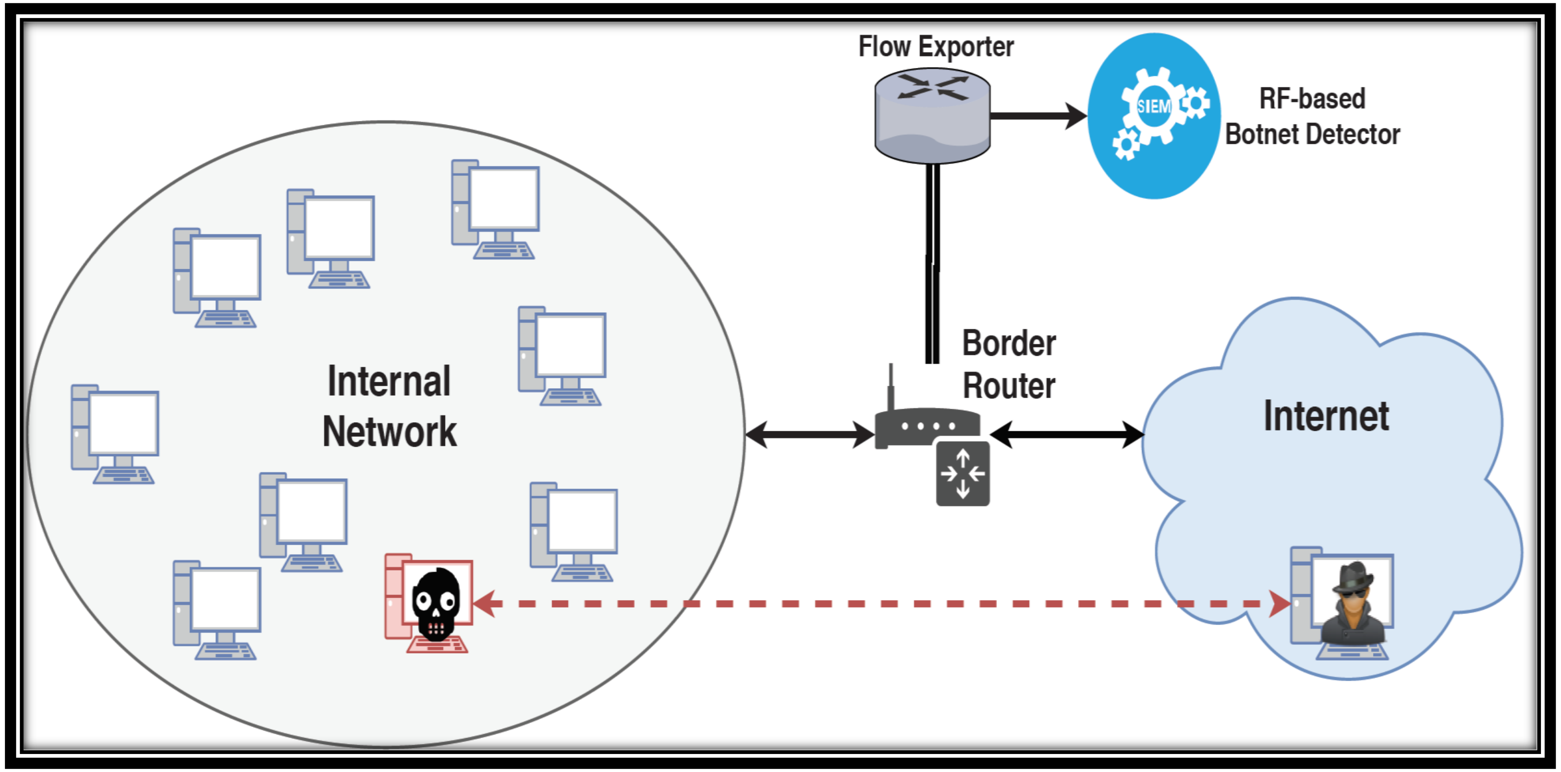
Dataset	Packets	Devices	Botnet Flows	Legitimate Flows	Botnet Families
CTU-13	855 866 143	150	443 906	19 199 170	6
IDS2017	5 776 888	111	1 966	189 067	1
CIC-IDS2018	13 486 990	450	283 429	760 824	1
UNB-CA Botnet	14 502 782	369	238 415	345 113	10

Each dataset is evaluated with the following 12 machine learning classifiers

Random Forest (RF)	Bagging (Bag)	Support Vector Machine (SVM)
Stochastic Gradient Descent (SGD)	Deep Neural Network (DNN)	Logistic Regression (LR)
Decision Tree (DT)	Naive Bayes (NB)	Gradient Boosting (GB)
AdaBoost (AB)	K-Nearest Neighbor (KNN)	Extra Trees (ET)

