Datasets used for Intrusion Detection using Machine Learning: A Survey

G.Prethija[1], V.Subbulakshmi[2], K.Devi[3]

[1]Assistant Professor, Dept. of Information Technology, Velammal Engineering College, Tamil Nadu ,India
[2]Assistant Professor, Dept. of Information Technology, Velammal Engineering College, Tamil Nadu ,India
[3]Assistant Professor, Dept. of Information Technology, Velammal Engineering College, Tamil Nadu ,India

---***---

Abstract - *In digital era, cybercrime become a business. Nowadays, cyberattacks cause loss of sensitive data and severe financial loss to organizations. Therefore, cybersecurity expert's role is very important to protect the data from attacks. Researchers focus on intrusion detection to detect those unknown attacks. Machine learning algorithms plays a vital role in intrusion detection since it detects attacks accurately. The datasets used in most of the literature for intrusion detection are KDD Cup 99, NSL-KDD, UNSW-NB15, Kyoto* and CSCIDS 2017*. The detailed analysis of the datasets is discussed. The performance metrics used for evaluating the machine learning algorithms are also discussed. This study will be helpful for researchers to develop an efficient intrusion detection system.*

Key Words: Intrusion Detection, Machine Learning, KDD Cup 99, NSL-KDD, Kyoto, UNSW-NB15, Accuracy

1. INTRODUCTION

Nowadays businesses and governments deal with huge amount of data which is stored in computers and transmit across various networks to other systems. Due to the usage of huge amount of data, there is a possibility of data breach. Cyber security which plays a major role in defending various resources such as computers, mobile devices, networks, servers and data from variety of malicious attacks. Cyber security is also named as information technology security or electronic communication security. Cyber security is defined as the variety of technologies, processes and methods to protect the confidentiality, integrity, and availability of resources, against cyber-attacks or unauthorized access. The main aim of cyber security is to protect all organizational assets from both internal threats and external threats as well as disruption caused due to various natural disasters.

1.1 Various domains in cyber-security

(i) Application Security: It focuses to keep the software resources as free of threats. It implements the system as a secure one by designing the secure application architecture, writing the code as more secure one, implement the strong data input validation process and various threat modeling in order to minimize the chances of any unauthorized access or modifying the application resources

(ii) Identity Management and Data Security: Within the organization, it includes the framework, processes and other activities which enable the process of authentication and the authorization of legitimate users to information systems.

(v)Mobile Security: Organizational information and personal information are stored in mobile devices like cell phones, laptops, tablets, etc. from various threats such as unauthorized access, device loss or theft, malware, etc. Mobile security which provides the technique to secure the data in various mobile devices.

(vi) Disaster recovery and business continuity: It deals with the process of how the organization is going to deal with the cyber-security incident or any other disaster which causes the loss of operations on data.

(vii)Cloud Security: It is used to secure the data which is stored in various cloud providers storage such as AWS, Google, Azure, etc.

1.2 Intrusion Detection

Intrusion detection system are classified based on source of data and detection methodology. The classification of intrusion detection is depicted in Fig. 1. If the intrusion is monitored on hosts or devices on the network, then it is called Host intrusion detection systems (HIDS). In network-based IDS, the intrusion is monitored on the whole network.

Based on detection methodology, IDS is classified into signature based and misuse-based approach. In a misuse detection approach, abnormal behavior of network is matched against known patterns of detected attacks. Anomaly based detection determines the unusual network traffic.

