A Detail Analysis of KDD 1999, NSL KDD 1999 and GureKDD Dataset of Intrusion Detection System

Devendra K. Singh[1], *Dr. Manish Shrivastava[2]
Assistant Professor[1](Research Scholar), Assistant Professor[2]
Dept. of Computer Sci. & Engg.
SoS, Engg. & Technology
devendra.singh1700@gmail.com[1], devendra.singh170@gmail.com[1],
manbsp@gmail.com[2]
Mob. No: 9827471404[1], Mob. No:9827116390[2]

Abstract

The objective of this study was to survey and find the state of the art in the research area in intrusion detection. In the field of security requirement in between of computer network, Intruder is breaking the system of security. An intruder is dangerous for the health of computer security. Study of an intruder is very important for the security field. The KDD'99 dataset is providing the data set in the field of 41 features. These features are the information of data packet and that data packet is the collection of intruder's data. NSL KDD'99 data is the refined data of the KDD'99 dataset. In 1998 DARPA organized intrusion detection evaluation program by MIT Lincoln Labs (MIT). In this paper, we are trying to complete study of the KDD'99 / NSL-KDD'99 dataset.

Keywords: *KDD, NSL-KDD, DARPA, IDS.*

(1) Introduction: By through the intrusion detection system we can monitor the network traffic and monitor the unauthorized and suspicious activity in the network, after finding the information of attack we can alert the administrator of the network and alert the system. When we find the attacks then we get the IP address of that machine and send that information to the network administrator and stop or break the connection of network and save the machine. Administrators have the records of the attacker and manage the table in form of a white box and black box list. Administrators have the power of stop the connection or break the connection. There is number of algorithms available for finding the intruders available in the network. By the dataset, we can identify the types of attacks. KDD'99 / NSL-KDD'99 dataset are intruder's dataset. That dataset collected by the DARPA at MIT Lincoln Labs.

In 1998 DARPA organized at MIT Lincoln Lab for online competition for finding the different types of attack available in computer network on different – 2 machines (i.e. UNIX/LINUX). DARPA providing a platform for participation at MIT Lincoln Lab (under the sponsorship of DARPA)[3]. DARPA 1998 is about 4 gigabytes of compressed raw data of TCP dump data of seven weeks of network traffic. This will be processed into 5 million connection records, each is 100 bytes. The KDD 1999 training dataset used to develop the model for finding intruder of the computer network and try to develop the best model with more efficient in finding of all types of attacks. This dataset is raw TCP dump data. This dataset is collected over a period of nine weeks on Local Area Network (LAN). Training dataset [1] was processed in five million connections records from seven weeks of network traffic and two weeks of testing data records around two million connections. There are 41 features is either normal or an attack [6].

(2) Intrusion Detection Dataset: In this paper we are trying to analysis in between of KDD'99, NSL KDD'99 and GureKDD dataset [5].

(i) **KDD'1999 dataset**: This dataset is collected by the DARPA. This dataset is the prepared and managed by the MIT Lincoln Labs organized by DARP to collect the intrusion detection data and manage the collected dataset. Above data is collected by the nine weeks of TCP raw data for LAN. Above data available in UCI Repository of KDD99

intrusion detection dataset. That data is divided into 5 categories (i.e. Normal, DoS, U2R, R2L, and Prob). In KDD'1999 full dataset that is 4898431 dataset, KDD'1999 10% dataset is 494021 dataset and KDD'1999 corrected dataset is 3111029 dataset. The KDD'99 dataset contains 24 attack types in training and 14 more attack types in testing for the total of 38 attacks available in this dataset. These 14 new types of attacks theoretically test IDS capability to generalize to unknown attacks. At the same time, it is hard for machine learning based IDS to detect these 14 new attacks [6,7].

(ii) **NSL KDD'99 dataset**: As par record of NSL KDD 99 there is 125973 samples dataset and test dataset is 22544 samples of the dataset.

