PERFORMANCE ANALYSIS COMPARISON ON VARIOUS CYBER-ATTACK DATASET BY RELATING A DEEP BELIEF NETWORK MODEL ON AN INTRUSION DETECTION SYSTEM (IDS)

S.Priya [1], Dr.K.Pradeep Mohan Kumar [2]
[1] *Research scholar,SRMIST,Kattankulathur*
[2] *Assistant Professor,SRMIST*

Abstract: In latest eon there are prompt development in Deep learning techniques which is the subset of Machine learning and AI which leads to instinctive innovation of growth in technologies. In general not only in Engineering, researches but also major improvement in medical fields. So deep learning techniques is being applied on emphasizing network security applications.Most observed in network security systems are intruders, which can be viruses,Dos attacks and Penetration among the network makes the difference in the activities of networks.So dynamic methods can be followed to detect and prevent the attack by intruders.In terms, intrusion detection system(IDS) has so many static datasets which was analyzed for traffic alignments. In that aspects for more accuracy and analyzing the deep learning techniques.IDS are classified such as pattern based intrusion detection,time interval based intrusion detection.We focus on antivirus related signature based intrusions. The datasets such as KDDcup 99 and UNSW-NB15 are the pre-existing databases that have variety of patterns.Main focus on generating the False alarm rate (FAR) and nominal IDS using Deep Belief Network (DBF). This DBF identifies the unpredictable and unanticipated cyber-attacks in both static and dynamic methods. A performance analysis using malware IDS datasets such as KDD dataset DARPA/KDDcup, ADFA-LD and NSL-KDD, CICIDS2017,Iot Device Network logs and UNSW-NB15 features are identified and passed into hidden layers by applying a softmax classifier

***Keywords*: Cyber-attacks, Deep Learning, FAR, IDS, KDDcup99, UNSW-NB15**

1. Introduction

Cyber security is the latest technology that has all over attention worldwide. The Wireless networks via cloud vulnerability are becoming the huge issues that make up constitute a threat. The organization are collecting through internet not only the information's required but also all sensitive data's such as date of birth, aadhar number etc. those data's are turned into business usage such as hacking the passwords, Dos attacks, phis icing attacks[2].

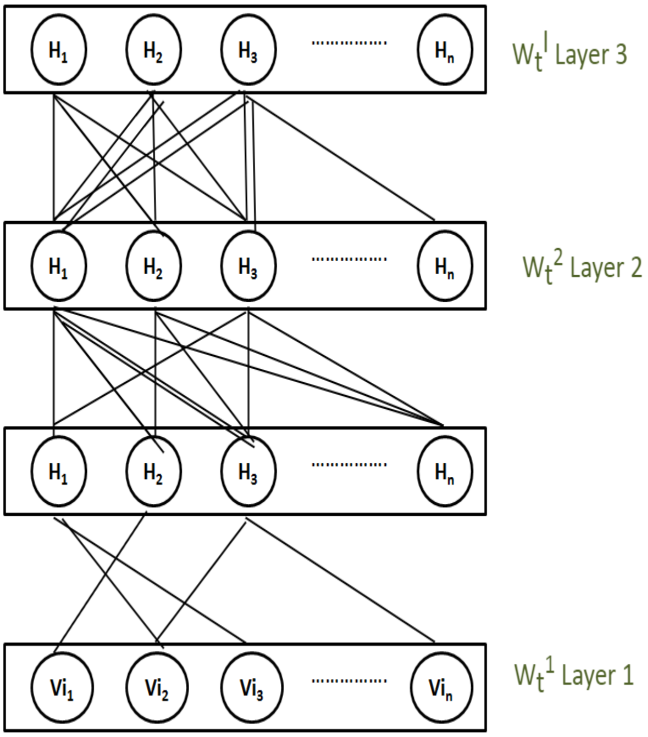


Fig.1.A DBN approach that represents the intrusion features labeled with weights

Figure 1 shows the deep belief networks (DBN) which has representation of direct connections with visible layer as input (vi) that is depicted as propagative explicit model. These visible layer detects the feature which are multiple layer of variable that terms many hidden layers(h1)only and not connected with its units[7].

