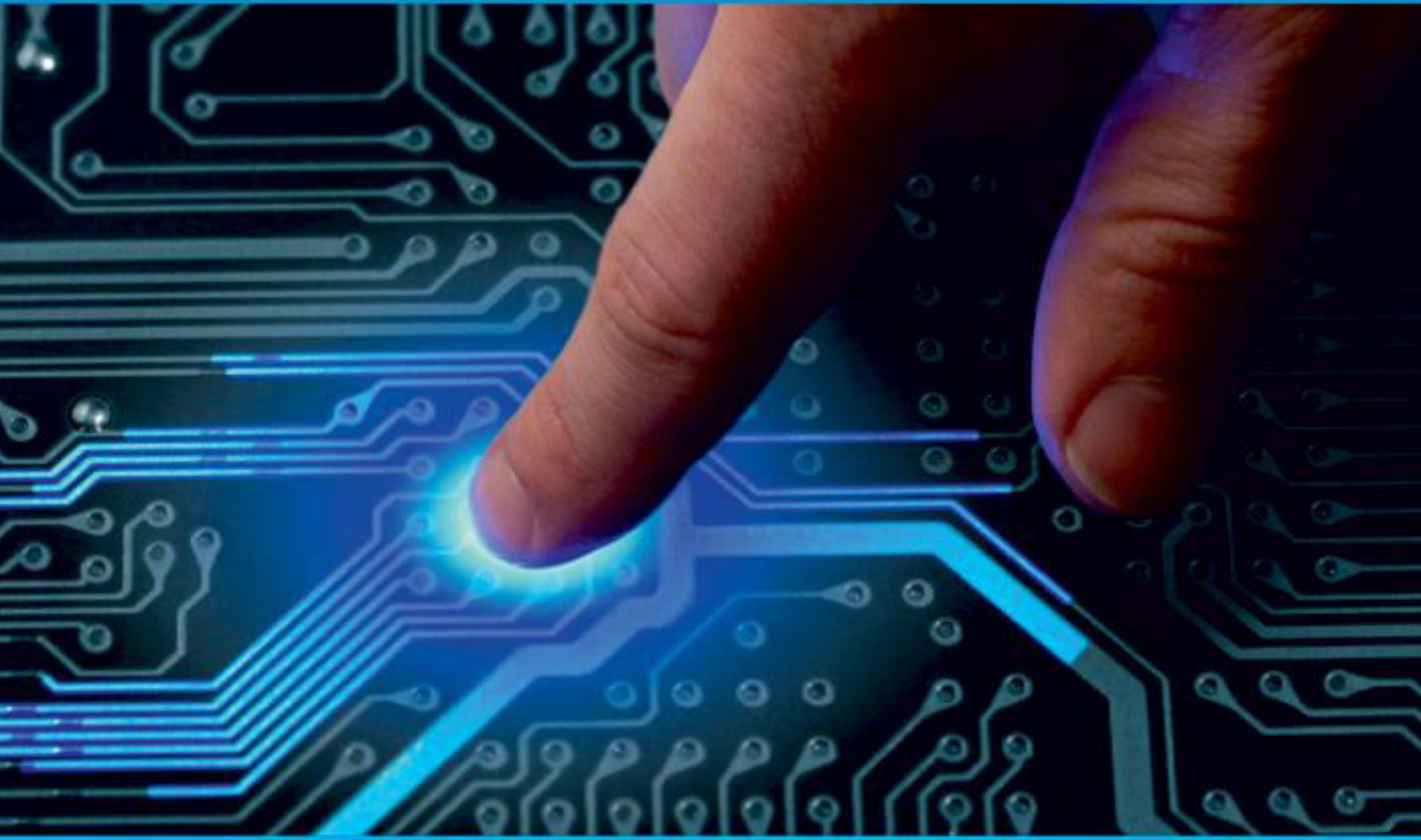
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 10, October 2021

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 7.542

| DOI: 10.15680/IJIRCCE.2021.0910012 |

Classification of Intrusion Detection System Dataset: A Study

Bheem Singh Saini [1], Seema Rani Gadai[2]

M.Tech Research Scholar, Department of Computer Science, Keystone Group of Institution, Surajgarh, Rajasthan, India [1]

Assistant Professor, Department of Computer Science, Keystone Group of Institution, Surajgarh, Rajasthan, India[2]

ABSTRACT: The basic purpose of an IDS (Intrusion Detection System) is to secure the system by analyzing and anticipating user activity. These behaviors will then be classified as either an attack or a normal response. Since the technology's inception in the mid-1980s, researchers have been working to improve the ability to detect attack without sacrificing network speed. As the popularity of the internet grows among users across the globe, so does the need of maintaining security and keeping the system informed of dangerous activity. The major goal of this work is to provide a comprehensive analysis of intrusion detection, including different types of attacks and finally the development of an IDS tool for research purposes that tool is capable of detecting and preventing intruder intrusion.

KEYWORDS: Intrusion detection system, Classification , DARPA 1999, DARPA 2000, KDD Cup 99, NSL-KDD, GureKDD, Anomaly detection, IDS dataset

I. Introduction

We can monitor network traffic and unauthorised and suspicious behaviour in the network using an intrusion detection system, and once the information about an attack is discovered, we can notify the network administrator and the system. When we discover an attack, we obtain the system's IP address and communicate it to the network administrator, who will subsequently terminate or break the network connection and save the machine. Administrators have access to the attacker's records and administer the table using a white and black box list. Administrators have the ability to suspend or terminate a connection. There are a variety of techniques available for detecting intruders in the network. We can determine the types of attacks using the dataset. The intruder dataset is the KDD99 / NSLKDD 99 dataset. The DARPA data was gathered at MIT Lincoln Labs.

