OnRampII - TAMU Camps Focused Research

Data-Driven Network Analysis And Discovery Of System Vulnerabilities In Real Scenarios

Presented by
Shumon Alam
Research Scientist
SECURE Cybersecurity Center of Excellence
Prairie View A&M University, TX



Project Description

In the cybersecurity area, Machine Learning (ML) and Deep Learning (DL) algorithms have contributed to effective approaches for network intrusion detection, but these techniques also introduce new threat approaches such as adversarial machine learning, model extraction, data inference and pre-processing, and distributed learning. The detection performance depends on the training datasets and the use of relevant features. There are synthetic or simulation-based datasets available with various limitations. To overcome the limitations of existing datasets, a realistic IDS dataset will be developed for various use cases using Spirent's CyberFlood-CF20 emulator and other attack tools using complete network configurations.

Project Deliverable:

- **Raw traffic:** Emulate various attack scenarios and capture raw traffic
- **IDS dataset:** Create a labeled dataset from the captured emulated traffic
- **Performance Analysis:** Evaluate ML and DL algorithms for network intrusion detection using the developed IDS dataset

Available Datasets

DARPA (Lincoln Laboratory 1998-99)	KDD99/KDDCup99(University of California, Irvine 1998-99)	NSL-KDD (2009)	Gure-KddCup
CDX (United States Military Academy 2009)	Kyoto (Kyoto University 2009)	AWID2018	CICIDS2017
UNSW-NB15	ISCX IDS 2012	ASNM-CDX (2009)	Lawrence Berkeley National Laboratory (LBNL, 2004-05):
Twente (University of Twente in 2009)	CTU-13	CAIDA (Center of Applied Internet Data Analysis 2000-2016)	DEFCON (2000-2002)
UMASS (University of Massachusetts 2004-2018):	DDoS2019 (CIC)		

Limitations:

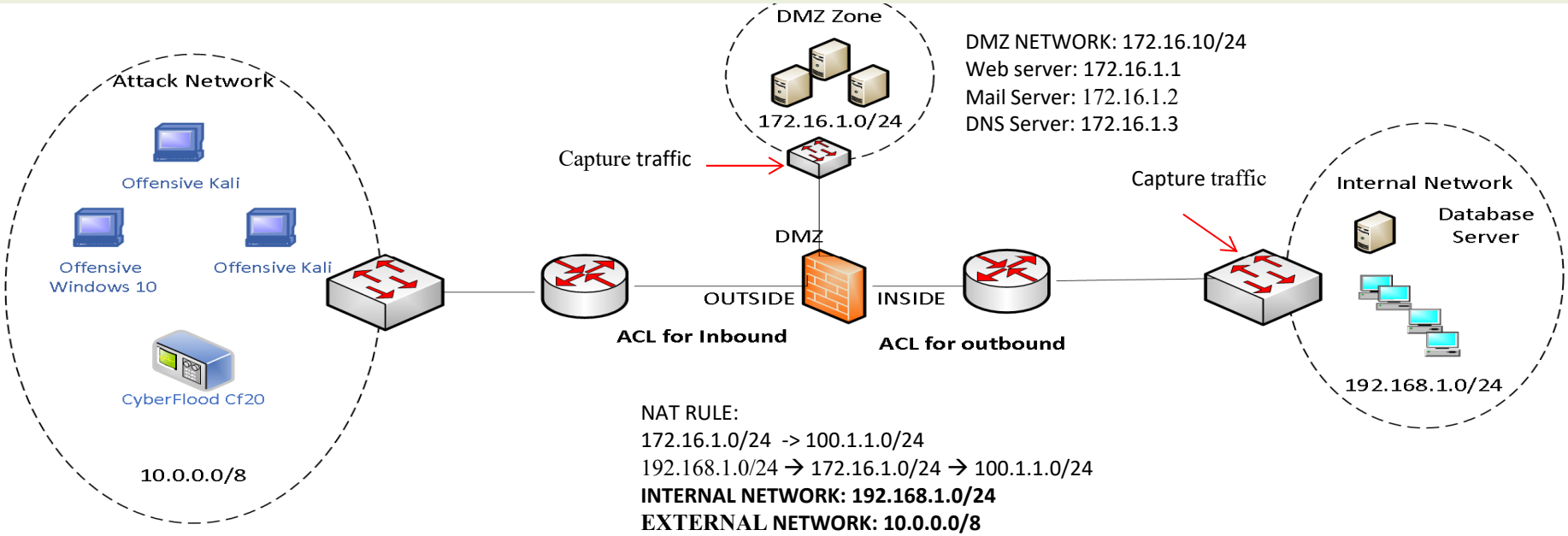
- Size of the dataset
- Lack of diverse Traffic
- Lack of diverse attacks
- Unrealistic scenarios
- Missing full network configurations