1.1 A Deep Learning Approach On IDS

The development of all complicated to even simple tasks is changed into advanced techniques such as Artificial intelligence, ML and Deep learning. Since then the results compared to out date techniques predictions are perfect with these methodologies. Network security systems are future focused on all aspects by Engineering field to research till date. The general attacks such as brute force, DOS or viruses

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

also considered as intrusion systems[1]. Even the network every now and then gets update and nowhere the traffic changes frequently or the state of changes that happened every time is recorded as parameters of networks. In such cases we name or mention those as intrusion when some anonymous or malicious changes takes place and monitored in the network. Either the network must be prevented without attacks or there must be detection of the entire systems. There the techniques called Intrusion detection System (IDS)[3]. The techniques such as Deep learning which is a most powerful technique that are shape into a automatic feature extraction that implies the network with sturdy IDS.

The main objective of this paper is to performance analysis of various data sets that are static and dynamic approach on intrusion detection systems in cyber-attacks. The deep learning technique DBF(a generative model) which will analyses the various dataset in IDS[15]. We focus our research work on UNSW-NB15 as input and passes into multiple layer generative model which helps to identify the labeled features by applying the softmax classifier. The further work is based on different parameters from the same database from intrusion detection systems.

The paper is organized as follows: Section 2 deals with the related work, Section 3 discusses about the UNSW-NB15 dataset, Section 4 deals with the discussions and Section 5 concludes the paper.

2. Related Work

Detecting intrusion in systems is further classified as N_T_IDS which is based on networks and also as H_T_IDS which is based on the hosts [9]. Research focus towards the intrusion systems has nowadays involves the solutions based on machine learning approaches. Both deep learning as well as machine learning methods goes hand in hand while dealing with the solutions for intrusions in the system [6].

N_T_IDS (System to detect intrusion based on Networks)

N_T_IDS is one of the important research topic for researchers of today's world. Such systems are keen on usage of metrics such as length of packets, arrival time of packets, size of flow of the packets including several other metrics[5]. The main limitation associated with them is both the false positive as well as false negative seems to be high. The former limitation indicates that attack alerts can be received even at times when such incidents have not taken place. The latter limitation indicates that the system may not be able to identify the attack in an evident way. Such situations make these solutions to become invalid in the context of attacks.

Systems based on self learning comes to rescue in order to avoid the attacks. Such systems can be further classified as supervised as well as unsupervised systems that focus on learning of patterns from the attacks made in the systems. Though machine learning methods are available to solve the issues, still they suffer from certain limitations including the higher cost for computation as well as false positive rates at a higher value. The main reason behind this is the learning capability of TCP/IP attributes by the classifiers which happens locally. The subset of machine learning is deep learning which passes the TCP/IP data to layers which are hidden in the deep learning networks that play a major role in representation of features [14]. Functionalities related to artificial intelligence for processing images as well as recognition of speech and many more are the fields in which deep learning has attained good results [4]. In addition to these, such works have taken a new shape as detecting intrusion, analysis of malware and its classification, predicting the traffic of network, detecting ransomware, categorizing the text that is encrypted, detecting URL which seems to be malicious, and so on[13]. Our work is focused on analyzing the existing methods from machine learning as well as deep learning with the available dataset for N_T_IDS[5].

Our work mainly concentrates on efficacy of the deep learning methods while applying dataset related to IDS such as UNSW-NB15. KDDCUP 99 is yet another dataset predominantly used in IDS systems. But it suffers from certain limitations such as redundancy of data as a result of which bias may occur while detecting. Another limitation may be the data missed at some places as a result of which problems keep on arising. Yet another improvised version of KDDCUP 99 is dataset named as NSLKDD which can handle the missing data problems. But it also suffers from certain limitations when attacks are being encountered [8].