DARPA held an online competition in 1998 at MIT Lincoln Lab to discover different sorts of attacks possible in computer networks on different – 2 systems (i.e. UNIX/LINUX). DARPA has set up a platform for participation at MIT Lincoln Lab (sponsored by DARPA) [30]. DARPA 1998 contains approximately 4 GB of compressed raw TCP dump data from seven weeks of network traffic. This will be broken down into 5 million 100-byte connection records. The KDD 1999 training dataset was used to create a model for detecting computer network intruders, with the goal of creating the most efficient model for detecting all sorts of attacks. This is a raw TCP dump dataset. This data was gathered during a nine-week period on the Local Area Network (LAN). The training dataset [29] was split into five million records based on seven weeks of network traffic and two million records based on two weeks of testing data. There are 41 features that are either normal or attack [31].

Intrusion detection system (IDS) is a kind of security management system for computers systems and networks. An Intrusion Detection System gathers the information from certain areas within a network or computers and analyzes it to find potential security breaches, that contain the both intrusions (outdoor attacks) and misuse (indoor attacks) [1]. The need of security problem for the data has been increasing every day along with the rapid development of the computer network. Security means degree of protection given to the network or system. The primary goal of security are confidentiality, availability and integrity [2]. Attacks on network also be known as intrusion. Intrusion implies that any set of malicious programs that try to cause the security goals of the important information. IDS assist the system in resisting outside attacks. Intrusion detection system gathers data through the network, then monitors and analyzes this data and after that separate it into malicious & normal programs, produce the result to the system administrator [3].

An IDS monitor all internal and external network event and also detect suspicious behavior that may possibly show a network or system attack from someone trying to break into or even cause a system. IDS primary design and function is to protect the organization's important information from an intruder. The IDS analyzes the collected data from different sources and compares it to wide databases of attack signatures.

An intrusion detection system (IDS) is used as a tool to identify unauthorized intrusions or malicious programs i. e. various attacks placed into computer systems and networks. These types of system are often tend to generate alerts or signify the area where intrusions are placed. The following common terms used for detection and identification of attack and normal behaviour. Figure 1 shows anomaly detection process

1. True positive (TP): Detection of attack when its correctly labeled as attacked;
2. True negative (TN): Detection of normal when its correctly labeled as normal;
3. False positive (FP): Detection of attacks when its correctly labeled as normal called as false alarm;
4. False negative (FN): Detection of normal when its correctly labeled as attacked.

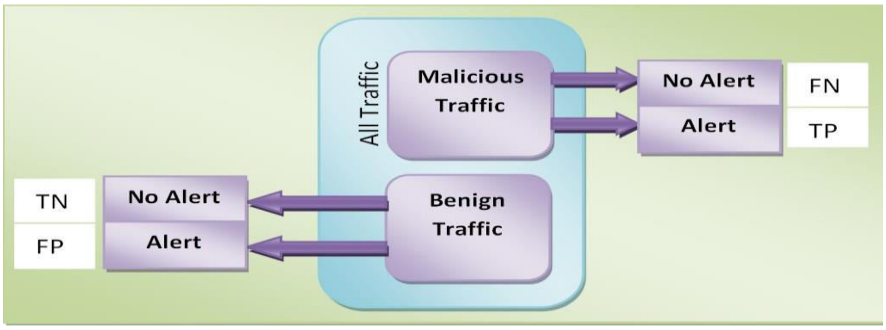


Figure 1. Anomaly Detection Process

Intrusion detection aspects consist of:

- Analyzing and monitoring system and user's behavior.
- Analyzing system configurations and vulnerabilities.
- Analyze file integrity and system.
- Capability to identify typical pattern of attacks.
- Analysis of anomalous activity patterns.
- Monitoring user policy violations [4].

Intrusion detection systems are intentionally mounted on a network to recognize threats and track packets. The IDS carry out this by gathering information from number of network and system sources and analyzing the data for potential threats [5]. The functions of the IDS providing information on threats, taking out corrective measures whenever it identify threats and capturing important activities within a network [6]. Figure 2 shows a Intrusion detection system model.

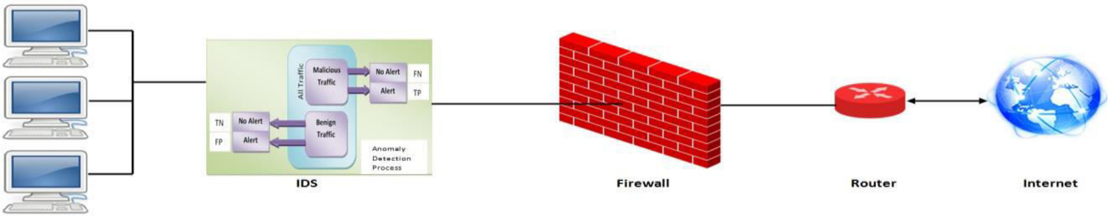


Figure 2. Intrusion Detection System

Protecting networks and computer security, attacks is a apprehension of computer security. Sensitive and confidential data transfer and information exchange is the part of a network traffic that leads open way to attacks. Although it's too well known that the dependency of network are also rising rapidly. Because of this, the network problem are very critical now a days and it will become more complex in coming time. This traffic may result in massive damage of network system and its related resources. To detect malicious and unwanted attacks, anomaly detection is a technique to analyze the network traffic on the basis of traffic pattern.[7].

Network behaviours that cannot be specify by any model for such situations non-model based procedures are primarily used. Non-model based procedures can be additional categorization based on the unambiguous implementation and accuracy constraints which have been imposed on the detection system. Malicious activity can be detected by analyzing the identity of intrusion in Misuse Detection. Misuse detection technique monitor and analyze host or network activity, looking out for events that match patterns of known (signatures) attacks. Initially a reference database of attack signatures is built, and then monitored various activities from sensors data are compared against this attack database (signature) for evidence of intrusions [8].