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Application Scenario



Attacker Model

- Goal: evade the botnet detector
- Knowledge: Limited
- Capabilities: Limited
- Strategy: alter the bot(s) communications

Realistic assumptions

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Experiments – outline

I. Develop botnet detectors <u>with good performance</u>
   - $(F1\text{-}score, Precision, Recall) > 90\%$

II. Generate **realistic** adversarial samples

III. Evaluate the detectors against the generated adversarial samples
   - Measured through the *Attack Severity* (AS): $AS = 1 - \frac{Recall\ (attack)}{Recall\ (no\ attack)}$

Higher AS = higher impact

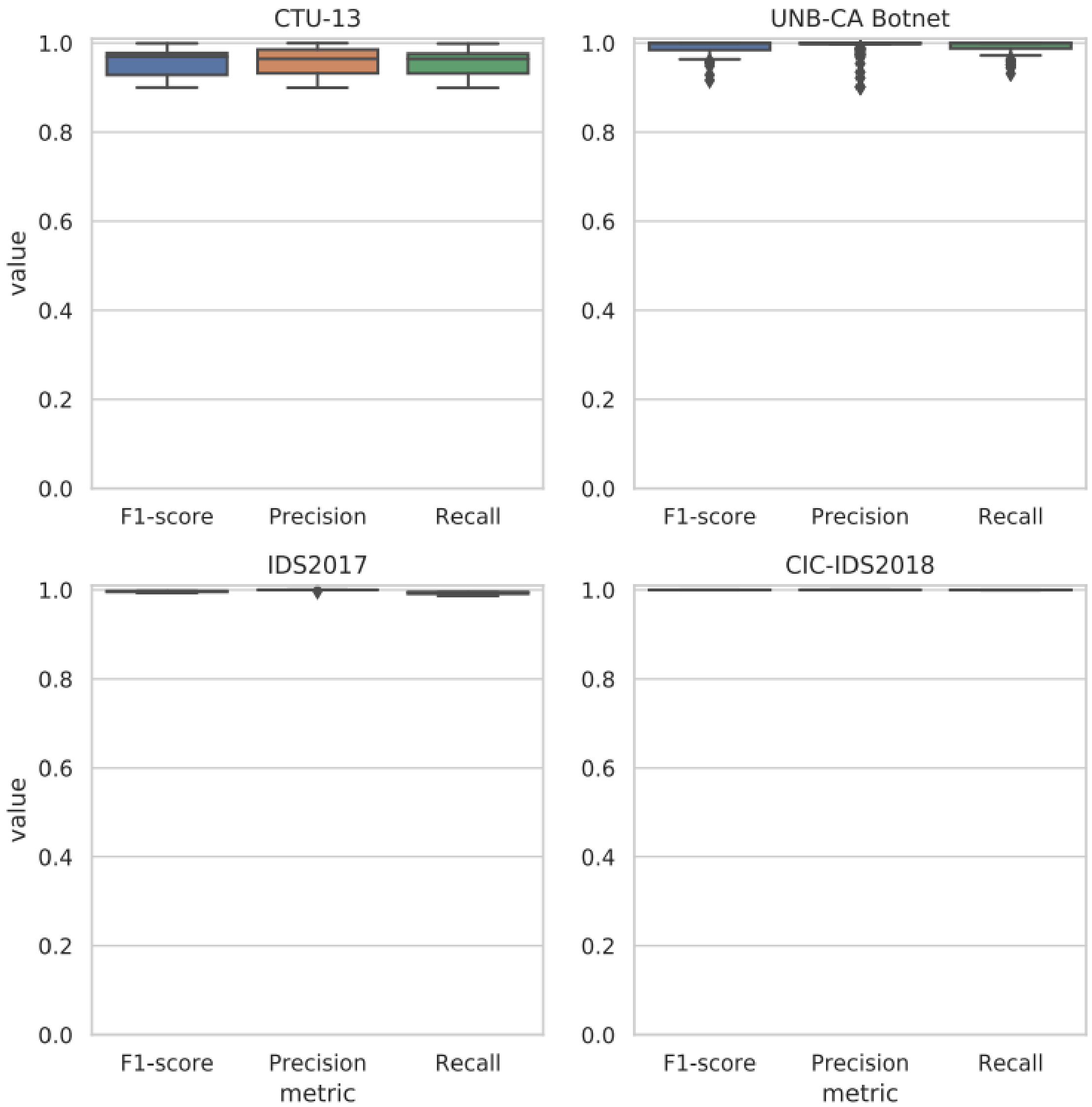
IV. Test the effectiveness of *feature removal* against these attacks
   - How much is the baseline performance affected?