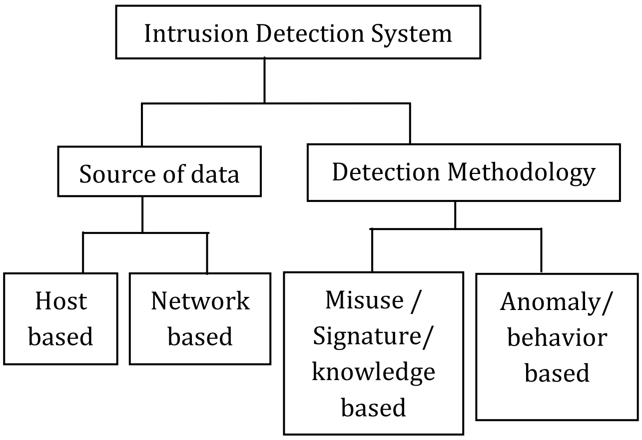

Fig- 1: Classification of Intrusion detection

2. METRICS USED TO EVALUATE INTRUSION DETECTION SYSTEM

The performance evaluation of any intrusion detection system can be done by the metrics such as: accuracy (ACC), Recall (REC), Precision (PRE), True Negative Rate (TNR), False Alarm Rate (FAR), False Negative Rate (FNR), F-Measure, Mathews Correlation Coefficient (MCC), ROC Graph and Kappa Statistics. The metrics required for evaluation are computed from confusion matrix (Table-1). A matrix that describes the performance of a given classification model (or "classifier") is called confusion matrix. It denotes true and false classification results. The ways in which confusion is made when a prediction is done by the classification model is depicted by confusion matrix.

		Predicted Class	
		Class 1 (Positive)	Class 2 (Negative)
Actual Class	Class 1 (Positive)	TP	FN
	Class 2 (Negative)	FP	TN

Table- 1: Confusion matrix

True positive (TP): It is the number of correctly identified anomaly records. False positive (FP): It represents the no. of incorrectly identified usual records that are detected as anomaly. True Negative (TN): It represents the number of correctly detected records. False Negative (FN): It shows the number of incorrectly detected anomaly records.

Accuracy (ACC): It is the ratio of correct classification done by a classifier.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Sensitivity, REC (Recall), hit rate, detection rate or True Positive Rate (TPR): It measures the proportion of positives that are exactly identified as positives. It gives the ratio of correctly identified records to the total number of abnormal records.

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

Precision (PRE): It is the ratio of correctly classified records over predicted positive cases.

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Specificity, True -ve Rate (TNR): It quantifies the fraction of negatives that are exactly identified as negatives.

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}$$

False +ve Rate (FPR) or False Alarm Rate (FAR): It gives the percentage of negative records that were incorrectly classified as positive.

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} = 1 - \mathrm{sensitivity}$$

False Negative Rate (FNR): It is the percentage of positive records that were incorrectly classified as negative.

$$\mathrm{FNR} = \frac{\mathrm{FN}}{\mathrm{FN} + \mathrm{TN}} = 1 - \mathrm{TPR}$$

F-measure (F-Score): It gives the sensitivity and precision of the harmonic mean.

$$F = \frac{2}{\left(\frac{1}{\mathrm{recall}}\right) + \left(\frac{1}{\mathrm{Precision}}\right)} = \frac{2 \times \mathrm{Recall} \times \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}} = \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

General formula for positive real β is:

$$F = \frac{(1+\beta)^2 \times \mathrm{Recall} \times \mathrm{Precision}}{\beta^2 \mathrm{Precision} + \mathrm{Recall}} = \frac{(1+\beta)^2 \times \mathrm{P} \times \mathrm{TP}}{\beta^2 \times \mathrm{P} + \mathrm{TP}}$$

G-Mean1: It is the geometric mean of precision and true positive rate.

$$\text{G-Mean1} = \sqrt{\mathrm{TP} * \mathrm{P}}$$

G-Mean2: It is the geometric mean of true positive rate and true negative rate.

$$\text{G-Mean2}=\sqrt{TP * TN}$$

Matthews correlation coefficient (MCC): It measures the quality of binary classifications. It returns a value between −1 and +1.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

Receiver operating characteristic (ROC) Graph: A ROC graph examines the performance of classifiers. A ROC graph plots false alarm rate in the horizontal(X) axis and the sensitivity in the vertical(Y) axis.

Kappa statistic: It is the comparison between observed accuracy and expected accuracy (random chance). Observed accuracy is the count of exactly classified instances throughout the confusion matrix. Expected accuracy gives the accuracy that any classifier can achieve from the confusion matrix.

$$Kappa = \frac{\text{observed accuracy} - \text{expected accuracy}}{1 - \text{expected accuracy}}$$

3. DATASETS USED FOR INTRUSION DETECTION RESEARCH

Most researchers used the datasets DARPA, KDD (Knowledge Discovery and Data Mining) Cup and NSL-KDD (Network Security Laboratory-KDD), UNSW-NB15, Kyoto and CSCIDS 2017 for intrusion detection. The datasets used for intrusion detection by researchers have both training data and testing data. The data set size comparison of training and test data for different datasets is shown in Table- 2.