Attack Class	Attack Type (57)
Probe	portsweep, queso, msscan, lsdomain, illegal-snifer, ipsweep ntinfoscan, satan,
DoS	selfping, dosnuke, back, tcpreset, syslogd, arppoison, mailbomb, teardrop, processtable, neptune, udpstorm, land, warezclient, apache2, crashiis, smurf, pod
R2L	imap, xlock, sshtrojan, ppmacro, netbus, sendmail, snmpget, ncftp, httptunnel, xsnoop, named, dict, framespoof, netcat, guest, ftpwrite, phf
U2R	sechole, ps, secret, perl, fdformat, casesen, ntfsdos, yaga, ppmacro, eject, loadmodule, nukepw, sqlattack, xterm, ffbconfig,

TABLE I

DARPA 1999 DATASET ATTACK CLASS AND TYPES

We usually used DARPA1999 dataset for performance evaluation of anomaly detection system [9]. The DARPA Dataset was generated for the analysis purpose of network security through data centric intrusion detection. The KDD Cup1999 Dataset was created by processing the tcpdump parts of the DARPA1998 Intrusion Detection System assessment dataset [10].

This paper gives detail about following area. Section 1 gives detail Introduction about Intrusion detection system. Section 2 gives brief description of the various intrusion detection system datasets and the next section is of conclusion.

II. INTRUSION DETECTION DATASET DESCRIPTION

The datasets play a vital function in the testing and validation of the anomaly detection method in networks or system. A decent quality dataset not only allows us to detect the capability of a technique or a system to find abnormal behavior, however additionally permit us to provide potential effectiveness when deployed in real operating environments [11].

DARPA

DARPA datasets (1999 and 2000) generated in MIT Lincoln Laboratories. The Dataset is created by introducing manually generated network based attacks [14]. The classification of the different attacks discovered within the network traffic is defined in detail [12] with regards to DARPA intrusion detection assessment dataset[13].

DARPA 1999

The test data of the DARPA1999 included 190 samples of the 57 attacks which included 8 Probes, 17 DoS attacks, 17 R2L attacks and 15 U2R attacks with details of attack types given in Table I [10].

The attacks classified into four main classes specifically, Denial of Service attack (DoS), Probe attack, User to Remote attack (U2R) and Remote to local attack (R2L).

The probe attacks automatically scan a system or network in attempt [11] to accumulate records of private systems or a DNS server to locate legitimate IP addresses (ipsweep, mscan, lsdomain), host operating system sorts (mscan, queso) active ports (mscan, portsweep), and recognized vulnerabilities (satan) [10].

The DoS attacks are intend to confuse a host or network service toward off valid users from using a service provided by the system [11]. These consist of the Solaris operating system crash (selfping), actively terminate all TCP connections for a particular host (tcpreset), corrupt ARP cache entries for a victim not in others caches (arppoison), crash the web server Microsoft Windows NT (crashiis) and crash Windows NT (dosnuke) [10].

In R2L attacks, [15] an attacker who does not have an account or any access on a victim machine and takes benefits of bugs or weakness in machine to gains local access to the machine (guest, dict), remove files from the machine (ppmacro) or changes data in transit to the machine(framespoof). New R2L attacks include an a webbrowser attack called a man-in-middle (framespoof), NT power point macro attack (ppmacro), a Linux trojan SSHserver (sshtrojan), an NT trojan-installed remote administration tool (netbus) and a version of a Linux FTP file access-utility with a bug that permits remote commands to run on a local machine (ncftp) [10].

In U2R attacks, a local user on a system has the ability to acquire privileges usually available for the unix super user or Windows NT administrator. The Data attack is to remove special files which the security policy specifies and need to stay with the victim hosts. These include secret attacks, where a user who is authorized to get right of entry to the special files removes them (ntfsdos, sqlattack) [10].

DARPA 2000

Two attack situations were simulated in the DARPA 2000 assessment contest, namely Lincoln Laboratory scenario DDoS (LLDOS) 1.0 and LLDOS 2.0. To gain variations, these two attack scenarios had been completed over numerous network and audit scenarios.

It contains four separated files which constitute two forms of simulated scenarios (Scenario One and Scenario Two) of Distributed Denial of Services (DDoS) network attack on two distinct networks. (a)probing, (b)breaking into the machine with the aid of exploiting vulnerabilities, (c) the installation of DDoS software application for the compromised system so (d) launching DDoS attack against a different target. LLDOS 2.0 scenarios varying from LLDOS 1.0 scenario in that attack in Scenario Two were stealthier than Scenario One [16].

KDD CUP '99

KDD Cup'99 intrusion detection datasets that are based totally on DARPA '98 dataset [17] provides labelled dataset for researcher running within the area of intrusion detection and represent the publicly available labelled dataset. The detailed description of KDD dataset is given in the next phase. The KDD'99 dataset is created the usage of a simulation of an army network. In the end, there is a sniffer which records all transmitted network traffic data by using the Tcpdump format. KDD training [18] dataset contains around 4,900,000 single connection vectors, every one of which includes 41 attributes and is categories as either an attack or normal, with precisely one specified attack type. The simulated attacks classified amongst the subsequent four classes: Denial of Service (Dos), Probe, Remote to Local (r2l) and User to Root (u2r) attacks [19]. Features are labelled into four listed types:

- **Basic Features:** These characteristics tend to be derived from packet headers while no longer analyzing the payload.
- **Content Features:** To analyze the actual TCP packet payload, Domain knowledge is used and this encompasses features which includes the large variety of unsuccessful login attempts.
- **Time-based Traffic Features:** These features are created to acquire properties accruing over a 2 second temporal window. An example of such a feature will be the wide range of connections to the exact same host over the interval of 2 second.
- **Host-based Traffic Features:** Make use of a historical window calculated over the number of connections and in this case it is 100. Thus Host based attributes are created to analyze attacks, which time frame longer than 2 seconds [20].

There are 41 features for each and every TCP/IP connection, 41 different quantitative (continuous data type) and qualitative (discrete data type) features were extracted among the 41 attributes, 34 attributes (numeric) and 7 attributes (symbolic) [23], which are mentioned in Table II.