V. Repeat this process for all considered datasets

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Experiments I – Baseline Performance Results

Dataset	F1-Score (std. dev.)	Precision (std. dev.)	Recall (std. dev.)
CTU-13	0.957 (0.029)	0.958 (0.031)	0.956 (0.028)
IDS2017	0.996 (0.002)	0.999 (0.001)	0.993 (0.003)
CIC-IDS2018	0.999 (< 0.001)	0.999 (< 0.001)	0.999 (< 0.001)
UNB-CA Botnet	0.991 (0.017)	0.992 (0.021)	0.991 (0.017)
Average	0.986 (0.011)	0.987 (0.012)	0.985 (0.011)



Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Experiments II – Generation of Realistic Adversarial Samples

Goal: generate adversarial samples through small and easily attainable modifications

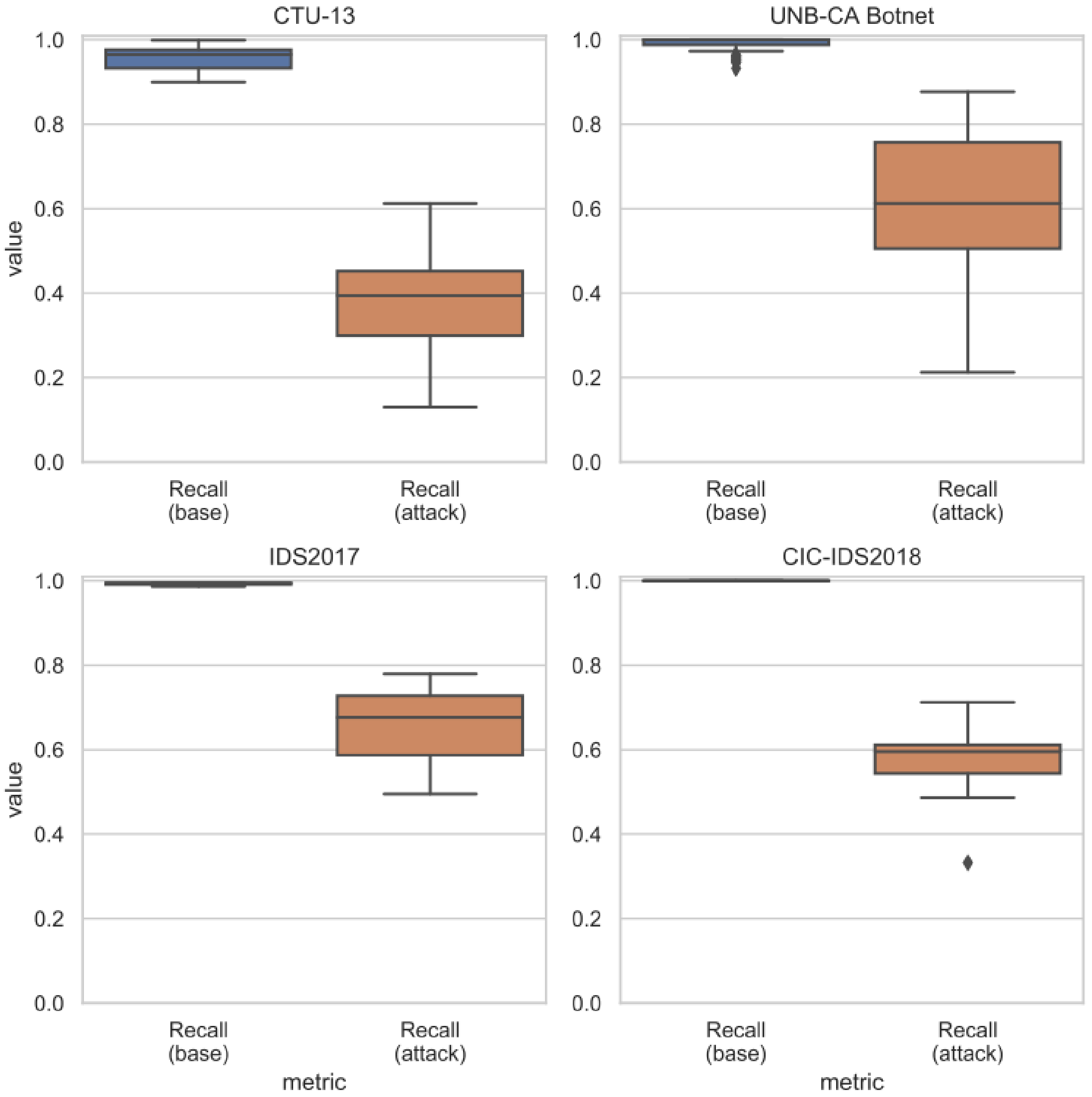
Group	Altered features
1a	Duration (s)
1b	Src_bytes
1c	Dst_bytes
1d	Tot_pkts
2a	Duration, Src_bytes
2b	Duration, Dst_bytes
2c	Duration, Tot_pkts
2e	Src_bytes, Tot_pkts
2d	Src_bytes, Dst_bytes
2f	Dst_bytes, Tot_pkts
3a	Duration, Src_bytes, Dst_bytes
3b	Duration, Src_bytes, Tot_pkts
3c	Duration, Dst_bytes, Tot_pkts
3d	Src_bytes, Dst_bytes, Tot_pkts
4a	Duration, Src_bytes, Dst_bytes, Tot_pkts