(iii) **GureKDD dataset:** This dataset is UCI repository dataset. This is also TCP dump files. The size of the dataset is 9.3 GB and 6% dataset size is 4.2 GB.

Different types of attack are divided into five categories that are following:

(i) **Normal Attack:** In this attack, there are no attacks computer networks. This is real user or normal user connection in the computer network.

(ii) **DoS Attack:** This one is Denial of Services. In this attack user unable the use of services. Users feel that there are unable to access the system. Example is (a) ping-of-death, (b) teardrop, (c) smurf, (d) syn flood, etc.

(iii) **U2R Attack:** Attacker attacks the local user machine by unauthorized and gets the privileges of the user machine. An example is (a) buffer overflow attacks etc.

(iv) **R2L Attack:** Unauthorized access by through the root user. Attacker attacks in root level to user machine and gets the privileges of the machine. An example is (a) guessing password etc.

(v) **Probing Attack:** In this attack, attacker tries to get the information from target host machine. By probing attack attacker find the known vulnerabilities. Example is (a) port-scan, (b) ping-sweep, etc.

In KDD99 dataset there are 41 features are available in Table 1. These features are collected by UCI Repository KDD99 dataset for intrusion detection [3,8,7].

Table 1: Features of KDD99 Dataset [3,8,7].

S.No.	Name of Features	S.No.	Name of Features
1	duration	22	is_guest_login
2	protocol_type	23	count
3	service	24	srv_count
4	flag	25	serror_rate
5	src_bytes	26	srv_serror_rate
6	dst_bytes	27	rerror_rate
7	land	28	srv_rerror_rate
8	wrong_fragt	29	same_srv_rate
9	urgent	30	diff_srv_rate
10	hot	31	srv_diff_h_rate
11	num_fail_login	32	host_count
12	logged_in	33	host_srv_count
13	nu_comprom	34	h_same_sr_rate
14	root_shell	35	h_diff_srv_rate
15	su_attempted	36	h_src_port_rate
16	num_root	37	h_srv_d_h_rate
17	nu_file_creat	38	h_serror_rate
18	nu_shells	39	h_sr_serror_rate
19	nu_access_files	40	h_rerror_rate
20	nu_out_cmd	41	h_sr_rerror_rate
21	is_host_login		

The KDD99 intrusion detection dataset consists of three benchmark components, which are shown in Table 2. Only 10% KDD99 IDS dataset is used for training purpose. In Table 3, maintain the 23 attached with the classification of 4 types (i.e. DoS, U2R, R2L, and Prob).

Table 2: Basic Characteristics of KDD99 intrusion detection dataset [8]

S.No.	Dataset	DoS	Probe	U2R	R2L	Normal	Total Dataset
1	**10% KDD99 IDS Dataset**	391458	4107	52	1126	97277	494020
2	**Corrected KDD99**	229853	4166	70	16347	60593	311019
3	**Whole data of KDD99**	3883370	41102	52	1126	972780	4898430

In Fig 1 we are showing the DoS attck with the diagram in form of (a) 10% KDD'99 IDS dataset DoS attack is 391458 (b) Corrected KDD'99 IDS dataset DoS attack is 229853 and (c) Whole data of KDD'99 IDS dataset DoS attack is 3883370 [8].

In Fig 2 we are showing the Prob attck with the diagram in form of (a) 10% KDD'99 IDS dataset Prob attack is 4107, (b) Corrected KDD'99 IDS dataset Prob attack is 4166, and (c) Whole data of KDD'99 IDS dataset Prob attack is 41102 [8].

In Fig 3 we are showing the **U2R** attck with the diagram in form of (a) 10% KDD'99 IDS dataset **U2R** attack is 52 (b) Corrected KDD'99 IDS dataset **U2R** attack is 70 and (c) Whole data of KDD'99 IDS dataset **U2R** attack is 52 [8].