PRAIRIE VIEW A&M UNIVERSITY | SECURE | OnRampII - TAMU Camps Focused Research | TAMU

Work Methodology

Network Traffic → Capture Traffic (PCAP format) → Pre-processing raw data → Processed Dataset → ML Based IDS

Network Architecture: Testbed



NAT RULE:
172.16.1.0/24 -> 100.1.1.0/24
192.168.1.0/24 → 172.16.1.0/24 → 100.1.1.0/24
INTERNAL NETWORK: 192.168.1.0/24
EXTERNAL NETWORK: 10.0.0.0/8

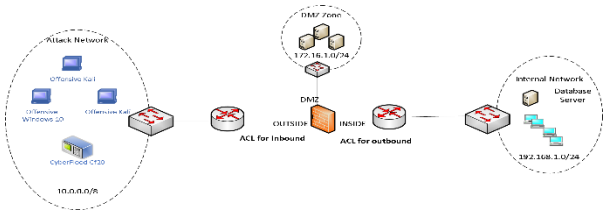
Unified Attack Models*

Type	Description	Subtype (Per DARPA Dataset)	Attack Tool
Communication Passive	Gathering information without active damage	Probe	Nmap, CF20
Communication Active	Attacks conducted actively and damage the system	DOS, DDoS	CyberFlood CF20
Host	Attacks to a host by installing malicious code into it	R2L (unauthorized access from a remote machine)	Njrat, vantom
		U2R (unauthorized access to local superuser (root) privileges)	Http tunnel Buffer overflow medusa ftp-write
Application	Attacks which targets a given application aiming at executing malicious code by penetrating interfaces	R2L	SQL Injection Cross Site Scripting CF20

*Adapted from Falcao, Zoppi, et. Al, "Quantitative Comparison of Unsupervised Anomaly Detection," in Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, April 2019

Communication Passive Class: Probing

- Only probing attack
- Probing attack with background HTTP traffic (used CF20, Traffic load 100 Mbps)



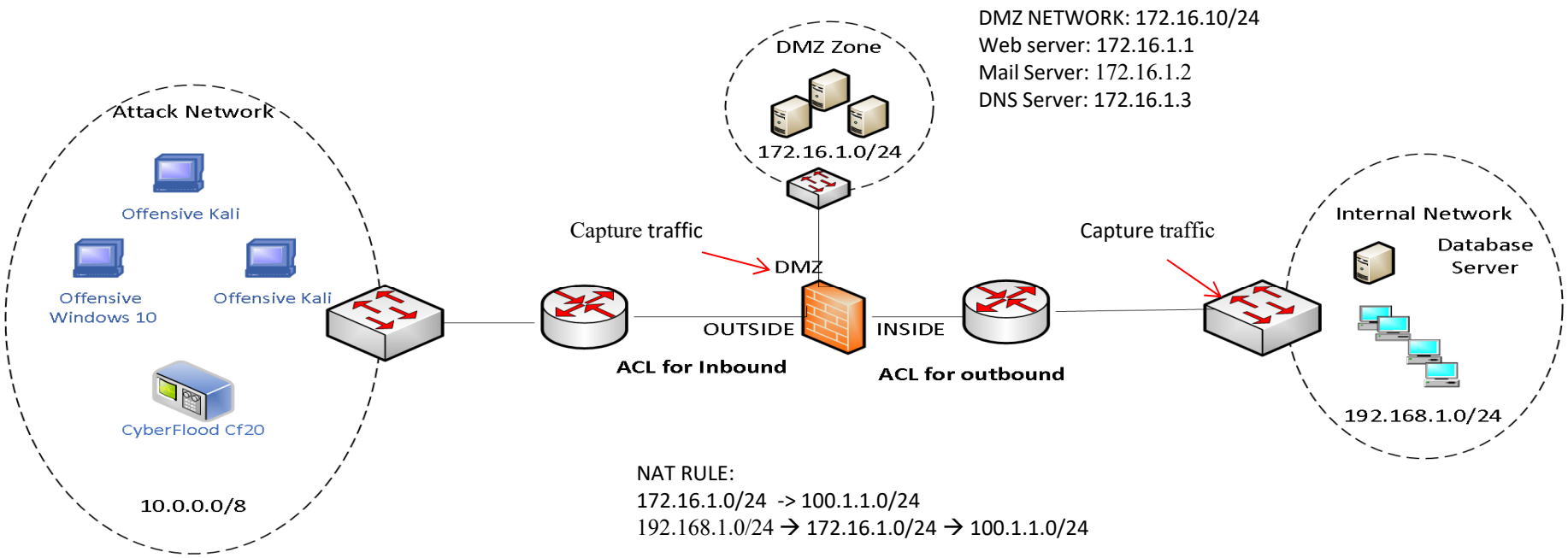
Probe Type
Host scan
Ping scan
TCP SYN scan
TCP connect scan
TCP ACK scan
TCP window scan
UDP scan
Null scan
FIN scan
XMAS scan
OS-fingerprinting

