Các bộ dữ liệu dùng cho phát hiện bất thường và xâm nhập mạng

1 Bộ dữ liệu NSL-KDD

Bộ dữ liệu NSL-KDD là một bộ dữ liệu phổ biến dùng để đánh giá lưu lượng truy cập internet. Bộ dữ liệu gồm 4 tập con: KDDTest+, KDDTest-21, KDDTrain+, KDDTrain+_20Percent, trong đó KDDTest-21 và KDDTrain+_20Percent là tập con của KDDTrain+ và KDDTest+. Mỗi bản ghi trong bộ dữ liệu sẽ bao gồm 43 đặc trưng, trong đó 41 đặc trưng là lưu lượng đầu vào, 2 thông tin còn lại là nhãn (là tấn công mạng hay bình thường) và mức độ quan trọng của các thông tin đầu vào đó.

Trong bộ dữ liệu có 4 kiểu tấn công mạng:

1. Probe attack: đây là một kiểu tấn công quét mạng (network) để tìm thông tin. Mục tiêu tìm ra cổng port đang mở để tấn công.
2. R2L (Remote to Local): đây là kiểu tấn công cố gắng tiếp cận máy nội bộ từ xa khi không có tài khoản hoặc quyền truy cập.
3. U2R (User to Root): đây là kiểu tấn công chiếm quyền toàn bộ hệ thống (root).
4. Dos: đây là cuộc tấn công làm gián đoạn mạng bằng việc làm quá tải hệ thống mạng.

Classes:	DoS	Probe	U2R	R2L
Sub-Classes:	• apache2 • back • land • neptune • mailbomb • pod • processtable • smurf • teardrop • udpstorm • worm	• ipsweep • mscan • nmap • portsweep • saint • satan	• buffer_overflow • loadmodule • perl • ps • rootkit • sqlattack • xterm	• ftp_write • guess_passwd • httptunnel • imap • multihop • named • phf • sendmail • Snmpgetattack • spy • snmpguess • warezclient • warezmaster • xlock • xsnoop
Total:	11	6	7	15

Hình 1.1. Những lớp con của từng kiểu tấn công.

Hình ảnh bên dưới đây mô tả phân bố của các kiểu tấn công trên bộ dữ liệu. Bộ dữ liệu có phân bố tương đối lệch, nhãn normal chiến chủ yếu, trong khi đó đối với các kiểu tấn công thì DoS chiếm đã số.

Dataset	Number of Records:					
	Total	Normal	DoS	Probe	U2R	R2L
KDDTrain+20%	25192	13449 (53%)	9234 (37%)	2289 (9.16%)	11 (0.04%)	209 (0.8%)
KDDTrain+	125973	67343 (53%)	45927 (37%)	11656 (9.11%)	52 (0.04%)	995 (0.85%)
KDDTest+	22544	9711 (43%)	7458 (33%)	2421 (11%)	200 (0.9%)	2654 (12.1%)

Hình 1.2: Phân bố dữ liệu theo các kiểu tấn công.

Bộ dữ liệu gồm 43 đặc trưng như mô tả ở bảng bên dưới:

#	Feature Name	Description	Type	Value Type	Ranges (Between both train and test)	
1	Duration	Length of time duration of the connection	Continuous	Integers	0 - 54451	
2	Protocol Type	Protocol used in the connection	Categorical	Strings		
3	Service	Destination network service used	Categorical	Strings		
4	Flag	Status of the connection – Normal or Error	Categorical	Strings		
5	Src Bytes	Number of data bytes transferred	Continuous	Integers	0 - 137996388	

		from source to destination in single connection			8	
6	Dst Bytes	Number of data bytes transferred from destination to source in single connection	Continuous	Integers	0 - 309937401	
7	Land	If source and destination IP addresses and port numbers are equal then, this variable takes value 1 else 0	Binary	Integers	{ 0 , 1 }	*** Values within {} are exact/possible values.
8	Wrong Fragment	Total number of wrong fragments in this connection	Discrete	Integers	{ 0,1,3 }	
9	Urgent	Number of urgent packets in this connection. Urgent packets are packets with the urgent bit activated	Discrete	Integers	0 - 3	
10	Hot	Number of “hot“ indicators in the content such as: entering a system directory, creating programs and executing programs	Continuous	Integers	0 - 101	
11	Num Failed Logins	Count of failed login attempts	Continuous	Integers	0 - 4	
12	Logged In	Login Status : 1 if successfully	Binary	Integers	{ 0 , 1 }	

		logged in; 0 otherwise				
13	Num Compromised	Number of "compromised" conditions	Continuous	Integers	0 - 7479	
14	Root Shell	1 if root shell is obtained; 0 otherwise	Binary	Integers	{ 0 , 1 }	
15	Su Attempted	1 if "su root" command attempted or used; 0 otherwise	Discrete (Dataset contains '2' value)	Integers	0 - 2	
16	Num Root	Number of "root" accesses or number of operations performed as a root in the connection	Continuous	Integers	0 - 7468	
17	Num File Creations	Number of file creation operations in the connection	Continuous	Integers	0 - 100	
18	Num Shells	Number of shell prompts	Continuous	Integers	0 - 2	
19	Num Access Files	Number of operations on access control files	Continuous	Integers	0 - 9	
20	Num Outbound Cmds	Number of outbound commands in an ftp session	Continuous	Integers	{ 0 }	
21	Is Hot	1 if the login	Binary	Integer	{ 0 , 1 }	