In Fig 4 we are showing the **R2L** attck with the diagram in form of (a) 10% KDD'99 IDS dataset **R2L** attack is 1126 (b) Corrected KDD'99 IDS dataset **R2L** attack is 16347 and (c) Whole data of KDD'99 IDS dataset **R2L** attack is 1126 [8].

In Fig 5 we are showing the **Normal** data with the diagram in form of (a) 10% KDD'99 IDS dataset **Normal data** is 97277 (b) Corrected KDD'99 IDS dataset **Normal** is 60593 and (c) Whole data of KDD'99 IDS dataset **Normal** is 972780 [8].

In Fig 6 we are showing the data samples with the all type of attack availables in a KDD'99 dataset. By the bar diagram we can easily identify the position of all types of attack available in KDD'99 dataset.

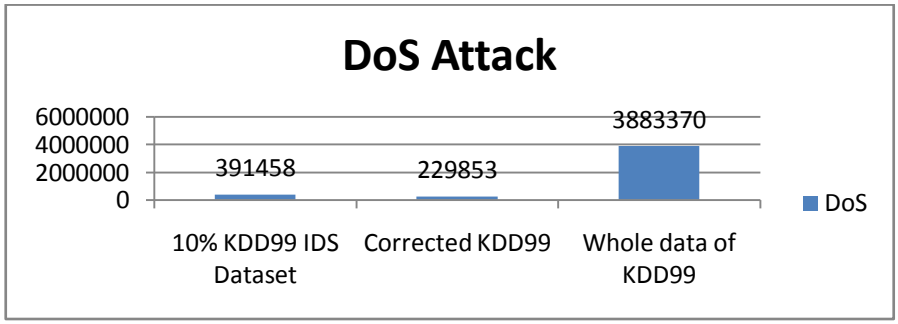

Fig 1: DoS Attack showing in KDD'99 IDS dataset

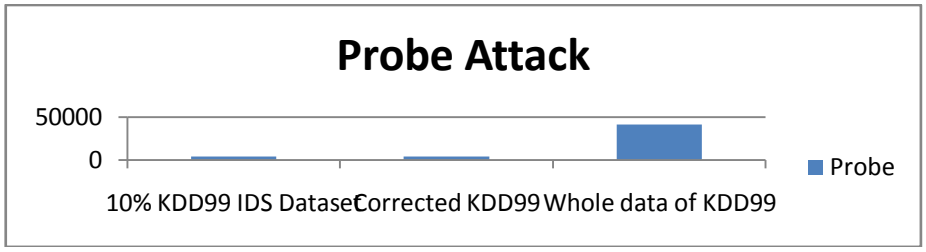

Fig 2: Probe Attack showing in KDD'99 IDS dataset

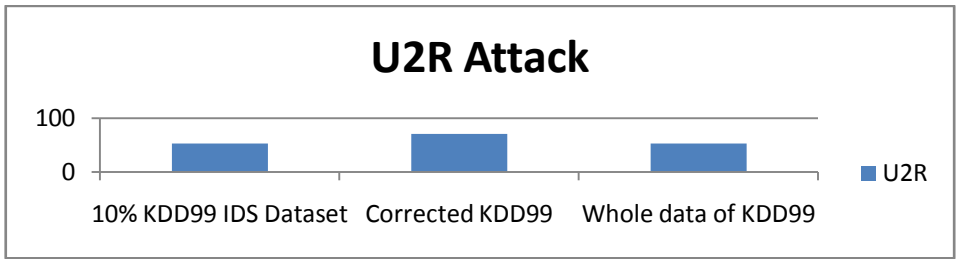

Fig 3: U2R Attack showing in KDD'99 IDS dataset

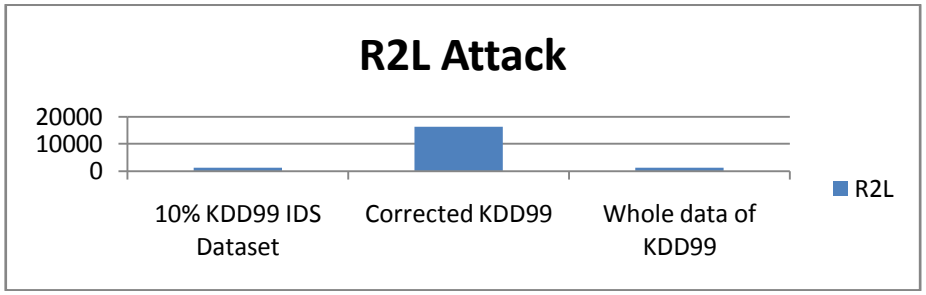