The features extracted in KDDCUP 99 dataset were further divided into three separate groups including vectors of 4 types for attacks. It also included few number in types of attacks both training as well as testing dataset. As the dataset was available publicly, it was used by various research people across the globe. In case of attacks, the data packets's TTL gets affected here. Another limitation is that concerned with the distribution of probability differences occur among both the training as well as testing dataset [16]. Biased decision may occur due to these differences. It also does not perform well in case of attack projections.

NSLKDD, as already mentioned is another version of KDD in an updated manner. Compared to KDDCUP 99 it was designed to handle several characteristics such as elimination of duplicates, selecting several data records at the same time and also deals with the elimination of problems related to unbalance in both the training as well as testing dataset. Due to these features, FAR named as False Alarm Rate gets decreased [9]. Also NSLKDD does not represent attack projections.

3. UNSW-NB15 Dataset

IXIA tool was used to create the dataset UNSW-NB15 mainly for extracting the attack related works in a traffic. Its features along with attacks as well are discussed in detail.Attacks in UNSW-NB15 includes 9 types as depicted in Fig.2.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Fuzzers
Analysis
Backdoor
DoS
Exploit
Generic
Reconnaissance
Shellcode
Worm

Fig.2. Attack Types

Features of the dataset is discussed in detail along with the purpose of the features (Fig.3).

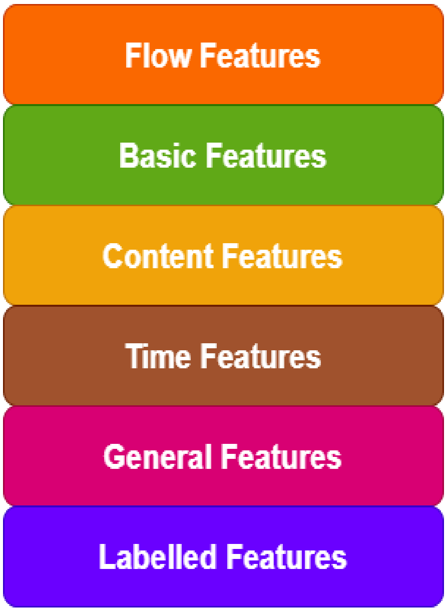
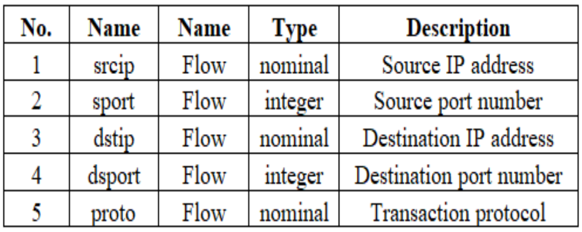


Fig.3 UNSW-NB15 Dataset Features

3.1 Features Related to Flow

Table 1 shows the features related to flow. These features mainly focus on the IP address of the source as well as destination along with the port number.