Attack Network IP address	Victim Network	Responding servers after NAT translation	Type	Duration	File Size, GB (Only Attack)	File Size, GB (Attack w/ background traffic)
10.0.0.21-200	100.1.1.0 Server Address: 172.16.1.100-15	172.16.1.20 Raw	host scan ping scan SYN scan connect scan ACK scan TCP window scan UDP scan Null scan FIN scan XMAS scan OS-fingerprinting	3 to 6 mins	2-4 GB	3-5 GB

Communication Active Class: Distributed Denial of Service: Volumetric & Protocol based

Type	Description
ICMP Flood	ICMP Flood (L3 based attack) consumes computing power and saturate bandwidth. This flood-attack is simple, but it can quickly bring down perimeter devices or saturate bandwidth. ICMP has various types. In the dataset, we included ICMP Type 3, 8, and 11 floods.
UDP Flood	A UDP flood saturates network bandwidth and brings down network services. The attack includes **a rapid succession of UDP datagrams (garbage payload) with spoofed IPs to a server using different ports (53, 123).** It forces the server to respond with ICMP traffic. Both ingress and egress bandwidth is saturated with this attack. We included UDP Flood Port 53 record in the dataset.
TCP SYN flood	**TCP SYN flood is for exhausting states by which legitimated requests are denied**, or the server is unable to respond to any valid request. A TCP SYN flood attack occurs when the attacker floods the system with SYN requests, by which the **TCP layer is saturated and prevent the completion of the TCP three-way handshake** between the client and server on every port. The target server's communications port becomes a half-open state. We used ports 25, 80, and 443 for the SYN flood.
XMAS Tree flood	A Christmas tree attack sends **a large number of packets with all the options turned on so that any protocol can be used**. This kind of packet requires **more processing by end devices and routers**. If a large number of this packet is sent, it keeps the **receivers busy** in processing, and it is nearly impossible to do other tasks, and thus the receivers deny the legitimate traffic.
Benign Traffic	HTTP1.0, HTTP1.1, HTTPs, FTP, SSH-Login, Telnet, SMTP, POP3, RTSP, SIP, BitTorre, Exchange, CIFS/SMB, IMAP4, SSH, Facebook traffic.