Fig 4: R2L Attack showing in KDD'99 IDS dataset

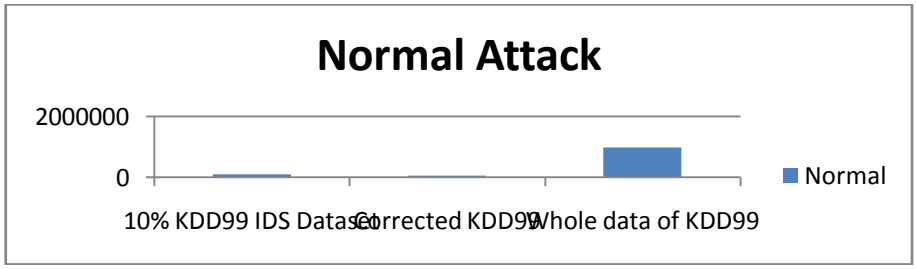

Fig 5: Normal Attack showing in KDD'99 IDS dataset [20].

Table 3: All dataset divided in Class labels that appears in 10% KDD'99 IDS dataset

S.No.	Attack	Data Samples	Category
1	**smurf**	280790	DoS
2	**neptune**	107201	DoS
3	**back**	2203	DoS
4	**teardrop**	979	DoS
5	**pod**	264	DoS
6	**land**	21	DoS
7	**normal**	97277	Normal

8	**satan**	1589	Probe
9	**ipsweep**	1247	Probe
10	**portsweep**	1040	Probe
11	**nmap**	231	Probe
12	**warezclient**	1020	R2L
13	**guess_passwd**	53	R2L
14	**warezmater**	20	R2L
15	**imap**	12	R2L
16	**ftp_write**	8	R2L
17	**multihop**	7	R2L
18	**phf**	4	R2L
19	**spy**	2	R2L
20	**buffer_overflow**	30	U2R
21	**rootkit**	10	U2R
22	**loadmodule**	9	U2R
23	**perl**	3	U2R
Total Dataset =		102583	

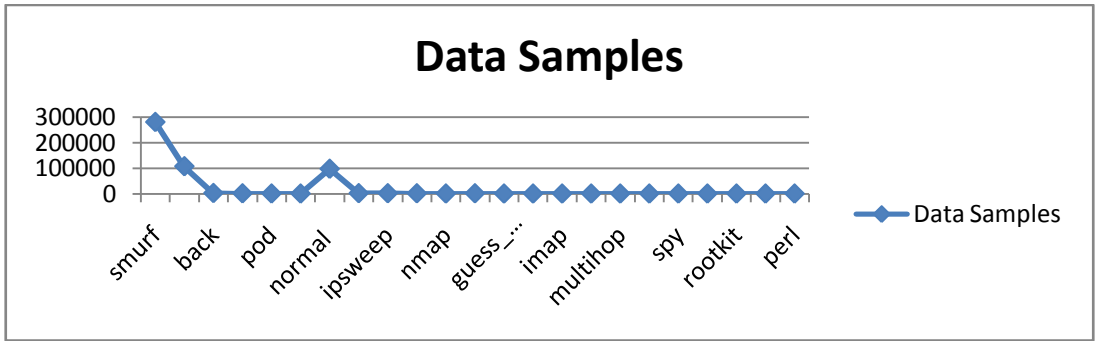

Fig 6: Data Samples with all Attack showing in KDD'99 IDS dataset

In Table 4 we are trying to show the feature ranking of KDD'99 IDS dataset.