TABLE II

DESCRIPTION OF KDD CUP 99 FEATURES TYPE AND ATTACK

No.	Feature Name	Type	Feature Description	Class
1	Duration	Continuous	Duration of the connection	Normal
6	Dst_Bytes	Continuous	Bytes sent by the target to source	
12	Logged_In	Discrete	1 if successfully logged in; Otherwise 0	
15	Su_Attempted	Continuous	1 if the command "su root" attempted; otherwise 0	
16	Num_Root	Continuous	Number of accesses "root"	
17	Num_File_Creations	Continuous	Number of file creation operations	

18	Num_Shells	Continuous	Number of requests for shell	
19	Num_Access_Files	Continuous	Number of transactions in access control files	
31	Srv_Diff_Host_Rate	Continuous	% of connections to different hosts	
32	Dst_Host_Coun	Continuous	Count of connections having the similar destination host	
37	Dst_Host_Srv_Diff_Host_Rate	Continuous	% of connections to the same service from different hosts	
4	Flag	Discrete	Connection Status flag	Smurf
25	Serror_Rate	Continuous	% of connections that have “SYN”errors	
26	Srv_Error_Rate	Continuous	% of connections that have “SYN”errors	
29	Same_Srv_Rate	Continuous	% of connections to the same Service	
30	Diff_Srv_Rate	Continuous	% of connections to different Services	
33	Dst_Host_Srv_Count	Continuous	Count of connections having the same destination host and service	
34	Dst_Host_Same_Srv_Rate	Continuous	% of connections having the same destination host and service	
35	Dst_Host_Diff_Srv_Rate	Continuous	% of different services on the current host	
38	Dst_Host_Serror_Rate	Continuous	% of connections to the current host presenting an error S0	
39	Dst_Host_Srv_Serror Rate	Continuous	% of connections to the current host and particular service that have an S0 error	
2	Protocol_Type	Discrete	Connection protocol (e.g. TCP, UDP, ICMP)	Neptune
3	Service	Discrete	Destination service	
5	Src_Bytes	Continuous	Bytes sent from the source to the target	
23	Count	Continuous	count number of connections to the same host as the current connection in the past two seconds	
24	Srv_Count	Continuous	Count Number of connections to the same service as the current connection in the past two seconds	
27	Rerror_Rate	Continuous	% of connections that have REJerrors	
28	Srv_Ressor_Rate	Continuous	% of connections that have REJerrors	
36	Dst_Host_Same_Src_Port_Rate	Continuous	% of connections to the current host have the same src port	
40	Dst_Host_Rerror_Rate	Continuous	% of connections to the current host which have an RST error	
41	Dst_Host_Srv_Rerror Rate	Continuous	% of connections to the current host and particular service which have an RST error	
10	Hot	Continuous	Numbers of “hot” indicators	Back
13	Num_Compromised	Continuous	Numbers of condition “compromised”	
7	Land	Discrete	1 if the connection is from/to the port/same host; otherwise 0	Land
8	Wrong Fragment	Continuous	Number of wrong fragments	Terdrop
9	Urgent	Continuous	Numbers of urgent packets	Ftp_Write
11	Num_Failed_Logins	Continuous	Number of failed logins	Guess_Pwd
14	Root_shell	Continuous	1 if root shell is obtained; Otherwise 0	Buffer_Overflow
22	Is_Guest_Login	Discrete	1 if the login is the “guest” login; otherwise 0	Warezclient

The KDD cup 99 intrusion detection dataset made up of three parts, which are illustrated in Table III. In the International Knowledge Discovery and Data Mining Tools Competition, only "10% KDD" dataset is used for the purpose of training [21]. It is a concise form of "Whole KDD". This dataset contain mainly 22 attack types and they are mostly of denial of service (DoS) category. It shows more number of attack than normal. Whereas "Corrected KDD" dataset provides a dataset with different statistical distributions compared to "10% KDD" or "Whole KDD". It contains 37 type of attacks. Table 3 gives number of instances in each attack category.

TABLE III

NUMBER OF SAMPLES IN KDD 99 DATASET

Dataset	Normal	Dos	Probe	R2L	U2R
KDD Corrected	60593	229855	4166	16345	70
10% KDD	97278	391458	4107	1126	52
WholeKDD	972780	3883370	41102	1126	52

Corrected KDD and 10%KDD has been analyze which shows that there are 37 and 22 types of attacks in the datasets with varying percentage of different attacks which is shown in Table IV and Table V.

TABLE IV

ATTACK FREQUENCY IN KDD CORRECTED

Attack Type	Value	Count	Percent
Normal	normal.	60593	19.48%
Dos	smurf.	164091	52.76%
	pod.	87	0.03%
	apache2.	794	0.26%
	udpstorm.	2	0.00%
	processtable.	759	0.24%
	neptune.	58001	18.65%
	back.	1098	0.35%
	worm.	2	0.00%
	mailbomb.	5000	1.61%
	teardrop.	12	0.00%
	land	9	0.00%
Probe	ipsweep.	306	0.10%
	saint.	736	0.24%
	portsweep.	354	0.11%
	satan.	1633	0.53%
	mscan.	1053	0.34%
	nmap.	84	0.03%
R2L	snmpgetattack	7741	2.49%
	named.	17	0.01%
	xlock.	9	0.00%
	multihop.	18	0.01%
	xsnoop.	4	0.00%
	sendmail.	17	0.01%
	guess_passwd.	4367	1.40%
	phf.	2	0.00%
	warezmaster.	1602	0.52%
	imap.	1	0.00%

	httptunnel.	158	0.05%
	ftp_write.	3	0.00%
	snmpguess.	2406	0.77%
U2R	buffer_overflow.	22	0.01%
	perl.	2	0.00%
	xterm.	13	0.00%
	ps.	16	0.01%
	rootkit.	13	0.00%
	loadmodule.	2	0.00%
	sqlattack.	2	0.00%
Total		**311029**	**100%**