Table- 2: Data Size comparison for different datasets

Dataset	Training size	Testing size
DARPA 99	6.2GB	3.67GB
KDD99	4898431bytes	311029 bytes
NSL-KDD	125973 bytes	22444 bytes
UNSW-NB15	175,341 bytes	82,332 bytes
AWID	1,795,575 bytes	575,643 bytes

3.1 DARPA

The first standard corpus for the evaluation of intrusion detection system was created by MIT Lincoln Laboratory's in 1998 under the sponsorship of DARPA [1].

Lippmann et al. [2] developed a normal traffic scenario which is quite analogous to users of nearly 100's working on 1000's of workstation using intrusion detection evaluation test bed. The observations showed that detection rates were worse for new and novel R2L and DoS attacks.

The second DARPA off-line intrusion detection evaluation was done in 1999. Lippmann et al. [2] analyzed training data for three weeks and test data for two weeks and they found that over 200 instances of 58 attack types were launched in UNIX and Windows NT hosts. The major drawback is among 58 attack types; nearly ten attacks were not identified by any system because TCP services protocols and were not properly analyzed.

3.2 KDD Cup 99

The tcpdump portions (about 4 GB compressed tcpdump data for network traffic of 7 weeks of the 1998 DARPA Intrusion Detection System (IDS) Evaluation dataset were processed to create the dataset namely KDD Cup 99 [3] which was created by Lincoln Lab under contract to DARPA [4]. The data contains four main categories of attacks namely DoS (Denial of Service), U2R (User to Root), R2L (Remote to Local) and probing attack. DOS attack is an attack which denies resources. In U2R, user attacks gain root access through user account. R2L is a type of attack which sends packet through the network by gaining local access of the host as a user. Probing Attack attempts to gather information about a network of computers. Each connection records include traffic, intrinsic and content features.

Tavalee et al. [5] have analyzed KDD dataset in detail. KDD'99 features can be classified into three groups' namely basic, traffic and content features. The major problem in this dataset is the enormousness of duplicate records. They analyzed both test and training sets and reported that nearly 75% and 78% of the records are redundant. Due to these huge redundant records, during training the learning algorithms are biased towards more repeated records. It stops learning from records that are used infrequently which can cause harm to networks. The detection rates for these frequent records are better.

Atilla O et al. [6] analyzed KDD dataset and found that it consists two weeks of attacks-free instances and five weeks of attack instances. The dataset has a total of 38 attacks which includes 24 attack types in training and 14 more attack types in testing. Therefore, machine learning based IDS find it's difficult to detect these 14 new attacks. Only limited attacks are found under U2R and R2L.Since KDD99 is a large dataset for most machine learning algorithms, most researchers used a small percentage of it. This dataset is mainly used for anomaly type of intrusion. They tabulated the comparisons of training and testing size of different attacks.
Out of the 42 features/attributes in this data set (Table- 3), 41 attributes can be categorized into 4 different classes namely basic, content, traffic and host features. The value type is either continuous(C)/ discrete(D).The different types of attacks in KDD dataset are shown in Fig-3.