Table 4: Ranking information of KDD99 IDS dataset [4]

CLASS	No of Features	Feature Ranking of KDD99 IDS Dataset
DoS	12	f23,f5,f3,f6,f32,f24,f12,f2,f37,f36,f8,f31
Probe	19	f5,f27,f3,f35,f40,f37,f33,f17,f41,f30,f34,f28,f22,f4,f 24,f25,f19,f32,f29
U2R	23	f37,f17,f8,f18,f16,f1,f4,f15,f7,f22,f20,f21,f31,f19, f12,f13,f14,f6,f32,f29,f3,f40,f2
R2L	15	f3,f15,f5,f10,f9,f32,f33,f22,f1,f17,f24,f11, f23,f8,f6

The information of attack categories of NSL KDD'99 and GureKDD dataset are shown in Table 5. We can see the information on different types of attack is shown in this table.

Table 5: Attack categories and total samples present in NSLKDD and GureKDD Dataset [5].

S.No.	Attack Class	NSL KDD Train	NSL KDD Train	KDD Test+	GureKDD Dataset	Instances On GureKDD	After Removing Duplicate	% rate of reduction

		Full	20%			Original	samples	
1	**Normal**	67343	13449	9711	**Normal**	174873	157048	10.19
2	**Apache2**	0	0	737	**Anomaly**	9	9	0.00
3	**Back**	956	196	359	**Dict**	880	878	0.23
4	**Buffer Overflow**	30	6	20	**Dict_simple**	2	1	50.00
5	**Ftp_write**	8	1	3	**Eject_fail**	2	1	50.00
6	**Guess_pwd**	53	10	1231	**Eject**	12	11	8.33
7	**HttpTunnel**	0	0	133	**Ffb**	11	10	9.09
8	**Imap**	11	5	1	**Ffb_clear**	2	1	50.00
9	**Ipsweep**	3599	710	141	**Format-fail**	2	1	50.00
10	**Land**	18	1	7	**Format**	7	6	14.29
11	**Loadmodule**	9	1	2	**Format_clear**	2	1	50.00
12	**MailBomb**	0	0	293	**Ftp_write**	9	8	11.11
13	**Mscan**	0	0	996	**Guest**	51	50	1.96
14	**Multihop**	7	2	18	**Lmap**	8	7	12.50
15	**Named**	0	0	17	**Land**	36	17	52.78
16	**Neptune**	41214	8282	4657	**Load_clear**	2	1	50.00
17	**Nmap**	1493	301	73	**Loadmodule**	9	6	33.33
18	**Perl**	3	0	2	**Multihop**	9	9	0.00
19	**Phf**	4	2	2	**Perl_clear**	2	1	50.00
20	**Pod**	201	38	41	**Perl_magic**	5	4	20.00
21	**Portsweep**	2931	587	157	**Phf**	6	5	16.67
22	**ProcessTable**	0	0	685	**Rootkit**	30	29	3.33
23	**Ps**	0	0	15	**Spy**	3	2	33.33
24	**Rootkit**	10	4	13	**Sys_log**	5	3	40.00
25	**Saint**	0	0	319	**Teardrop**	1086	1083	0.28
26	**Satan**	3633	691	735	**Warez**	2	1	50
27	**SendMail**	0	0	14	**Warezclient**	1750	1692	3.31
28	**Smurf**	2646	529	665	**Warezmaster**	20	19	5.00
29	**SnmpGuess**	0	0	331	**Total**	178835	160904	10.03%
30	**Spy**	2	1	0				
31	**SqlAttack**	0	0	2				
32	**Teardrop**	892	188	12				
33	**UdpStorm**	0	0	2				
34	**Warezmaster**	20	7	944				
35	**Warezclient**	890	181	0				
36	**Worm**	0	0	2				
37	**Xlock**	0	0	9				
38	**Xsnoop**	0	0	4				
39	**Xterm**	0	0	13				
	Total	125973	25192	22544				

Table 6: Divided in list of feature elements of KDD'99 IDS dataset [5].