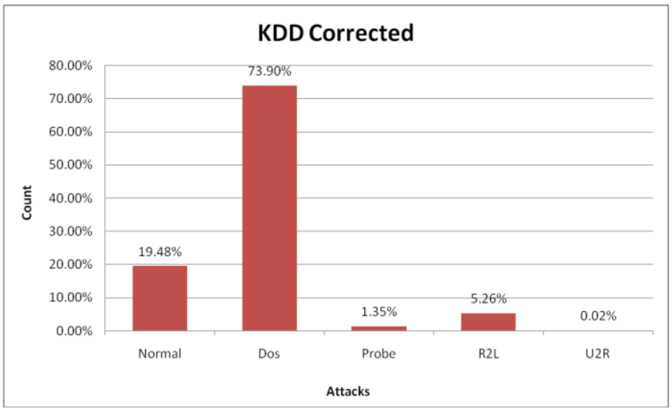


Figure 3. Statistics for Normal and Attack type in KDD corrected

TABLE V

ATTACK FREQUENCY IN 10%KDD

Attack Type	Value	Count	Percent
Normal	normal.	97278	19.69%
Dos	smurf.	280790	56.84%
	pod.	264	0.05%
	neptune.	107201	21.70%
	back.	2203	0.45%
	teardrop.	979	0.20%
	land	21	0.00%
Probe	ipsweep.	1247	0.25%
	portsweep.	1040	0.21%
	satan.	1589	0.32%
	nmap.	231	0.05%
R2L	multihop.	7	0.00%
	guess_passwd.	53	0.01%
	phf.	4	0.00%
	warezclient.	1020	0.21%
	warezmaster.	20	0.00%
	imap.	12	0.00%
	spy.	2	0.00%
	ftp_write.	8	0.00%

U2R	buffer_overflow.	30	0.01%
	perl.	3	0.00%
	rootkit.	10	0.00%
	loadmodule.	9	0.00%
Total		**494021**	**100%**

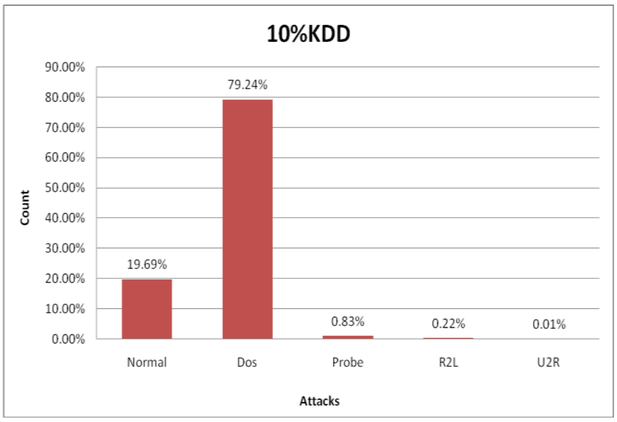


Figure 4. Statistics for Normal and Attack type in 10%KDD

NSL KDD

The NSL KDD dataset is offline network data based totally on KDD'99 dataset [22]. The NSL-KDD data set trained to solve many different immanent issues of the KDDCUP'99 data set. KDD CUP'99 is said to be broadly used data set for anomaly detection [24] for locating accuracy in intrusion detection [25]. The deficiency found in the KDD CUP'99 [17] data set is the extensive quantity of duplicate record of approximately 78% in train set and 75% in test set, respectively. Which makes the learning algorithm rule biased, that makes U2R much more vulnerable to the network. To resolve these types of problems, a new edition of KDD dataset NSL-KDD is offered.

Advantages of NSL-KDD dataset over the original KDD dataset:

- NSL-KDD dataset contains no duplicate data within the train set, then the classifiers do not produce the result biased.
- The proposed test sets does not contain any duplicate records, due to this the learners' performance will not be prevented and gives better detection rates.
- The small number of record selected by each level of difficulty is inversely proportional to the proportion of records in the KDD dataset.
- The dataset contains a reasonable number of samples by train as well as test sets, that makes it convenient to run experiments on complete sets without any requirement to randomly consider a small part [26].

Number of datasets available in NSL-KDD, which consist of two parts: (i) KDDTrain+ and (ii) KDDTest+. The KDDTrain+ part of the dataset NSL-KDD is used to train a system to detect network intrusions or the detection method. It consist of four classes of attacks and a normal class data set. The KDDTest+ part of NSLKDD dataset is used for testing a detection method or a system when it is evaluated for performance. It additionally contains the same classes of attack traffic within the training set [11]. The dataset NSL-KDD has 41 attribute and a class attribute. From the once 41 attribute some attribute have no role and some have minimal role in detecting attacks [14].

41 attributes are included three types of features: Binary, Numeric, and Nominal. Table VI indicates Features name and types [24].

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 10, October 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0910012 |

TABLE VI

NSL-KDD FEATURES AND TYPES

Type	Features
Nominal	Service(3), Protocol_Type(2), Flag(4)
Binary	Su_Attempted(15), Is_Host_Login(21) , Root_Shell(14), Is_Guest_Login(22), Land(7), Logged_In(12)
Numeric	Duration(1), Dst_Bytes(6), Urgent(9), Src_Bytes(5), Num_Failed_Logins(11), Num_Root(16), Hot(10), Count(23), Wrong_Fragment(8), Rerror_Rate(27), Dst_Host_Srv_Serror_Rate(39), Dst_Host_Srv_Count(33), Srv_Diff_Host_Rate(31), Num_File_Creations(17), Dst_Host_Diff_Srv_Rate(35), Num_Shells(18), Num_Access_Files(19), Dstdst_Host_Rerror_Rate(40), Num_Compromised(13), Num_Outbound_Cmds(20), Serror_Rate(25), Dst_Host_Count(32), Dst_Host_Same_Srv_Rate(34), Diff_Srv_Rate(30), Dst_Host_Same_Src_Port_Rate(36), Srv_Rerror_Rate(28), Dst_Host_Srv_Diff_Host_Rate(37), Srv_Serror_Rate(26), Dst_Host_Serror_Rate(38), Same_Srv_Rate(29), Dst_Host_Srv_Rerror_Rate(41), Srv_Count(24),

NSL KDD Train+, NSL KDD Test+ and NSL KDD 20%Train has been analyze which shows that there are 22, 37 and 21 types of attacks in the datasets with varying percentage of different attacks which is shown in Table VII, Table VIII and Table IX.