Step	Duration	Src_bytes	Dst_bytes	Tot_pkts
I	+1	+1	+1	+1
II	+2	+2	+2	+2
III	+5	+8	+8	+5
IV	+10	+16	+16	+10
V	+15	+64	+64	+15
VI	+30	+128	+128	+20
VII	+45	+256	+256	+30
VIII	+60	+512	+512	+50
IX	+120	+1024	+1024	+100

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Experiments III – Impact of the Adversarial Attacks

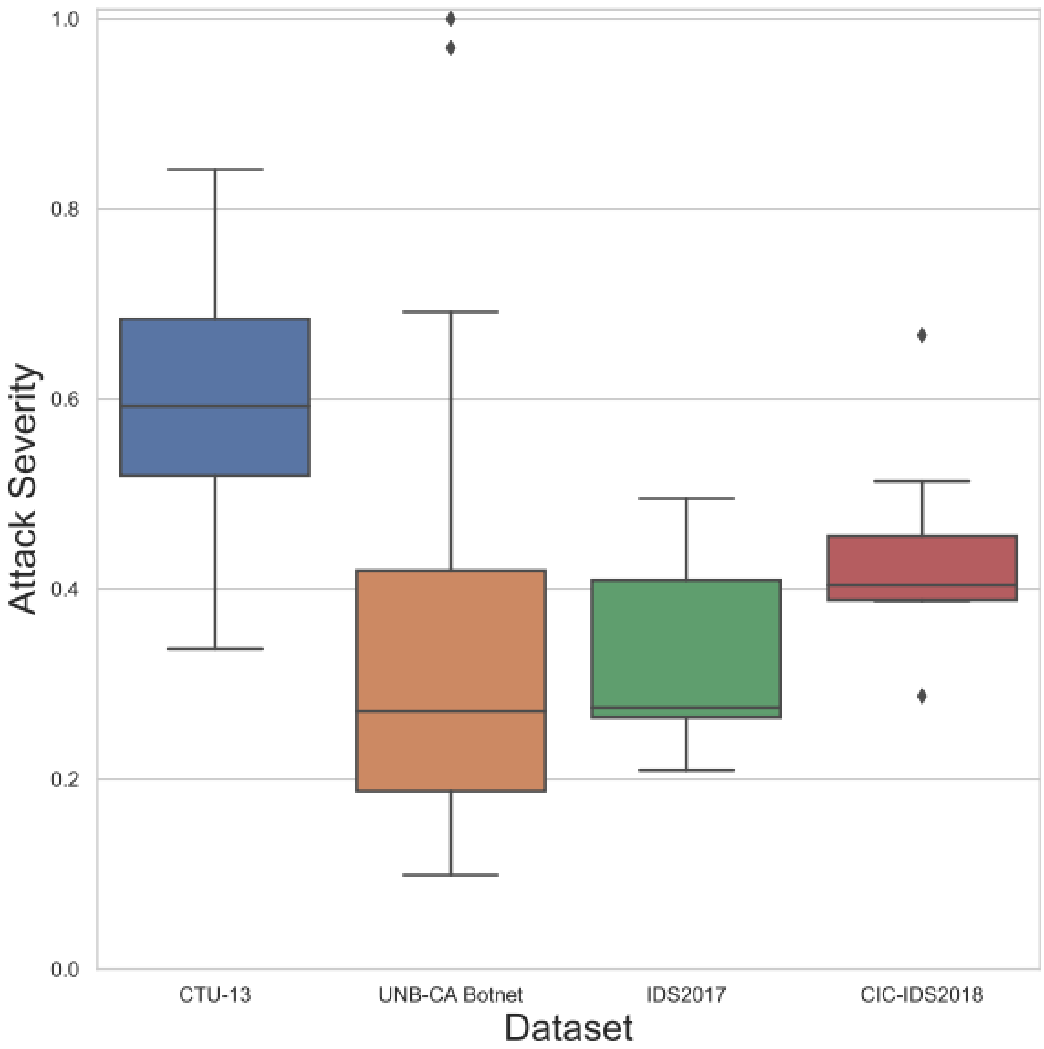
Dataset	Recall baseline (std. dev)	Recall adversarial (std. dev)	Attack Severity (std. dev)
CTU-13	0.956 (0.028)	0.372 (0.112)	0.609 (0.110)
IDS2017	0.993 (0.003)	0.656 (0.102)	0.327 (0.103)
CIC-IDS2018	0.999 (< 0.001)	0.564 (0.112)	0.436 (0.112)
UNB-CA Botnet	0.991 (0.017)	0.588 (0.218)	0.328 (0.212)
Average	0.985 (0.011)	0.545 (0.136)	0.425 (0.134)



Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Experiments III – Impact of the Adversarial Attacks

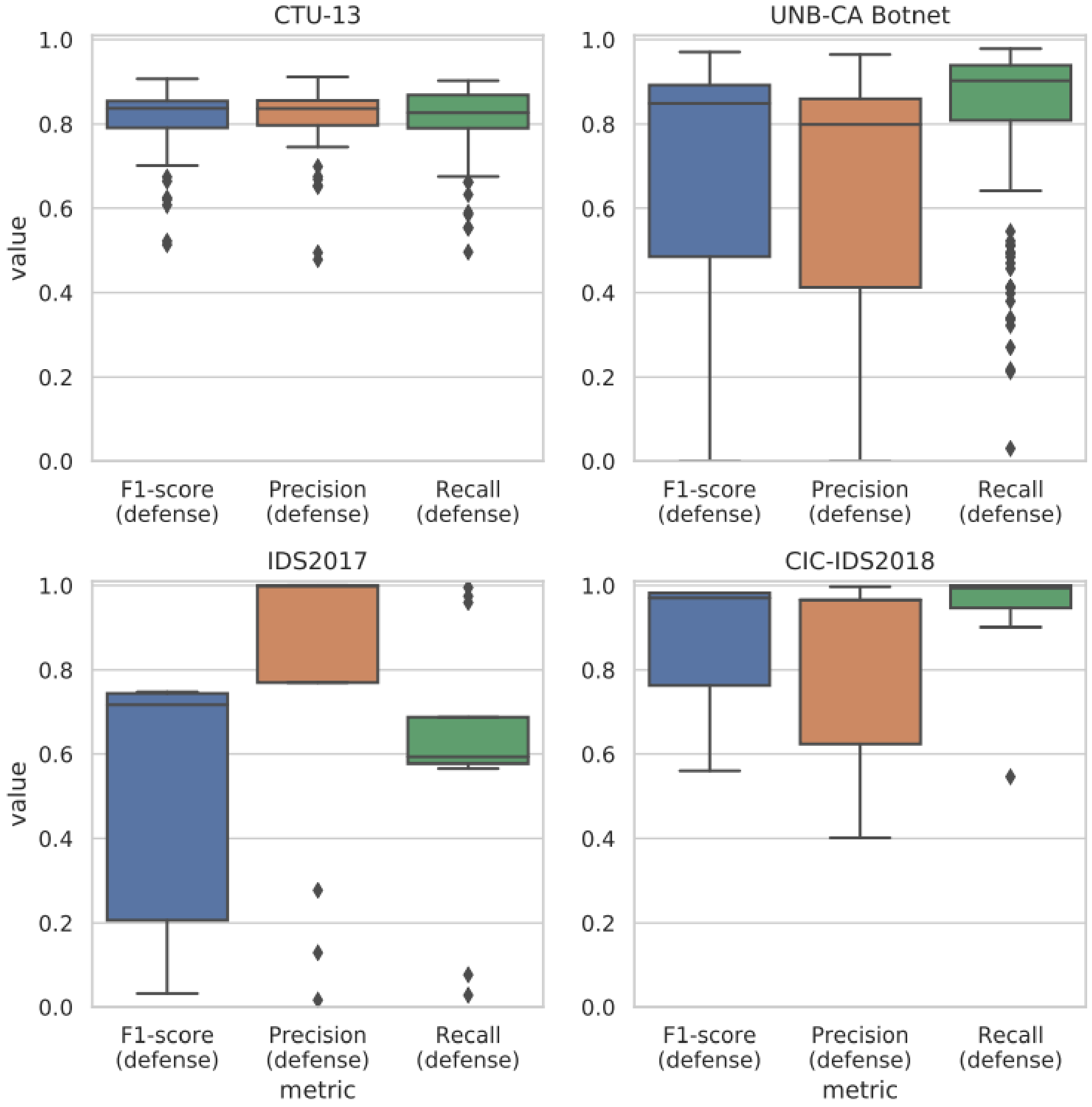
Dataset	**Recall baseline** (std. dev)	**Recall adversarial** (std. dev)	**Attack Severity** (std. dev)
CTU-13	0.956 (0.028)	0.372 (0.112)	0.609 (0.110)
IDS2017	0.993 (0.003)	0.656 (0.102)	0.327 (0.103)
CIC-IDS2018	0.999 (< 0.001)	0.564 (0.112)	0.436 (0.112)
UNB-CA Botnet	0.991 (0.017)	0.588 (0.218)	0.328 (0.212)
Average	0.985 (0.011)	0.545 (0.136)	0.425 (0.134)



Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Experiments IV – Countermeasure effectiveness

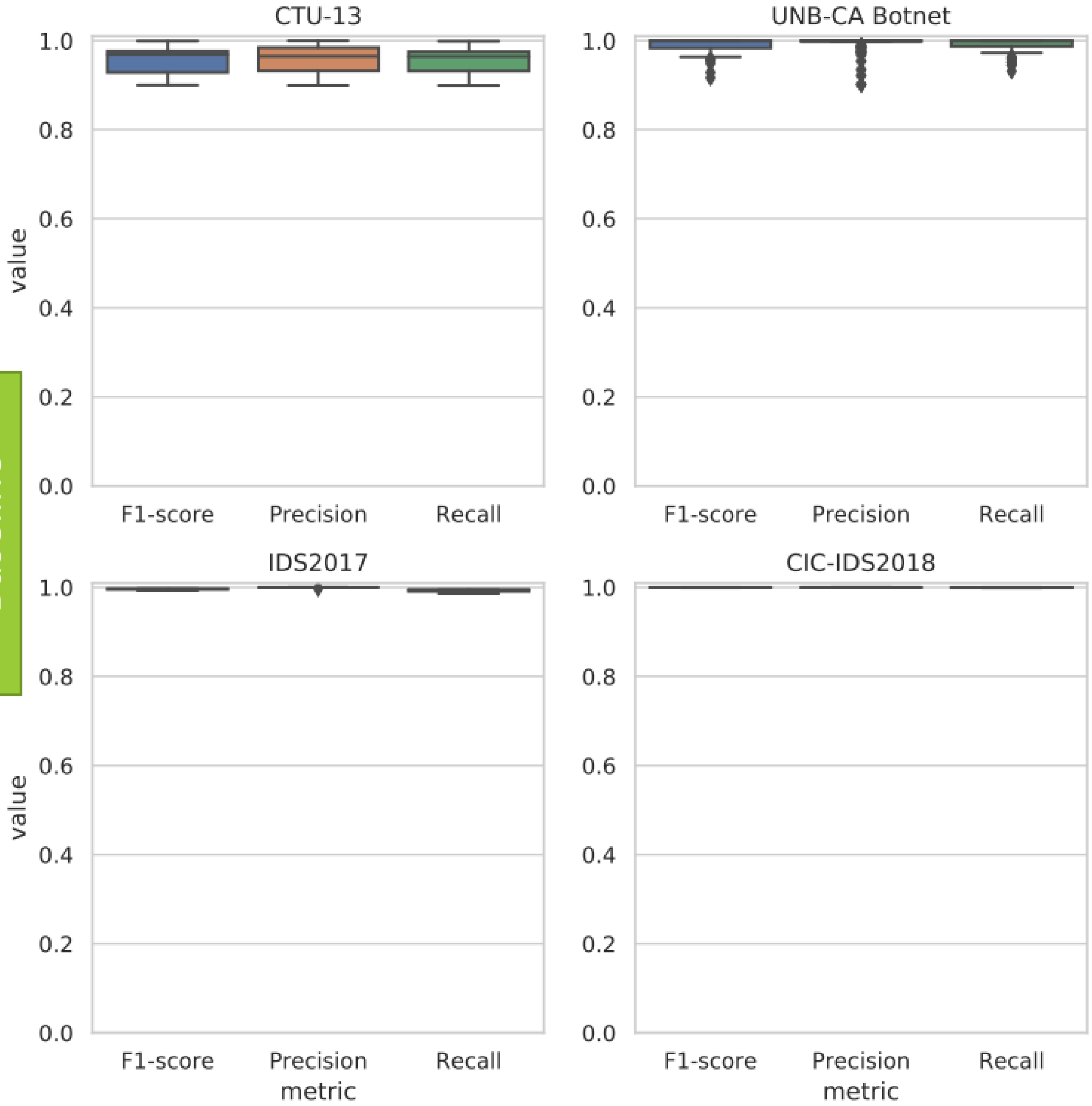
Dataset	F1-Score (std. dev.)	Precision (std. dev.)	Recall (std. dev.)
CTU-13	0.803 (0.092)	0.810 (0.089)	0.799 (0.101)
IDS2017	0.503 (0.304)	0.777 (0.388)	0.596 (0.306)
CIC-IDS2018	0.859 (0.164)	0.814 (0.212)	0.942 (0.128)
UNB-CA Botnet	0.691 (0.276)	0.645 (0.285)	0.808 (0.209)
Average	0.714 (0.209)	0.761 (0.2235)	0.786 (0.186)