List of feature	
Basic Features	1)Duration, 2) Protocol Type, 3) Service, 4) Flag, 5) Source bytes, 6) Destination bytes
Content Features	7)Land, 8) Wrong Fragment, 9) Urgent, 10) Hot, 11) Failed Logins, 12) Logged in, 13) Compromised, 14) Root shell, 15) Su attempted, 16) Root 17) File creations, 18) Shells, 19) Access files, 20) Outbound cmds, 21) Is host login, 22) Is guest login

Traffic Features	23) Count, 24) Srv count, 25) Serror rate, 26) Srv serror rate, 27) Rerror rate, 28) Srv rerror rate, 29) Same srv rate, 30) Diff srv rate, 31) Srv diff host rate
Host Based Traffic Features	32) Dst host count, 33) Dst host srv count, 34) Dst host same srv rate, 35) Same Srv rate, 36) Dst host same src port rate, 37) Dst host srv diff host rate, 38) Dst Host rate, 39) Dst host srv serror rate, 40) Dst host rerreor rate, 41) Dst host srv Rerror rate.

(3) Related Work in KDD'99 IDS dataset: From year 1999 there are many researchers work in IDS and they are trying to develop the efficient model of IDS for find the Intruders. They are also using KDD'99 dataset of IDS and develop the efficient model of IDS. Our KDD'99 IDS dataset is approximately 18 years old dataset.

Laheeb et.al.[14] author worked on comparison between the KDD cup 99 IDS dataset and NSL-KDD99 IDS dataset by using self organization map (SOP) on artificial neural network of soft computing. Finding of detection rate of KDD'99 in this paper is 92.37% and detection rate of NSL KDD'99 is 75.49%.

Lee et. al. [12] authors worked on DARPA 1998 dataset of KDD'99 features by using the data mining techniques. In this paper author saying KDD 99 dataset are much more dataset for observation and finding result is not so easy, it is very difficult to find and develop the model for find IDS.

H.S.Hota et. al.[13] authors worked in feature selection of KDD'99 dataset. They have worked in different data mining rules and find the result of reduced features. In this paper they have find the accuracy by artificial neural network is 99.56% and by using Bayesian Net is 99.51%.

Zargari.et.al.[16] author worked on KDD cup 99 and NSL KDD99 dataset. In this paper use the Data mining techniques for finding significant features on both above dataset.
Sabhnani et. al.[18] author worked on Decision Tree and statistical method for find the R2L attacks by using KDD'99 IDS dataset.

Vipin Kumar et. al. [15] author worked on NSL KDD'99 IDS dataset by using K-Mean clustering algorithm. He worked on complete analysis of NSL KDD'99 IDS dataset.

Ali et. al. [19] author worked for compared the different dataset. They are worked in realistic traffic of network with the different dataset.

Olusola et. al. [17] author finding the relevant features of KDD'99 IDS dataset. They have also worked on 10% Kdd99 data of intrusion detection dataset (IDS). They have also found the some features are not relevant in any attack.

(4) Requirement of Analysis: If we are using some dataset in our research work than analysis is too much important for do your research work. Without study of KDD'99 dataset of IDS we can't work anything in research field. We motivate for the study of KDD'99 dataset after that we can use in our research work. So before start research work analysis is must for research. After the study of dataset we can find the important point in dataset than we do valuable work in our research work.