TABLE VII

ATTACK FREQUENCY IN NSL KDD TRAIN

Attack Type	Value	Count	Percent
Normal	normal	67343	53.46%
Dos	neptune	41214	32.72%
	teardrop	892	0.71%
	smurf	2646	2.10%
	pod	201	0.16%
	back	956	0.76%
	land	18	0.01%
Probe	ipsweep	3599	2.86%
	portsweep	2931	2.33%
	nmap	1493	1.19%
	satan	3633	2.88%
R2L	warezclient	890	0.71%
	guess_passwd	53	0.04%
	ftp_write	8	0.01%
	multihop	7	0.01%
	imap	11	0.01%
	warezmaster	20	0.02%
	phf	4	0.00%
	spy	2	0.00%
U2R	rootkit	10	0.01%

	buffer_overflow	30	0.02%
	loadmodule	9	0.01%
	perl	3	0.00%
Total		**125973**	**100%**

TABLE VIII

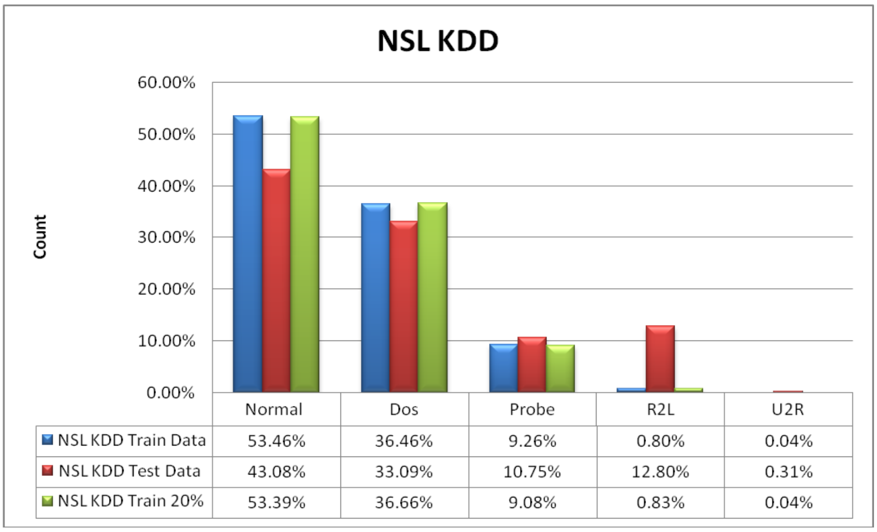
ATTACK FREQUENCY IN NSL KDD TEST

Attack Type	Value	Count	Percent
Normal	normal	9711	43.08%
Dos	neptune	4657	20.66%
	smurf	665	2.95%
	apache2	737	3.27%
	back	359	1.59%
	processtable	685	3.04%
	pod	41	0.18%
	mailbomb	293	1.30%
	worm	2	0.01%
	teardrop	12	0.05%
	land	7	0.03%
	udpstorm	2	0.01%
Probe	saint	319	1.42%
	mscan	996	4.42%
	satan	735	3.26%
	nmap	73	0.32%
	ipsweep	141	0.63%
	portsweep	157	0.70%
R2L	guess_passwd	1231	5.46%
	warezmaster	944	4.19%
	snmpgetattack	178	0.79%
	httptunnel	133	0.59%
	snmpguess	331	1.47%
	multihop	18	0.08%
	named	17	0.08%
	sendmail	14	0.06%
	xlock	9	0.04%
	xsnoop	4	0.02%
	ftp_write	3	0.01%
	imap	1	0.00%
	phf	2	0.01%
U2R	buffer_overflow	20	0.09%
	ps	15	0.07%
	loadmodule	2	0.01%
	xterm	13	0.06%
	rootkit	13	0.06%
	perl	2	0.01%
	sqlattack	2	0.01%
Total		**22544**	**100%**

TABLE IX

ATTACK FREQUENCY IN NSL KDD20% TRAIN

Attack Type	Value	Count	Percent
Normal	normal	13449	53.39%
Dos	neptune	8282	32.88%
	teardrop	188	0.75%
	smurf	529	2.10%
	pod	38	0.15%
	back	196	0.78%
	land	1	0.00%
Probe	ipsweep	710	2.82%
	portsweep	587	2.33%
	nmap	301	1.19%
	satan	691	2.74%
R2L	warezclient	181	0.72%
	guess_passwd	10	0.04%
	ftp_write	1	0.00%
	multihop	2	0.01%
	imap	5	0.02%
	warezmaster	7	0.03%
	phf	2	0.01%
	spy	1	0.00%
U2R	rootkit	4	0.02%
	buffer_overflow	6	0.02%
	loadmodule	1	0.00%
Total		**25192**	**100%**



	Normal	Dos	Probe	R2L	U2R
🟦 NSL KDD Train Data	53.46%	36.46%	9.26%	0.80%	0.04%
🟥 NSL KDD Test Data	43.08%	33.09%	10.75%	12.80%	0.31%
🟩 NSL KDD Train 20%	53.39%	36.66%	9.08%	0.83%	0.04%

Figure 5. NSL KDD attack classification

DARPA, KDD99, and NSL-KDD in figure 6 give a general overview for sets of related data in this study. DARPA is a set of raw dataset. KDD99 is the feature extracted edition of DARPA dataset. NSL-KDD is duplicates removed and reduced size version of KDD99 dataset [27].

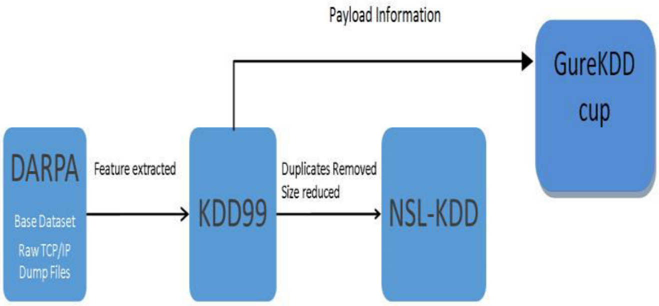


Figure. 6 The relation between main and extracted datasets.