Table 1. Features Related to Flow

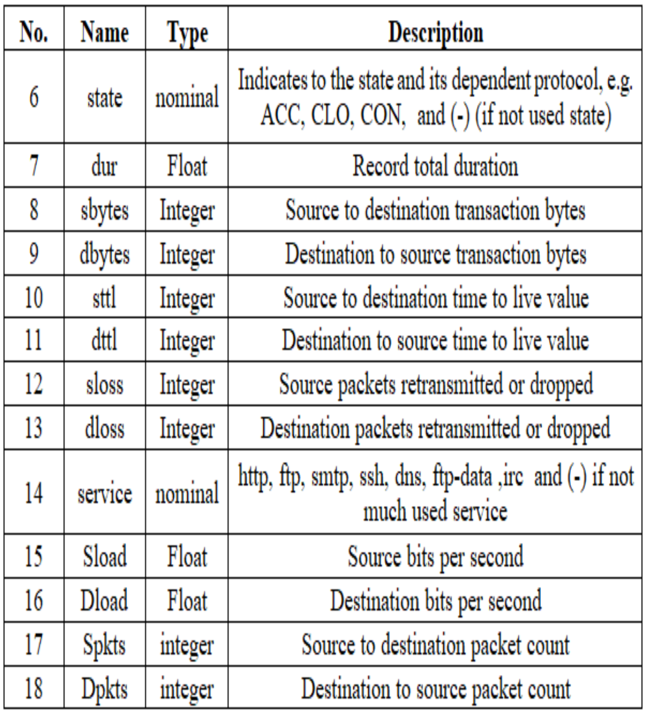


No.	Name	Name	Type	Description
1	srcip	Flow	nominal	Source IP address
2	sport	Flow	integer	Source port number
3	dstip	Flow	nominal	Destination IP address
4	dsport	Flow	integer	Destination port number
5	proto	Flow	nominal	Transaction protocol

3.2 Basic Features

The features related to TTL both on source and destination side along with count of packets during transmission are taken as basic features.

Table 2 Basic Features

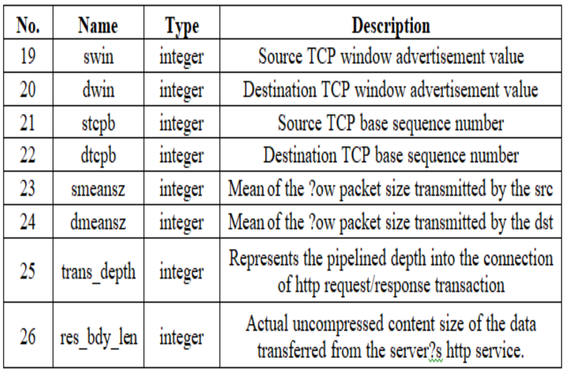


No.	Name	Type	Description
6	state	nominal	Indicates to the state and its dependent protocol, e.g. ACC, CLO, CON, and (-) (if not used state)
7	dur	Float	Record total duration
8	sbytes	Integer	Source to destination transaction bytes
9	dbytes	Integer	Destination to source transaction bytes
10	sttl	Integer	Source to destination time to live value
11	dttl	Integer	Destination to source time to live value
12	sloss	Integer	Source packets retransmitted or dropped
13	dloss	Integer	Destination packets retransmitted or dropped
14	service	nominal	http, ftp, smtp, ssh, dns, ftp-data ,irc and (-) if not much used service
15	Sload	Float	Source bits per second
16	Dload	Float	Destination bits per second
17	Spkts	integer	Source to destination packet count
18	Dpkts	integer	Destination to source packet count

3.3 Features related to Content

The features are mostly created on generation of header packet and Table 3 shows both source as well as destination TCP window related features[17].

Table 3 Features related to Content

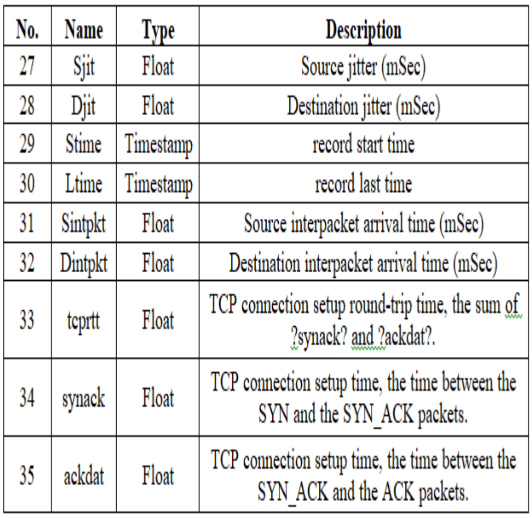


No.	Name	Type	Description
19	swin	integer	Source TCP window advertisement value
20	dwin	integer	Destination TCP window advertisement value
21	stcpb	integer	Source TCP base sequence number
22	dtcpb	integer	Destination TCP base sequence number
23	smeansz	integer	Mean of the ?ow packet size transmitted by the src
24	dmeansz	integer	Mean of the ?ow packet size transmitted by the dst
25	trans_depth	integer	Represents the pipelined depth into the connection of http request/response transaction
26	res_bdy_len	integer	Actual uncompressed content size of the data transferred from the server?s http service.

3.4 Features related to Time

Table 4 shows jitter related features which are categorized in relation with time for both source and destination [12]. Till the features related to time, all are created from the packets through which data is sent.

Table 4 Features related to Time

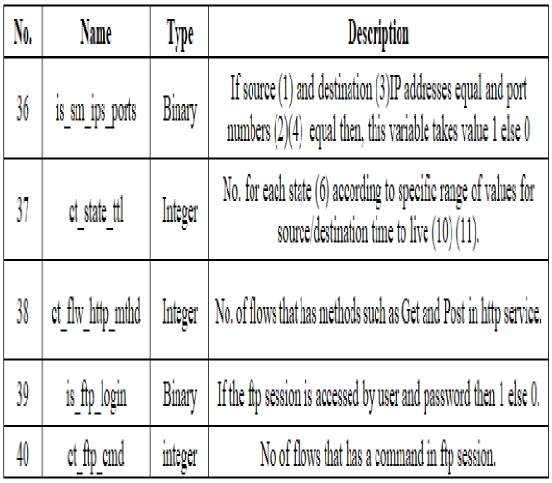


No.	Name	Type	Description
27	Sjit	Float	Source jitter (mSec)
28	Djit	Float	Destination jitter (mSec)
29	Stime	Timestamp	record start time
30	Ltime	Timestamp	record last time
31	Sintpkt	Float	Source interpacket arrival time (mSec)
32	Dintpkt	Float	Destination interpacket arrival time (mSec)
33	tcprtt	Float	TCP connection setup round-trip time, the sum of ?synack? and ?ackdat?.
34	synack	Float	TCP connection setup time, the time between the SYN and the SYN_ACK packets.
35	ackdat	Float	TCP connection setup time, the time between the SYN_ACK and the ACK packets.

3.5 General Features

Features are generated additionally and further classified as those which help in general works and those which defend during connection issues. Table 5 shows features that serve several works related to any other general purpose [12].

Table 5 General Features

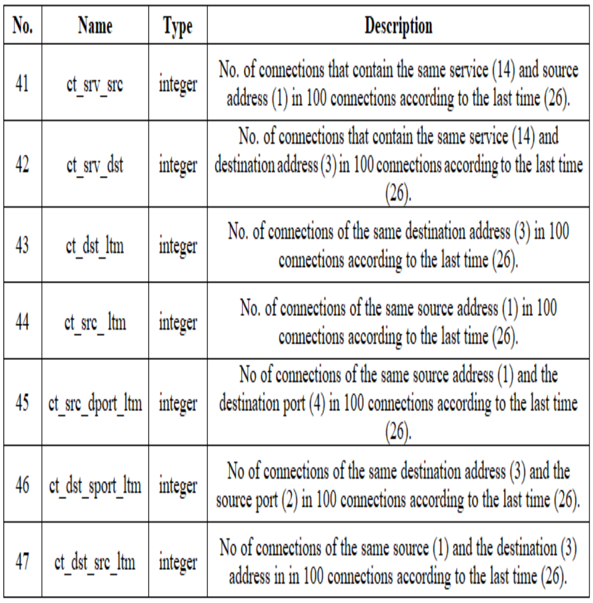


No.	Name	Type	Description
36	is_sm_ips_ports	Binary	If source (1) and destination (3)IP addresses equal and port numbers (2)(4) equal then, this variable takes value 1 else 0
37	ct_state_ttl	Integer	No. for each state (6) according to specific range of values for source/destination time to live (10) (11).
38	ct_flw_http_mthd	Integer	No. of flows that has methods such as Get and Post in http service.
39	is_ftp_login	Binary	If the ftp session is accessed by user and password then 1 else 0.
40	ct_ftp_cmd	integer	No of flows that has a command in ftp session.