(5) Conclusion:

By the analysis of intrusion detection dataset, we have found the type of attack categories of KDD99, NSL KDD99, and GureKDD Dataset. In KDD99 there is 23 attack available in this dataset but in NSL KDD99 there are 39 attacks is available in this dataset [5] and by the GureKDD dataset there is 28 attack available in this dataset. The objective of this analysis is finding the attack categories available in the different dataset. We can use that information to create or develop the IDS model for secure your computer network for safe use.

References:

[1]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).

[2]Jenifar Dayana. A et. al. ," Analysis of Intrusion Detection in KDD'99 Dataset", IJSART, Vol. 1, Issue 8, August (2015).

[3]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).

[4] Mohammed A. A. et. al.," Building an Intrusion Detection System Using a Filter-Based Feature Selection Algorithm", IEEE Transaction on Computers, Issue 10, vol. 65, Oct. (2016).

[5]Santosh Kumar Sahu et. al.," A Detail Analysis on Intrusion Detection Dataset", IEEE International Advance Computing Conference (IACC), (2014).

[6] 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).

[7]The 1998 intrusion detection off-line evaluation plan. MIT Lincoln Lab., Infromation Systems Technology Group. http://www.ll.mit.edu/IST/ideval/docs/1998/id98-eval-11.txt 25 March (1998).

[8]Knowledge discovery in databases DARPA archive. Task Description. http://kdd.ics.uci.edu,/databases/kddcup99/task.html.

[9]Kayacik et. al.," Analysis of three Intrusion Detection System Benchmark Datasets Using Machine Learning Algorithms", Proceedings of the IEEE ISI 2005 Atlanta, USA, May (2005).

[10]Wong et. al. ," Synthesizing Statistical Knowledge from incomplete mixed-model data", IEEE Transaction on Pattern Analysis and Machine Intelligence, vPAMI-9,no. 6, November (1987), pp 796-805,(1987).

[11] Mahbod Tavallaee et. al., "Detail Analysis of the KDD CUP 99 Data Set", Proceeding of the 2009 IEEE Symposium on Computational Intelligence in Security and Defense application (CISDA 2009), (2009).

[12]W.Lee et. al.,"A framework for constructing features and model's for intrusion detection systems", ACM Transaction information and system security, vol 3(4), Nov. (2000), pp. 227-261.

[13]H.S.Hota, Akhilesh Kumar Sriwas and S.K. Singhai " An Ensemble Classification Model for Intrusion Detection System With Feature Selection", published in International Journal of Decision Science & Information Technology, Vol. 3, no. 1, (2011), ISSN 1937-9013. School of Engg., Taylor's University, pp. 13-24.

[14]Laheeb m. Ibrahim et. al.,"Comparison study for intrusion database(KDD,NSL KDD) based on self organization (SOM) Artificial Neural Network", Journal of Engineering Science and Technology, vol 8, No 1, (2013), pp, 107-119.

[15]Vipin Kumar et. al.," K-Means Clustering Approach to Analyze NSL-KDD intrusion Detection Dataset", International Journal of Soft Computing and Engg. (IJSCE), ISSN: 2231-2307, vol. 3, sept. (2013).

[16] Zargari S. et.al.,"Feature selection in the corrected KDD dataset, Emerging intelligent Data and Web Tech. (EIDWT)", Third international conference, 19-21 Sept. (2012), pp. 174-180.

[17]Olusola et.al.,"Analysis of KDD'99 intrusion detection dataset for selection of relevance features".

[18] Mnanahesh kumar sabhnani et. al.,"KDD feature set complaint heuristic rule for R2L attack detection".

[19] Ali shiravi et.al.,"Toward developing a systematic approach to generate benchmark datasets for intrusion detection".

[20]H.Gunes Kayacik, A. Nur Zincir- Heywood, Malcolm I. Hey Wood.: "Selecting features for intrusion detection: A feature Relevance Analysis on KDD 99 Intrusion Detection Dataset".