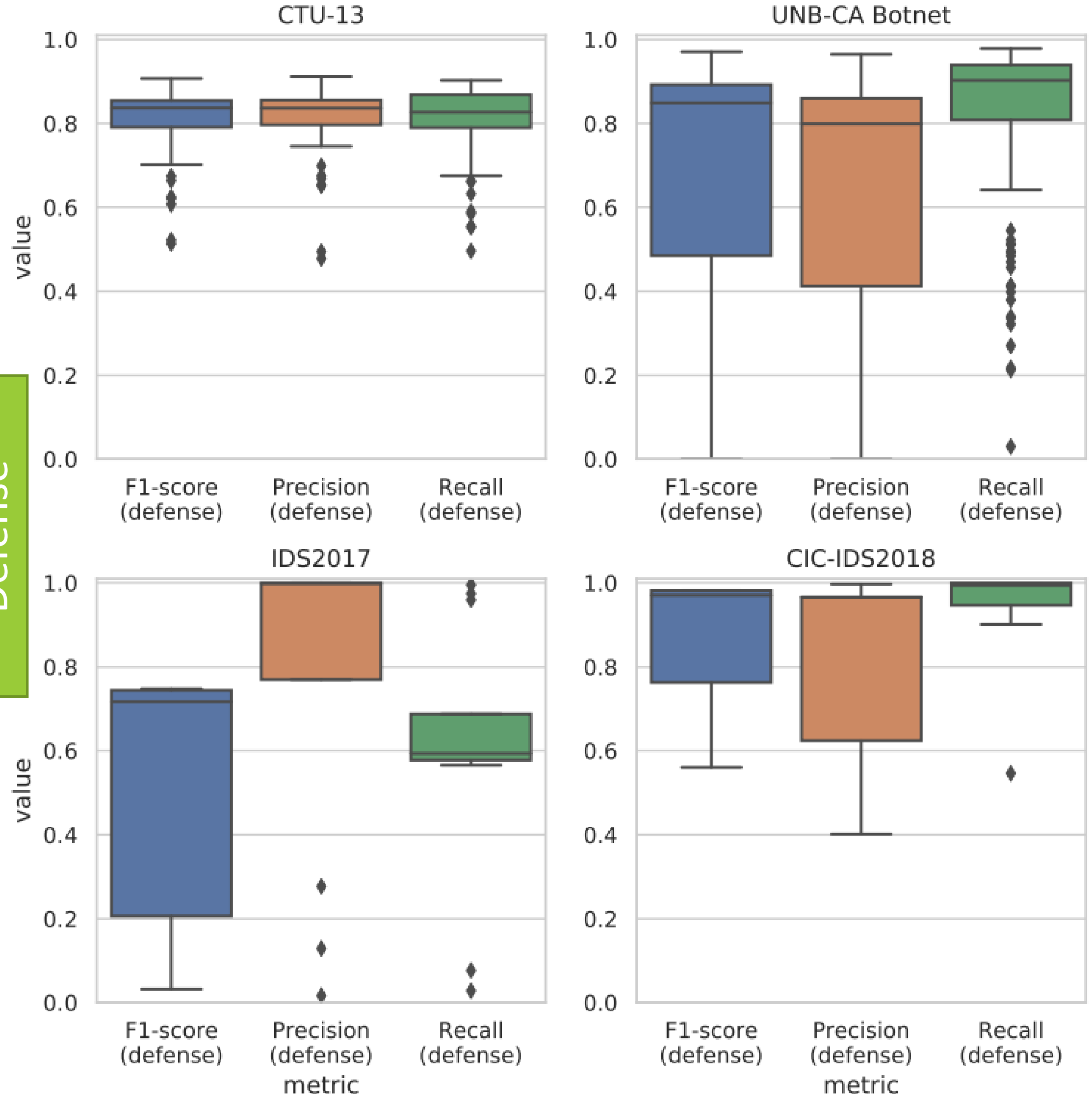
Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Experiments IV – Countermeasure effectiveness