	Logins	belongs to the "hot" list i.e., root or admin; else 0		s		
22	Is Guest Login	1 if the login is a "guest" login; 0 otherwise	Binary	Integers	{ 0 , 1 }	
23	Count	Number of connections to the same destination host as the current connection in the past two seconds	Discrete	Integers	0 - 511	
24	Srv Count	Number of connections to the same service (port number) as the current connection in the past two seconds	Discrete	Integers	0 - 511	
25	Serror Rate	The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in count (23)	Discrete	Floats (hundredths of a decimal)	0 - 1	
26	Srv Serror Rate	The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in srv_count (24)	Discrete	Floats (hundredths of a decimal)	0 - 1	

27	Rerror Rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in count (23)	Discrete	Floats (hundredths of a decimal)	0 - 1	
28	Srv Rerror Rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in srv_count (24)	Discrete	Floats (hundredths of a decimal)	0 - 1	
29	Same Srv Rate	The percentage of connections that were to the same service, among the connections aggregated in count (23)	Discrete	Floats (hundredths of a decimal)	0 - 1	
30	Diff Srv Rate	The percentage of connections that were to different services, among the connections aggregated in count (23)	Discrete	Floats (hundredths of a decimal)	0 - 1	
31	Srv Diff Host Rate	The percentage of connections that were to different destination machines among the connections aggregated in srv_count (24)	Discrete	Floats (hundredths of a decimal)	0 - 1	

32	Dst Host Count	Number of connections having the same destination host IP address	Discrete	Integers	0 - 255	
33	Dst Host Srv Count	Number of connections having the same port number	Discrete	Integers	0 - 255	
34	Dst Host Same Srv Rate	The percentage of connections that were to different services, among the connections aggregated in dst_host_count (32)	Discrete	Floats (hundredths of a decimal)	0 - 1	
35	Dst Host Diff Srv Rate	The percentage of connections that were to different services, among the connections aggregated in dst_host_count (32)	Discrete	Floats (hundredths of a decimal)	0 - 1	
36	Dst Host Same Src Port Rate	The percentage of connections that were to the same source port, among the connections aggregated in dst_host_srv_count (33)	Discrete	Floats (hundredths of a decimal)	0 - 1	
37	Dst Host Srv Diff Host Rate	The percentage of connections that were to different destination	Discrete	Floats (hundredths of a	0 - 1	

		machines, among the connections aggregated in dst_host_srv_count (33)		decimal)		
38	Dst Host Serror Rate	The percentage of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in dst_host_count (32)	Discrete	Floats (hundredths of a decimal)	0 - 1	
39	Dst Host Srv Serror Rate	The percent of connections that have activated the flag (4) s0, s1, s2 or s3, among the connections aggregated in dst_host_srv_count (33)	Discrete	Floats (hundredths of a decimal)	0 - 1	
40	Dst Host Rerror Rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in dst_host_count (32)	Discrete	Floats (hundredths of a decimal)	0 - 1	
41	Dst Host Srv Rerror Rate	The percentage of connections that have activated the flag (4) REJ, among the connections aggregated in	Discrete	Floats (hundredths of a decimal)	0 - 1	

		dst_host_srv_count (33)				
42	Class	Classification of the traffic input	Categorical	Strings		
43	Difficulty Level	Difficulty level	Discrete	Integers	0 - 21	

Các đặc trưng của bộ dữ liệu được phân chia vào bốn nhóm:

- Intrinsic feature: có thể được truy xuất từ header của gói tin mà không cần nhìn vào nội dung. (feature 1-9).
- Content feature: chứa thông tin về gói tin ban đầu (gói tin gốc được chia thành các gói tin nhỏ và gửi đi nhiều lần). (feature 10-12).
- Time-based feature: chứa thông tin phân tích của lưu lượng đầu vào trong khoảng 2 giây và chứa thông tin như có bao nhiêu kết nối cố thực hiện tới host. (feature 23-31).
- Host-based feature: phân tích các kết nối (số lượng requests được thực hiện tới host trong x connections).(feature 32-41).

Kiểu dữ liệu của các đặc trưng được chia thành 4 kiểu:

- 4 categorical (features 2,3,4,42)
- 6 binary (features: 7, 12, 14, 20, 21,22)
- 23 discrete (features: 8,9,15,23-41,43)
- 10 contínuos (features: 1,5,6,10,11,13,16,17,18,19).

Dưới đây là một số ví dụ về các categorical feature.