Network Architecture: Testbed



NAT RULE:
172.16.1.0/24 -> 100.1.1.0/24
192.168.1.0/24 → 172.16.1.0/24 → 100.1.1.0/24

INTERNAL NETWORK: 192.168.1.0/24
EXTERNAL NETWORK: 10.0.0.0/8

Volumetric DDoS Test Outcomes

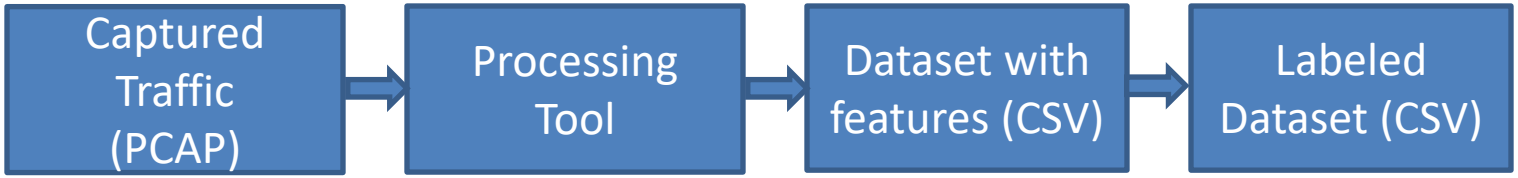
Attack Network	Victim Network	VDDoS Test	Type	Duration	File Size, GB (Only Attack)	File Size, GB (Attack background traffic)
10.0.0.101-200	172.16.1.0 Server Address: 172.16.1.100-15	Normal	ICMP 0,3,8,11	3 to 6 mins	11-16 GB	10-12 GB
		Malformed	TCP Flood		16.3	10-12 GB
			UDP Flood (Port: 53, 123)		16.5	10-12 GB
		Raw	IP 44		16.4	10-12 GB
			IP 58		16.5	10-12 GB
			IP 74		16.6	10-12 GB

Normal – Normal packets defined by the standard.
Malformed – Invalid TCP/UDP checksum, giving the packet the illusion of containing a TCP/UDP header along with application data.
Raw – Raw IP protocol packets are defined as IP packets with the IP protocol field set but no subsequent header defined

Protocol DDoS Test Outcomes

Attack Network IP address	Victim Network	PDDoS Test (Connection Oriented) Load 500Mbps	File Size, GB (Only Attack)	File Size, GB (Attack with background traffic)
10.0.0.101-200	172.16.1.101-115	SYN Flood – Port 25 SYN Flood – Port 80 SYN Flood – Port 443	7-10 GB	4-6 GB
		ACK Flood – Port 25 ACK Flood – Port 80 ACK Flood – Port 443	7-10 GB	4-6 GB
		RST Flood Type -1 Port 25 RST Flood Type -1 Port 80 RST Flood Type -1 Port 443	7-10 GB	4-6 GB
		RST Flood Type -2 Port 25 RST Flood Type -2 Port 80 RST Flood Type -2 Port 443	7-10 GB	4-6 GB
		XMAS Tree Flood – Port 25 XMAS Tree Flood – Port 80 XMAS Tree Flood – Port 443	7-10 GB	4-6 GB

Dataset Development



Features

- Used CICFLowMeter to extract the traffic features*
- The tool extracted 80 features (76 features below)

Flow Duration	Tot Fwd Pkts	Tot Bwd Pkts	TotLen Fwd Pkts	TotLen Bwd Pkts	Fwd Pkt Len Max	Fwd Pkt Len Min	Fwd Pkt Len Mean	Fwd Pkt Len Std
Bwd Pkt Len Max	Bwd Pkt Len Min	Bwd Pkt Len Mean	Bwd Pkt Len Std	Flow Byts/s	Flow Pkts/s	Flow IAT Mean	Flow IAT Std	Flow IAT Max
Flow IAT Min	Fwd IAT Tot	Fwd IAT Mean	Fwd IAT Std	Fwd IAT Max	Fwd IAT Min	Bwd IAT Tot	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 Header Len	Bwd Header Len	Fwd Pkts/s
Bwd Pkts/s	Pkt Len Min	Pkt Len Max	Pkt Len Mean	Pkt Len Std	Pkt Len Var	FIN Flag Cnt	SYN Flag Cnt	RST Flag Cnt
PSH Flag Cnt	ACK Flag Cnt	URG Flag Cnt	CWE Flag Count	ECE Flag Cnt	Down/Up Ratio	Pkt Size Avg	Fwd Seg Size Avg	Bwd Seg Size Avg
Fwd Byts/b Avg	Fwd Pkts/b Avg	Fwd Blk Rate Avg	Bwd Byts/b Avg	Bwd Pkts/b Avg	Bwd Blk Rate Avg	Subflow Fwd Pkts	Subflow Fwd Byts	Subflow Bwd Pkts
Subflow Bwd Byts	Init Fwd Win Byts	Init Bwd Win Byts	Fwd Act Data Pkts	Fwd Seg Size Min	Active Mean	Active Std	Active Max	Active Min
Idle Mean	Idle Std	Idle Max	Idle Min					