3.6 Feature related to connection

The second category in case of additional feature is those related to connections, that defend in case of scenarios where connections are attempted. Table 6 shows connection related features.

Table 6 Feature related to connection

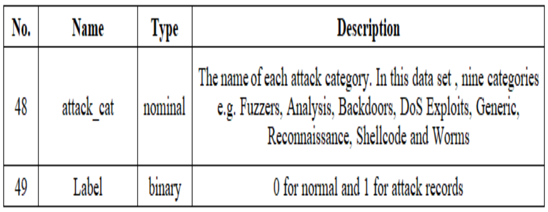


No.	Name	Type	Description
41	ct_srv_src	integer	No. of connections that contain the same service (14) and source address (1) in 100 connections according to the last time (26).
42	ct_srv_dst	integer	No. of connections that contain the same service (14) and destination address (3) in 100 connections according to the last time (26).
43	ct_dst_ltm	integer	No. of connections of the same destination address (3) in 100 connections according to the last time (26).
44	ct_src_ ltm	integer	No. of connections of the same source address (1) in 100 connections according to the last time (26).
45	ct_src_dport_ltm	integer	No of connections of the same source address (1) and the destination port (4) in 100 connections according to the last time (26).
46	ct_dst_sport_ltm	integer	No of connections of the same destination address (3) and the source port (2) in 100 connections according to the last time (26).
47	ct_dst_src_ltm	integer	No of connections of the same source (1) and the destination (3) address in in 100 connections according to the last time (26).

3.7 Features related to label

Features used to label can help in assigning values 0 or 1 and also attack_cat deals with the categories of attack.

Table 7 Features related to label



No.	Name	Type	Description
48	attack_cat	nominal	The name of each attack category. In this data set , nine categories e.g. Fuzzers, Analysis, Backdoors, DoS Exploits, Generic, Reconnaissance, Shellcode and Worms
49	Label	binary	0 for normal and 1 for attack records

4. Discussions

The novel attacks such as static and anomaly detection are general in IDS, the network traffic captures raw traffic and identified as features for configuring class labels that are in parallel processing and stored as CSV files. There are major attacks such as fuzzers, analyzer, denial of service attack, backdoor,shellcode,worm are considered and they are observed with its class label with its instances[11].According to the DB model the training and test process are undergone with its authentication for assessments. This helps to remove the repeated data's in both 60% of training data and 40% of test data[18].

4.1 Comparison of Various datasets in IDS

Table defines the various datasets in intrusion detection systems with its features, establishment of datasets year, online availability and its data capacity. IDS analysis with various data determines the feature correlations and avoid unit based interaction in the hidden layers.

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

Table 8 :Comparison of Various datasets in IDS

Static & Dynamic Methods	**Datasets In Intrusion Detection Systems**			
Name Of Datasets	**Establishment Year**	**No.Of Features**	**Online Data Existing**	**Data Capacity**
KDDCUP99	1998	99	yes	5m points
DARPA	1999	54	yes	Not specified
TWENTE	2008	71	yes	14M flows
UNIBS	2009	79	yes	79k flows
ISCX2012	2012	28	yes	2M flows
NGIDS-DS	2016	19	yes	1M packets
UNSW-NB15	2015	49	yes	2M points
CICIDS2017	2017	27	yes	3.1M flows

5. Conclusion:

In this paper, the performance analysis of IDS datasets have been represented along with its feature analysis by comparing various datasets which is applied over the deep belief network along with Sigmoid classifier[10].To highlight our focus on UNSW-NB15 datasets the most known attacks such as state are abnormal changes[8]. There are features compared from different datasets in intrusion system to show the attacks and information of packets which are not in use now.In future this can be classified based on the DBF model and applied classifier sigmoid which helps to identify the exact behavior of each features at its variables find the appropriate modeling

References

[1] Mukherjee, B., Heberlein, L. T., & Levitt, K. N. (1994).Network intrusiondetection.IEEE network, 8(3), 26-41.

[2] Larson, D. (2016). Distributed denial of service attacks-holding back theflood. Network Security, 2016(3), 5-7.

[3] Staudemeyer, R. C. (2015). Applying long short-term memory recurrentneural networks to intrusion detection. South African Computer Journal,56(1), 136-154.

[4] Venkatraman, S., Alazab, M. "Use of Data Visualisation forZero-Day Malware Detection," Security and CommunicationNetworks, vol. 2018, Article ID 1728303, 13 pages, 2018.https://doi.org/10.1155/2018/1728303

[5] Mishra, P., Varadharajan, V., Tupakula, U., & Pilli, E. S. (2018). Adetailed investigation and analysis of using machine learning techniquesfor intrusion detection. IEEE Communications Surveys & Tutorials.

[6] Azab, A., Alazab, M. &Aiash, M. (2016) "Machine Learning BasedBotnet Identification Traffic" The 15th IEEE International Conference onTrust, Security and Privacy in Computing and Communications (Trustcom2016), Tianjin, China, 23-26 August, pp. 1788-1794.

[7] Vinayakumar R. (2019, January 19). vinayakumarr/Intrusion-detection v1(Version v1). Zenodo. http://doi.org/10.5281/zenodo.2544036

[8] Tang, M., Alazab, M., Luo, Y., Donlon, M. (2018) Disclosure of cybersecurity vulnerabilities: time series modelling, International Journal ofElectronic Security and Digital Forensics. Vol. 10, No.3, pp 255 - 275.

[9] V. Paxson. Bro: A system for detecting network intruders in realtime.Computer networks, vol. 31, no. 23, pp. 24352463, 1999. DOIhttp://dx.doi. org/10.1016/S1389-1286(99)00112-7

[10] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning.nature,521(7553), 436.

[11] Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., ...&Wang, C. (2018).Machine Learning and Deep Learning Methods for Cybersecurity. IEEEAccess.

[12] Hofmeyr, S. A., Forrest, S., &Somayaji, A. (1998). Intrusion detectionusing sequences of system calls. Journal of computer security, 6(3),151180.

[13] Forrest, S., Hofmeyr, S. A., Somayaji, A., &Longstaff, T. A. (1996,May). A sense of self for unix processes. In Security and Privacy, 1996.Proceedings., 1996 IEEE Symposium on (pp. 120-128). IEEE.

[14] Hubballi, N., Biswas, S., & Nandi, S. (2011, January). Sequencegram:n-gram modeling of system calls for program based anomaly detection.In Communication Systems and Networks (COMSNETS), 2011 ThirdInternational Conference on (pp. 1-10). IEEE.

[15] Hubballi, N. (2012, January). Pairgram: Modeling frequency informationof lookahead pairs for system call based anomaly detection. In CommunicationSystems and Networks

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731

(COMSNETS), 2012 Fourth InternationalConference on (pp. 1-10). IEEE.

[16] Kozushko, H. (2003). Intrusion detection: Host-based and network-basedintrusion detection systems. on September, 11.

[17] W. Lee and S. Stolfo. A framework for constructing features and modelsfor intrusion detection systems.ACM transactions on information andsystem security, vol. 3, no. 4, pp. 227261, 2000.DOI http://dx.doi.org/10.1145/382912.382914

[18] Ozgur, A., Erdem, H.: A review of KDD99 dataset usage in intrusiondetection and machine learning between 2010 and 2015. PeerJPrePrints4 (2016) e1954

1S.PRIYA,
RESEARCH SCHOLAR,
DEPARTMENT OF CSE,
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY,
spriyampk@gmail.com

2Dr.K.Pradeep Mohan Kumar,
ASSISTANT PROFESSOR,
DEPARTMENT OF CSE,
SRM INSTITUTE OF SCIENCE AND TECHNOLOGY,
pradeepk@srmist.edu.in

ISSN (Print): 2204-0595
ISSN (Online): 2203-1731