Protocol Type (2)	Service (3)				Flag (4)
• icmp • tcp • udp	• other • link • netbios_ssn • smtp • netstat • ctf • ntp_u • harvest • efs • klogin • systat • exec • nntp • pop_3 • printer • vmnet • netbios_ns	• urh_i • ssh • http_8001 • iso_tsap • aol • sql_net • shell • supdup • auth • whois • discard • sunrpc • urp_i • Rje • ftp • daytime • domain_u • pm_dump	• time • hostnames • name • ecr_i • bgp • telnet • domain • ftp_data • nnsp • courier • finger • uucp_path • X11 • imap4 • mtp • login • tftp_u • kshell	• private • http_2784 • echo • http • ldap • tim_i • netbios_dgm • uucp • eco_i • Remote_job • IRC • http_443 • red_i • Z39_50 • Pop_2 • gopher • Csnet_ns	• OTH • S1 • S2 • RSTO • RSTRs • RSTOS0 • SF • SH • REJ • S0 • S3

Flag	Value	Flag	Description
SF	Normal establishment and termination. Note that this is the same symbol as for state S1. You can tell the two apart because for S1 there will not be any byte counts in the summary, while for SF there will be	RSTO	Connection reset by the originator
REJ	Connection attempt rejected	RSTR	Connection reset by the responder
S0	Connection attempt seen, no reply	OTH	No SYN seen, just midstream traffic (a “partial connection” that was not later closed)
S1	Connection established, not terminated	RSTOS0	Originator sent a SYN followed by a RST, we never saw a SYN-ACK from the responder
S2	Connection established and close attempt by originator seen (but no reply from responder)	SH	Originator sent a SYN followed by a FIN, we never saw a SYN ACK from the responder (hence the connection was “half” open)
S3	Connection established and close attempt by responder seen (but no reply from originator)	SHR	Responder sent a SYN ACK followed by a FIN, we never saw a SYN from the originator. (Not in NSL-KDD but still a flag)

2. Bộ dữ liệu CIC-IDS-2017

Bộ dữ liệu được thực hiện tại Viện An ninh mạng Canada trong năm ngày và được lưu thành 8 files khác nhau.

Name of Files	Day Activity	Attacks Found
Monday-WorkingHours.pcap_ISCX.csv	Monday	Benign (Normal human activities)
Tuesday-WorkingHours.pcap_ISCX.csv	Tuesday	Benign, FTP-Patator, SSH-Patator
Wednesday-workingHours.pcap_ISCX.csv	Wednesday	Benign, DoS GoldenEye, DoS Hulk, DoS Slowhttptest, DoS slowloris, Heartbleed
Thursday-WorkingHours-Morning-WebAttacks.pcap_ISCX.csv	Thursday	Benign, Web Attack – Brute Force, Web Attack – Sql Injection, Web Attack – XSS
Thursday-WorkingHours-Afternoon-Infilteration.pcap_ISCX.csv	Thursday	Benign, Infiltration
Friday-WorkingHours-Morning.pcap_ISCX.csv	Friday	Benign, Bot
Friday-WorkingHours-Afternoon-PortScan.pcap_ISCX.csv	Friday	Benign, PortScan
Friday-WorkingHours-Afternoon-DDos.pcap_ISCX.csv	Friday	Benign, DDoS

Hình 2.1: Mô tả các tệp trong bộ dữ liệu CICIDS2017

Bộ dữ liệu gồm 83 features và 15 lớp (1 normal và 14 attacks).

Class Labels	Number of instances
BENIGN	2359087
DoS Hulk	231072
PortScan	158930
DDoS	41835
DoS GoldenEye	10293
FTP-Patator	7938
SSH-Patator	5897
DoS slowloris	5796
DoS Slowhttptest	5499
Bot	1966
Web Attack – Brute Force	1507
Web Attack – XSS	652
Infiltration	36
Web Attack – Sql Injection	21
Heartbleed	11

Hình 2.2: Thông tin label của bộ dữ liệu CICIDS2017

Bộ dữ liệu CICIDS2017 rất đa dạng và đầy đủ nhiều loại thông tin như: cấu hình mạng hoàn chỉnh, lưu lượng, các kiểu tấn công đa dạng, thông tin về giao thức, meta data. Tuy nhiên bộ dữ liệu rất cân bằng.

Sl No	Normal / Attack Labels	Number of in-stances	% of preva-lence w.r.t. the majority class	% of preva-lence w.r.t. the total instances
1	BENIGN	2359087	1	83.34406
2	Bot	1966	0.000833	0.06946
3	DDoS	41835	0.017734	1.47799
4	DoS GoldenEye	10293	0.004363	0.36364
5	DoS Hulk	231072	0.09795	8.16353
6	DoS Slow-httptest	5499	0.002331	0.19427
7	DoS slowloris	5796	0.002457	0.20477
8	FTP-Patator	7938	0.003365	0.28044
9	Heartbleed	11	0.000005	0.00039
10	Infiltration	36	0.000015	0.00127
11	PortScan	158930	0.067369	5.61483
12	SSH-Patator	5897	0.0025	0.20833
13	Web Attack – Brute Force	1507	0.000639	0.05324
14	Web Attack – Sql Injection	21	0.000009	0.00074
15	Web Attack – XSS	652	0.000276	0.02303

Hình 2.3: Phân bố dữ liệu qua các lớp