*Arash Habibi Lashkari, Gerard Draper-Gil, Mohammad Saiful Islam Mamun and Ali A. Ghorbani, "Characterization of Tor Traffic Using Time Based Features", In the proceeding of the 3rd International Conference on Information System Security and Privacy, SCITEPRESS, Porto, Portugal, 2017

PCA Analysis on PVAMU DDoS Dataset 2020

Rank	Score	Feature
1	0.954063	Flow Byts/s
2	0.942186	ACK Flag Cnt
3	0.93532	Down/Up Ratio
4	0.934062	FIN Flag Cnt
5	0.932798	Fwd IAT Tot
6	0.932179	RST Flag Cnt
7	0.932171	Bwd URG Flags
8	0.932171	URG Flag Cnt
9	0.932076	PSH Flag Cnt
10	0.932076	Bwd PSH Flags
11	0.931256	Flow Duration
12	0.928051	Init Bwd Win Byts
13	0.9238	Fwd Pkts/s
14	0.923197	Flow Pkts/s

ML Algorithms

- Random Forest (RF)
- Iterative Dichotomiser 3 (ID3)
- Gaussian Naïve Bayes (GNB),
- Multilayer Perceptron (MLP, a class of feedforward artificial neural network (ANN))
- Quadratic Discriminant Analysis (QDA)
- Adaptive Boosting (AdaBoost),
- K-Nearest Neighbor (KNN)

Comparison of ML Algorithms for DDoS Detection*

Algorithm	Pr	Rc.	F1	Execution (Sec)
Gaussian Naïve Bayes (GNB),	0.998	0.998	0.998	13.37
Quadratic Discriminant Analysis (QDA)	0.811	0.554	0.534	13.72
Random Forest (RF)	0.893	0.873	0.863	14.98
Iterative Dichotomiser 3 (ID3)	1	1	1	18.25
Adaptive Boosting (AdaBoost),	1	1	1	3057.83
Multilayer Perceptron (MLP)	0.893	0,873	0.863	3967.68
K-Nearest Neighbor (KNN)	0.998	0.998	0.998	15950.61

Standard evolution metrics:
Precision (Pr), Recall (Rc), and F-Measure (F1 score)

$$Pr = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$Rc = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F1\ Score = 2 * \frac{Pr * Rc}{Pr + Rc}$$

*S. Alam, Y. Alam, S. Cui, C. Akujuobi and M. Chouikha, "Toward Developing a Realistic DDoS Dataset for Anomaly-based Intrusion Detection," 2021 IEEE International Conference on Consumer Electronics (ICCE), 2021, pp. 1-6, doi: 10.1109/ICCE50685.2021.9427660.

Tasks to Complete

- Complete following test scenarios
  - Host: **R2L** (unauthorized access from a remote machine)
  - Applications: **U2R** (unauthorized access to local superuser (root) privileges)
- Complete dataset analysis using unsupervised algorithms
  - One Class SVM
  - Deep autoencoder
  - Isolation Forest
  - CNN
  - PCA Based
- Share various captured raw traffic files and datasets

Student Experience

PVAMU Students:

- Willie Waters
- Toya Acharya
- Denedra Woods
- Faith Nwokoma
- Aaron Brown
- Tawab Attaie
- Gbenga Agunsoye

Student activities:

- Training on network equipment such as routers, switches, firewall and CyberFlood CF20
- Server configuration and management
- OS: Linux, Windows
- Traffic generation and collection
- Python Programming
- ML and DL algorithms

Conference Publications

- S. Alam, Y. Alam, S. Cui, C. Akujuobi and M. Chouikha, "Toward Developing a Realistic DDoS Dataset for Anomaly-based Intrusion Detection," 2021 IEEE International Conference on Consumer Electronics (ICCE), 2021

- Toya Acharya, Ishan Khatri, A. Annamalai, Mohamed Chouikha, “Efficacy of Heterogeneous Ensemble-Assisted Machine Learning Model for Binary and Multi-Class Network Intrusion Detection,” 2021 IEEE International Conference on Automatic Control & Intelligent Systems, June 2021.

- Toya Acharya, Ishan Khatri, A. Annamalai, M. Chouikha, "Efficacy of Machine Learning-Based Classifiers for Binary and Multi-Class Network Intrusion Detection," 2021 IEEE International Conference on Automatic Control & Intelligent Systems, June 2021.

Thank you!