Table- 3: Features of KDD dataset

No.	Feature	Value Type (C/D)	Description
			Basic/Intrinsic Features
1	Duration	C	Connection length
2	protocol_type	D	Protocol type
3	service	D	Network service(telnet, http provided on the destination
4	src_bytes	C	No. of bytes(data)routed from source to destination
5	dst_bytes	C	No. of bytes(data) routed from destination to source
6	Flag	D	Status of established connection(normal /error)
7	land	D	Value is 1 if connection is established between same host/port. otherwise , the value is 0
8	wrong_fragment	C	"wrong fragments" count
9	Urgent	C	" urgent packets" count
			Content Features
10	Hot	C	"hot indicators" count
11	num_failed_logins	C	Count of"login attempts" that are failed
12	logged_in	D	Sets a value 1, if login is done successfully. Otherwise, value 0 is set.
13	num_compromised	C	"compromised" conditions count
14	root_shell	D	Value set is 1 if "root shell" is obtained; Otherwise, value 0 is set.
15	su_attempted	D	Value set is 1 if "su root command" is attempted;Otherwise, value 0 is set.
16	num_root	C	"root accesses" count
17	num_fil_creations	C	Number of operations for file creation
18	num_shells	C	"shell prompts" count
19	num_access_files	C	Count of create,, delete andwrite operations on files for access control
20	num_outbound_cmds	C	Count of "outbound commands" in FTP session
21	is_hot_login	D	Value is set 1 if it is "hot list login"(e.g.,adm, root, , etc.); otherwise,value set is 0
22	is_guest_login	D	Value is set 1 if it is "guest login" (e.g., anonymous,guest, etc.); otherwise,value set is 0
			Traffic features
23	count	C	Number of connections to the same host as the current connection in the past 2 seconds
24	serror_rate	C	Percentage(%) of "SYN" errorconnections
25	rerror_rate	C	Percentage(%) of "REJ" errorconnections
26	same_srv_ rate	C	Percentage(%) of connections that have same service
27	diff_srv_rate	C	Percentage(%) of connections that have different services
28	srv_count	C	Number of connections to the same service as the current connection in the past 2 seconds
29	srv_serror _rate	C	Percentage(%) of "SYN" errorconnections
30	srv_rerror _rate	C	Percentage(%) of "REJ" errorsconnections
31	srv_diff_host_rate	C	Percentage(%) of connections provided to different hosts
			Host Features
32	dst_host_count	C	Count for destination host

33	dst_host _srv_cou nt	C	Number of connections to the same destination port
34	dst_host _same_ srv_rate	C	Percentage(%) of connections that have same service
35	dst_host _diff_ srv_rate	C	Percentage(%) of connections that have different service
36	dst_host _same_ src_port _rate	C	Percentage(%) of same "source port" connections
37	dst_host _srv_ diff_host _rate	C	Percentage(%) of connections provided to different host
38	dst_host _serror _rate	C	Percentage(%) of connections that have "SYN" errors Type
39	dst_host _srv_ serror_r ate	C	Percentage(%) of connections that have "SYN" errors
40	dst_host _rerror _rate	C	Percentage(%) of connections that have "REJ" errors Type
41	dst_host _srv_ rerror_r ate	C	Percentage(%) of connections that have "REJ" errors

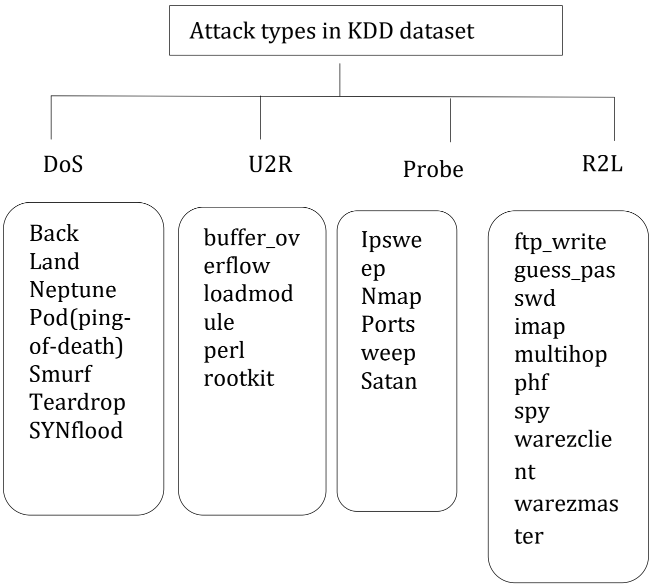

Fig- 3: Types of attacks in KDD dataset

3.3 NSL-KDD

Tavallaee et al [5] published the NSL-KDD dataset in their website [7] which is more beneficial than the original KDD data set. It eliminates duplicate records in training set thereby overcoming the drawback of classifiers gets biased towards more frequent records. Reasonably the number of records in train and test sets are selected which executes the complete set affordably. Only 20% of training data with total instances 25192 was identified as KDDTrain+_20Percent.The test data set named KDDTest+ has a total of 22544 instances.

3.4 UNSW-NB15 dataset

Due to absence of modern attack styles and traffic situations in KDD dataset, a new dataset (UNSW-NB15)[8] was developed by ACCS-an American Cyber security Center. This dataset has a 49-feature set and a total of 2,540,044 records [9]. The features in this dataset are tabulated in Table- 4. The types of attacks are shown in Table-5.

Table- 4: Features of UNSW-NB15 dataset

Feature No.	Feature Name	Description
Flow features		
1	Srcip	Source" IP address"
2	Sport	Source "Port address"
3	Dstip	Destination " IP address"
4	Dsport	Destination "Port number"
5	Proto	Type of protocol (UDP ,TCP)
Basic features		
6	State	Indicates "state " and " its dependent protocol" (CLO,ACC, and CON).
7	Dur	Total duration of connection
8	Sbytes	No. of bytes from source to destination
9	Dytes	No. of bytes from destination to source
10	Sttl	Time to live(TTL) of Source to destination
11	Dttl	TTL of destination to source
12	Sloss	Retransmitted / dropped source packets
13	Dloss	Retransmitted / dropped

		destination packets
14	Service	ftp ,http, ssh, smtp, dns service
15	Sload	Source bps(bits per second)
16	Dload	Destination bps(bits per second)
17	Spkts	Source to destination packet count
18	Dpkts	Destination to source packet count
Content features		
19	Swin	Source TCP window " advertisement value"
20	Dwin	Destination TCP window "advertisement value"
21	Stcpb	Source TCP base "sequence number"
22	Dtcpb	Destination TCP base "sequence number"
23	Smeansz	Mean of the " flow packet size" transmitted by source
24	dmeansz	Mean of the "flow packet size" transmitted by destination
25	trans_depth	Represents the pipelined depth into the connection of http request/response transaction
26	res_bdy_len	Actual uncompressed content size of the data transferred from the server's http service
Time features		
27	Sjit	Source jitter (mSec)
28	Djit	Destination jitter (mSec)
29	Stime	start time of record
30	Ltime	last time of record
31	sintpkt	Inter-packet arrival time of source (mSec)
32	dintpkt	Inter-packet arrival time of destination (mSec)
33	tcprtt	TCP connection setup RTT(round trip time)(synack + ackdat)
34	synack	TCP connection setup time (SYN_ACK packets - SYN packets)
35	Ackdat	TCP connection setup time

		(ACK packets - SYN_ACK)
Additional generated features		
36	is_sm_ips_ports	If srcip = dstip & sport = dsport, this variable is assigned to 1, otherwise value 0 is assigned
37	ct_state_ttl	No. for each state (6) according to specific range of values of sttl (10) and dttl (11)
38	ct_flw_http_mthd N	No. of flows that has methods such as Get and Post in http service
39	is_ftp_login	If ftp session is login using user and password then 1 is set. otherwise, 0 is set
40	ct_ftp_cmd	No of flows that has a command in ftp session
41	ct_srv_src	No. of records that have same" service" and "srcip" in "100 records" based on " ltime "
42	ct_srv_dst	No. of records that have same "service" and "dstip" in "100 records" based on "ltime"
43	ct_dst_ltm	No. of records of the same "dstip" in "100 records" based on "ltime"
44	ct_src_ltm	No. of records of "srcip" in "100 records" based on "ltime"
45	ct_src_dport_ltm	No of records of the same "srcip" and " dsport" in "100 records" based on "ltime"
46	ct_dst_sport_ltm	No. of records of the same "dstip" and "sport" in "100 records" based on "ltime"
47	ct_dst_src_ltm	No. of records of the same "srcip" and "dstip" in "100 records" based on the "ltime"

Table-5: Attack Types in UNSW-NB15 dataset

Attack Types
Fuzzers
Analysis
Backdoor
Dos
Exploit
Generic
Reconnaissance
Shellcode
Worm

3.5 Kyoto 2006+ dataset

Kyoto dataset [10] is created from real environment traffic data collected from honey pot over 3 years. It has 24 statistical features, 14 features derived from KDD dataset and 10 from other analysis done on NIDS [11]. Researchers are capable to obtain more accurate results during their evaluation. The description of features in Kyoto dataset are shown in Table -6.