GureKDD Cup

GureKDDcup dataset is consist of kddcup99 connections (UCI repository database) also payload added to (network packets contents) each and every connections. The GureKDDCup capture group employs the similar methods implemented to create kddcup99 [28]. They processed tcpdump data files with bro-ids and also obtained every connection with its proper features. And finally, the dataset is labeled each and every connection based on the connections-class files (tcpdump.list) which provided by MIT. The Original dataset size is too large i.e 9.3 GB and the size of 6 percent dataset is 4.2 GB.

GureKddcup (and gureKddcup6percent) contains 41 attributes same as the KDDcup'99. The gureKddcup is too big to be utilized in any learning process. Most of the research projects with *kddcup* database are carried out by using the 10% of the database available in UCI [28]. A reduced sample: gureKddcup6percent which consisting of only no-flood attacks matched with tcpdump.list along with a random subsample of normal connections matched with tcpdump.list. Particulars of the dataset which include number of samples, attack categories are mentioned in Table X and Figure 7 shows abnormal and normal class.

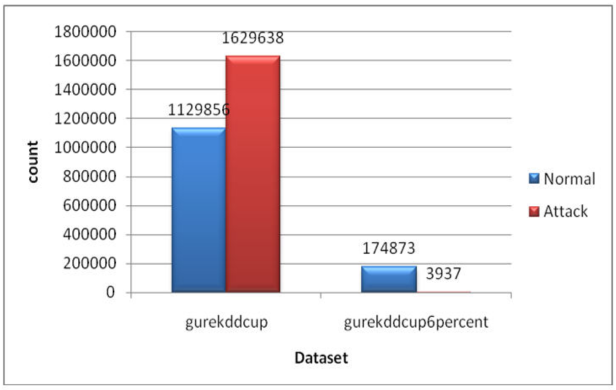


Figure 7. Attack and Normal instances in gurekddcup and gurekddcup6percent

TABLE X

ATTACK FREQUENCY IN GUREKDDCUP AND GUREKDD6PERCENT

Attack Name	gureKddcup		gureKddcup6percent	
	No. of Instances	%	*No. of Instances*	%
anomaly	9	0.00033	9	0.005
dict	879	0.03185	879	0.492
dict_simple	1	0.00004	1	0.001
eject	11	0.00040	11	0.006
eject-fail	1	0.00004	1	0.001

ffb	10	0.00036	10	0.006
ffb_clear	1	0.00004	1	0.001
format	6	0.00022	6	0.003
format_clear	1	0.00004	1	0.001
format-fail	1	0.00004	1	0.001
ftp-write	8	0.00029	8	0004
guest	50	0.00181	50	0.028
imap	7	0.00025	7	0.004
land	35	0.00127	35	0.020
load_clear	1	0.00004	1	0.001
loadmodule	8	0.00029	8	0.004
multihop	9	0.00033	9	0.005
perl_clear	1	0.00004	1	0.001
perlmagic	4	0.00014	4	0.002
phf	5	0.00018	5	0.003
rootkit	29	0.00105	29	0.016
spy	2	0.00007	2	0.001
syslog	4	0.00014	4	0.002
teardrop	1085	0.03932	1085	0.607
warez	1	0.00004	1	0.001
warezclient	1749	0.06338	1749	0.978
warezmaster	19	0.00069	19	0.011
pod	5	0.00018		
back (flood)	2248	0.08146		
ipsweep (flood)	15760	0.57112		
neptune (flood)	1526643	55.32329		
nmap (flood)	1995	0.07230		
portsweep (flood)	9973	0.36141		
satan (flood)	31411	1.13829		
smurf (flood)	37666	1.36496		
normal	1129856	40.94431	174873	97.7982
TOTAL	**2759494**	**100%**	**178810**	**100%**

The database gureKDDCup has been generated within the UADI project (Unsupervised Anomaly Detection for Intrusion detection system) in which a classifier that detects intrusions or attacks in network based systems was developed. The main distinctive feature of this project is that it uses the payload (body part of network packages) to detect attacks in network connections[28]. The analysis of the payload to classify the connections is not a deeply analysed field, however, it seems that it is essential to detect attacks such as R2L (Remote to Local, its goal is to use resources without permission) and U2R (User to Root, its goal is to get root or administrative privileges without having them). GureKDDCup has similar features to the ones in KDDCup99, but additional payload information and other features related to the connection such as IP address and port numbers. A new extension of the (KDDCup99+payload) that we called it gureKDDCup.

III. CONCLUSION

As the attacks and information threats are increasing rapidly there is a need for an improved intrusion detection system that can cope with the situation. In this paper, we have studied the DARPA, KDD CUP'99, NSL-KDD and

GureKDDcup dataset. While comparing this dataset, the survey shows that NSL-KDD dataset is most suitable for comparing different intrusion detection models. Using all 41 dataset features to the intrusive patterns might result in to time consuming and it also reduces the degradation of the system performance. Some of the features of KDD CUP 99 dataset are unnecessary and insignificant to the process. Gurkddcup dataset size is too big so due to this, only its reduce dataset gurekddcup6percent is used for practical implementation. NSL-KDD does not contain any duplicate records in train dataset and test dataset.