Baseline



Defense



Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Performance of the top5 algorithms for each dataset

CTU-13

	Baseline			Attack		Defense		
Algorithm	*F1-score*	*Precision*	*Recall*	*Recall*	*Attack Severity*	*F1-score*	*Precision*	*Recall*
RF	0.9694	0.9722	0.9668	0.4390	0.5461	0.8564	0.8498	0.8641
AB	0.9722	0.9748	0.9696	0.4074	0.5803	0.8446	0.8487	0.8410
MLP	0.9458	0.9454	0.9462	0.3141	0.7261	0.7235	0.7734	0.6886
KNN	0.9296	0.9273	0.9320	0.2982	0.6806	0.6992	0.7265	0.6767
Bag	0.9745	0.9799	0.9693	0.4007	0.5869	0.8477	0.8516	0.8442

IDS2017

	Baseline			Attack		Defense		
Algorithm	*F1-score*	*Precision*	*Recall*	*Recall*	*Attack Severity*	*F1-score*	*Precision*	*Recall*
AB	0.9972	1	0.9945	0.7455	0.2504	0.7172	0.9779	0.5663
MLP	0.9959	0.9972	0.9945	0.5991	0.3975	0.7169	0.9344	0.5816
KNN	0.9959	1	0.9918	0.5512	0.4442	0.4292	0.2764	0.9591
ET	0.9972	1	0.9945	0.7333	0.2626	0.7456	1	0.5943
GB	0.9945	1	0.9891	0.7221	0.2699	0.7476	1	0.5967

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Performance of the top5 algorithms for each dataset

CIC-IDS2018

	Baseline			Attack		Defense		
Algorithm	*F1-score*	*Precision*	*Recall*	*Recall*	*Attack Severity*	*F1-score*	*Precision*	*Recall*
RF	0.9999	0.9999	0.9999	0.5965	0.4034	0.9822	0.9653	0.9996
AB	0.9997	0.9999	0.9996	0.5632	0.4365	0.9709	0.9969	0.9463
MLP	0.9997	0.9999	0.9995	0.7123	0.2873	0.9696	0.9939	0.9465
KNN	0.9998	0.9999	0.9998	0.4866	0.5132	0.8225	0.7564	0.9012
ET	0.9999	0.9999	0.9999	0.6023	0.3976	0.9822	0.9653	0.9996

UNB-CA Botnet

	Baseline			Attack		Defense		
Algorithm	*F1-score*	*Precision*	*Recall*	*Recall*	*Attack Severity*	*F1-score*	*Precision*	*Recall*
RF	0.9974	0.9997	0.9951	0.6856	0.3110	0.8912	0.8584	0.9283
KNN	0.9496	0.9479	0.9516	0.6167	0.3507	0.8144	0.7555	0.8871
ET	0.9993	0.9999	0.9987	0.6831	0.3160	0.8897	0.8544	0.9294
MLP	0.9215	0.9113	0.9321	0.5978	0.2756	0.7393	0.6779	0.8325
AB	0.9955	0.9971	0.9939	0.6840	0.3118	0.8926	0.8595	0.9303

Giovanni Apruzzese
giovanni.apruzzese@unimore.it

Conclusion

- Machine Learning algorithms need to be evaluated against **adversarial attacks**, especially from a Cybersecurity perspective.

- We expose the fragility against *realistic* adversarial perturbations of botnet detectors:
  - based on **12** different ML algorithms;
  - evaluated on samples belonging to **4** different datasets.
- We show that *feature removal* defensive techniques are unfeasible in real-contexts.

TAKEAWAY: adversarial attacks represent a dangerous menace to ML security systems because they are: (i) highly effective; (ii) difficult to counter; (iii) easy to perform.

Our mission is to increase the awareness of this threat, so as to promote the development of appropriate countermeasures.

Evaluating the Effectiveness of Adversarial Attacks against Botnet Detectors

Giovanni Apruzzese
PhD Candidate in Information and Communication Technologies
University of Modena and Reggio Emilia
Department of Engineering "Enzo Ferrari"
giovanni.apruzzese@unimore.it
https://weblab.ing.unimo.it/people/apruzzese

Scuola di Ingegneria
Dipartimento di Ingegneria
"Enzo Ferrari"