Table-6: Features of Kyoto dataset

Feature	Description
1	Duration
2	Service
3	Source bytes
4	Destination bytes
5	Count
6	Same srv rate
7	Serror rate
8	Srvserror rate
9	Dst host count
10	Dst host srv count
11	Dst host same src port
12	Dst host serror rate
13	Dst host srvserror rate
14	Flag
15	IDS detection
16	Malware detection
17	Ashula detection
18	Label
19	Source IP Address
20	Source Port Number
21	Destination IP Address
22	Destination Port
23	Start Time
24	Duration

3.6 CSCIDS 2017

A reliable and real-world dataset namely CICIDS2017[12] has benign and seven common attack network flows namely Brute Force Attack, Heartbleed Attack, Botnet, DoS Attack, DDoS Attack, Web Attack and Infiltration Attack with 80 features. They used CICFlowMeter to extract the data from pcap file. The label for each flow is FlowID, SourceIP, DestinationIP, SourcePort, DestinationPort, and Protocol.

4. CONCLUSION

Mostly Intrusion detection based on Machine learning and deep learning literatures used the benchmark datasets such as KDD Cup 99, NSL-KDD, UNSW-NB15 , Kyoto and CSCIDS 2017. Most of the datasets used for research lack in real traffic data. Most of the organisations do not release the network traffic due to confidentiality issue. Therefore, there is a huge demand for real time network traffic data. The performance metrics is important in checking the effectiveness of an algorithm. Researchers can develop an efficient IDS only when a real time attack scenario are provided which incorporates innovative attacks.

REFERENCES

[1] K. Kendall, “A Database of Computer Attacks for the Evaluation of Intrusion Detection Systems,” Master Thesis, pp. 12–26, 1999, doi: citeulike-article-id:9077111.

[2]R. P. Lippmann et al., “Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation,” Proc. - DARPA Inf. Surviv. Conf. Expo. DISCEX 2000, vol. 2, 2000, pp. 12–26, doi: 10.1109/DISCEX.2000.821506.

[3] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/ databases/ kddcup 99/kddcup99.html, Ocotber 2007

[4] Lee, W., &Stolfo, S. J. A Framework for Constructing Features and Models for Intrusion Detection Systems (Vol. 3), 2001.

[5] [1] M. Tavallaee and E. B. and W. a. G. A. A. Lu, “‘A detailed analysis of the KDD CUP 99 data set,’ in Computational Intelligence for Security and Defense Applications,” no. Cisda, 2009, pp. 1–6.

[6] O. Atilla and E. Hamit, “A review of KDD99 dataset usage in intrusion detection and machine learning between 2010 and 2015,” PeerJ, 2016, pp. 0–21, doi: 10.7287/PEERJ.PREPRINTS.1954V1.

[7] NSL-KDD data set for network-based intrusion detection systems.” Available on: http://nsl.cs.unb.ca/NSL-KDD/, March 2009

[8] UNSW-NB15 DataSet for Network Intrusion Detection Systems. Available on: http://www.cybersecurity. unsw. adfa.edu.au /ADFA% 20NB15%20Datasets

[9] N. Moustafa and J. Slay, "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set," Inf. Secur. J., vol. 25, no. 1–3, 2016,pp. 18–31, doi: 10.1080/19393555.2015.1125974.

[10]Kyoto University Benchmark Dataset (2009). http://www.takakura.com/Kyoto_data/.

[11] S. S. Sivatha, S. Geetha, and A. Kannan, "Expert Systems with Applications Decision tree based light weight intrusion detection using a wrapper approach," Expert Syst. Appl., vol. 39, no. 1, 2012, pp. 129–141, doi: 10.1016 /j.eswa.2011. 06. 013.

[12] . Sharafaldin, I., Lashkari, A. H. & Ghorbani, A. :A. Toward generating a new intrusion detection dataset and intrusion traffic characterization. pp. 108–116 ICISSP 2018 - Proc. 4th Int. Conf. Inf. Syst. Secur. Priv. .