REFERENCES

[1] Ravi Jain and Ajith Abraham "Soft Computing Models for Network Intrusion Detection Systems" School of Information Science, University of South Australia, Australia ravi.jain@unisa.edu.au and Department of Computer Science, Oklahoma State University, USA ajith.abraham@ieee.org
[2] Anita John and Deepthy K Denatious "Survey on Data Mining Techniques to Enhance Intrusion Detection", International Conference on Computer Communication and Informatics (ICCCI-2012), 10 – 12, January, 2012, Coimbatore, INDIA
[3] Kai-Fan Cheng, Rung-Ching Chen and Chia-Fen Hsieh, "Using Rough Set And Support Vector Machine For Network Intrusion Detection", International Journal of Network Security & Its Applications (IJNSA), Vol 1, No 1, April 2009
[4] Mr. Suresh Kashyap, Ms. Pooja Agrawal, Mr.Vikas Chandra Pandey, Mr. Suraj Prasad Keshri "Soft Computing Based Classification Technique Using KDD 99 Data Set for Intrusion Detection System"*International Journal of Advanced Research in Electrical, Electronics and Instrumentation EngineeringVol. 2, Issue 4, April 2013*
[5] D. Rozenblum, "Understanding Intrusion Detection Systems," *SANS Inst.*, no. 122, pp. 11–15, 2001.
[6] "What it is Network intrusion detection system? | COMBOFIX." [Online]. Available: http://www.combofix.org/what-it-is-network-intrusion-detection-system.php. [Accessed: 10-Dec-2015].
[7] Sara Memar, Mohammadreza Ektefa, " Intrusion Detection Using Data Mining Techniques", IEEE 2010.
[8] Roesch Martin, 1999. Snort–lightweight intrusion detection for networks, 13th Systems Administration Conference (LISA), pages 229–238.
[9] R.P. Lipmann, D.J. Fried, I. Graf, et al., "Evaluating Intrusion Detection System: The 1998 DARPA offline intrusion Detection Evaluation," in Proceeding of the DARPA Information Survivability Conference and Exposition, pp. 12-26, 2000.
[10] Vishwas Sharma, Ciza Thomas, N. Balkrishan, "Usefulness of DARPA Dataset for Intrusion Detection System Evaluation", in proceeding of SPIE-The International Society of Optical Engineering, March 2008
[11] Monowar H. Bhuyan, Dhruba K Bhattacharyya, Jugal K. Kalita, "Towards Generating Real Life Datasets for Network Intrusion Detection" in International Journal of Network Security, Vol. 17, No.6, PP.675-693, Nov. 2015
[12] DARPA intrusion detection evaluation, http://www.ll.mit.edu/IST/ideval /data/ data index.html
[13] K. Kendall, A database of computer attacks for the evaluation of intrusion detection systems, Thesis, MIT, 1999
[14] Bhupendra Ingre, Anamika Yadav, "Performance Analysis of NSL-KDD Dataset using ANN" by International conference on signal processing and communication Engineering Systems (SPACES), 2015
[15]http://www.qnx.com/developers/docs/660/index.jsp?topic=%2Fcom.qnx.doc.neutrino.user_guide%2Ftopic%2Fsecurity_Remote_Local.html
[16] Information Systems Technology Group MIT Lincoln Lab, DARPA Intrusion Detection DataSets, Mar.2000. http://www.ll.mit.edu/mission/ communications /ist/corpora/ideval/data/2000data.html)
[17] KDD Cup 1999 Intrusion Detection Dataset. [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kdd cup99.html
[18] Ebrahim Bagheri, Wei Lu, Mahbod Tavallaee and Ali A. Ghorbani , "A Detailed Analysis of the KDD CUP 99 Data Set", in IEEE Symposium on Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009.
[19] P. GiftyJeya, M. Ravichandran, C. S. Ravichandran "Efficient Classifier for R2L and U2R Attacks" in International Journal of Computer Applications (0975 – 8887) Volume 45– No.21, May 2012
[20] H. GünesKayacık, A. NurZincir-Heywood, Malcolm I. Heywood, "Selecting Features for Intrusion Detection: A Feature Relevance Analysis on KDD 99 Intrusion Detection Datasets" in Third Annual Conference on Privacy, Security and Trust, October 12-14, 2005
[21] S. Hettich, S.D. Bay. 1999. The UCI KDD Archive. Irvine, CA: University of California, Department of Information and Computer Science. http://kdd.ics.uci.edu.
[22] P. Bhoria, K. Kanwal Garg. "Determining feature set of DOS attacks", in International Journal of Advanced Research in Computer Science and Software Engineering, vol.3 issue 5, May 2013, pp. 875-878.

[23] Aditya Shrivastava, Mukesh Baghel, Hitesh Gupta, “A Review of Intrusion Detection Technique by Soft Computing and Data Mining Approach”, in International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-3 Number-3 Issue-12 September-2013

[24] Hee-su Chae, Byung-oh Jo, Sang-Hyun Choi, Twae-kyung Park, “Feature Selection for Intrusion Detection using NSL-KDD” in Recent Advances in Computer Science

[25] S. Revathi, Dr. A. Malathi, “A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection” in International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 12, December – 2013

[26] http://www.unb.ca/research/iscx/dataset/iscx-NSL-KDD-dataset.html

[27] Atilla Ozgur, Hamit Erdem, “Review of KDD99 Dataset Usage in Intrusion Detection and Machine Learning between 2010 and 2015”, in PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1954v1 | CC-BY 4.0 Open Access | rec: 14 Apr 2016, publ: 14 Apr 2016

[28]https://addi.ehu.es/bitstream/handle/10810/20608/20160601_Txostena_gurek ddcup_InigoPeronaBalda.pdf?sequence=1

[29]Adetunmbi A. Olusola et. al.,” Analysis of KDD“99 Intrusion Detection Dataset for Selection of Relevance Features”, Proceeding of the World congress on Engineering and Computer Science, vol. 1, October 20-22,(2010), San Francisco, USA,(2010).

[30]S. Hettich et. al,” The UCI KDD Archive”. Irvine, CA: University of California, Department of Information and Computer Science, http://kdd.ics.uci.edu,1999.u/IST/ideval/docs/1998/id98-eval-11.txt 25 March (1998).

[31] Sabhnani M. et. al. ,” Why machine learning algorithms fail in misuse detection on kdd intrusion detection dataset”, Intell Data Anal 8:403-415, URL http://portal.acm.org/citation.cfm?id=1293805. 1293811, (2004).

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING