Network Intrusion Detection based on Feature set Selection using Back-Propagation Neural Network

[1]Supratim Paul, [2]Dr. R. Nagaraja, [3]Shivakumar B R

[1]Student, [2]Professor & PG Coordinator, [3]Assistant Professor
Department of Information Science and Engineering,
Bangalore Institute of Technology, Bengaluru, Karnataka, India

***Abstract*: With dynamic increase of network application and electronic gadgets such as PCs, cell phones, and so on attacks and detection of intrusion became most challenging task in cybercrime detection zone. Decades ago, because of better network technology and more utility of the Internet, it became all digital in worldwide. Parallel to these enhancements, the endeavors of hackers for intruding the networks are also increased. And these attacks adversely affect the network badly. In this project for intrusion detection, an Artificial Neural Network (ANN) classification algorithm is implemented. The Network Layer-Knowledge Discovery Database data set is analyzed that contains of 41 features to study the viability of classification algorithm and intrusion attack detection is made by using Back-Propagation method to achieve more accuracy.**

Artificial neural network model gives the best number of features of 22 attributes need for detecting the intrusion and which is further categorized into more and less significant attributes. Hence, improves the accuracy and saving of resources is done.

***Keywords*: Intrusion Detection System, Artificial Neural Network, Feature Selection**

I. INTRODUCTION

Nowadays communication became one of the important parts for every living people and these networks are utilized mostly in processing of data for business, learning and education purpose and so on. That led to development of robustious communication in the networking field. This adaptability of protocols make vulnerable to network attacks initiated by intruders. Thus it requires continuous observation to make the network safely. And all this observation is carried out by IDS. Since, then internet expanded its field in various application domains like banking operations, social network and so on. This is because of the system security weakness so that many PCs are been hacked by unauthorized one either by denial of service attack or probing attack and so forth. Generally, firewalls and other control mechanisms are used for prevention but it seems all are failed if the hackers are within the network itself. So, IDS must be implemented in such a way that it gets adapt to the updated attacks. IDSs are the device that observes the activities to distinguish the malicious or suspicious events. It acts like a sensor. It can perform variety of functions such as checking clients and system activities, inspecting system configuration, surveying the integrity of critical system and information records, distinguishing unusual activity and revising system configuration errors.

II. MOTIVATION

Since today, all are reliant on PC innovations in any way. So the utilization of innovations has been build, risk related with computer technology is additionally increases. System security has become enormous challenge. There are such a significant number of system security tools available for example antivirus, firewall, etc. Yet they are not able to cover all security risks in the network system. Intrusion detection is characterized to be the issue of identifying people who are utilizing PC without any kind of authorization and the individuals who have legitimate access to the system yet are exceeding their benefits.

III. PROBLEM STATEMENT

With the quick improvement of data innovation in recent years, system networks have been broadly utilized by industrialization, business purpose and in different areas of living people. Subsequently, constructing dependable network system is the main task for IT heads. Then again, these create several challenges that made very difficult task to build dependable network system. Numerous attack types are undermining the availability, integrity and confidentiality of the system. The Denial of Service attack (DoS) is known to be most widely recognized vulnerable attack.DoS attacks generally consume the available resources in the network and overload unnecessary requests that are not required. So, in this way DoS became a big platform for wide ranges of attack whose goal is to consume system and its resources. Moreover, R2L attack also form platform for wide ranges of attack whose intention for local right consents as a result of the accessibility of some network resources.

IV. LITERATURE SURVEY

Numerous functions for training are there in literature. The functions it took is the one which been used mostly in the model like trainbr, trainc, traincgp, trainlm, trainoss, trainr and trainscg. Thus, gives the relative outcomes among IDS approaching with multi-layered neural network [1]. In this way, it has tried 7 unique functions having similar data set size. In any case, because of the randomly selected data set it experienced various performance and execution values. Moreover, because of execution time was high few functions are tested 10 times and few tested 100 times.

In feature selection approach for IDS using the existing feature selection algorithms and compared the output using WEKA tool by conducting several experiments on KDDCup99 network intrusion dataset [2]. So, tests and comparison are done on KDDCup99 dataset. It could lessen 70% of the feature dimension space and roughly 55% decrease in training time. In any case, characterization precision expands from 61.39% to 66.80% in identifying attacks. So, 5% increment and 66% classification accuracy as overall which is quite low accuracy.

In [3] utilizes the NSL data for uncovering protocol which is vulnerable and much of the time utilized intruders to launch network based intrusion. The examination in [3] demonstrates that simulation and performance wise NSL data set are mostly used for IDS. It additionally uncovered that major attacks are initiated utilizing as in-built disadvantages of the TCP/IP protocol.

In [4], many implementations has been done and executed to assess the effectiveness for the following classifier based on machine learning: J48, Random Forest, Random Tree, Decision Table, Naive Bayes and so forth and all of these tests depends on the KDD dataset. This [4] shows no such algorithm based on machine learning that deal with proficiently for all types of attacks and in addition, Bayes classifier has highest precision rate of 93%.

V. METHODOLOGY

The system proposed a model known as artificial neural network. These neural networks are the system with data processed that are built and executed to demonstrate the brain of the people. The fundamental goal of the neural system is to build up a computational device for demonstrating or modeling the human brain to perform different computational undertaking at a faster rate than the traditional systems and furthermore to make aware of different ways it may offer solutions, like processing languages, recognition of characters, pattern recognition and so forth.

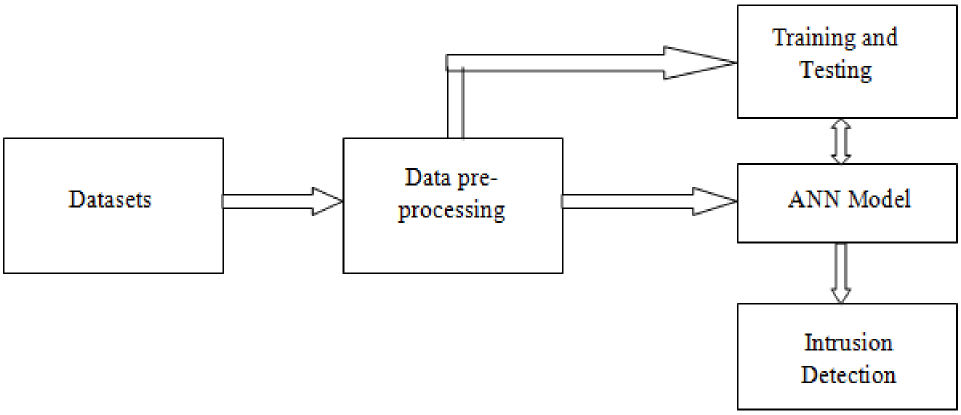


Fig. 1 The architecture of proposed system

Artificial neural network perform different undertaking for example matching patterns and categorization, optimize function, estimation, quantization, and clustering of data. This undertaking is extremely troublesome for PCs, which are faster in algorithmic computational errands and precise arithmetic operations. Therefore, for implementation of artificial neural networks, high-speed digital computers are used, which makes the simulation of neural processes feasible.

VI. IMPLEMENTATION

Here, in this section it describes the detail of each stages implemented.

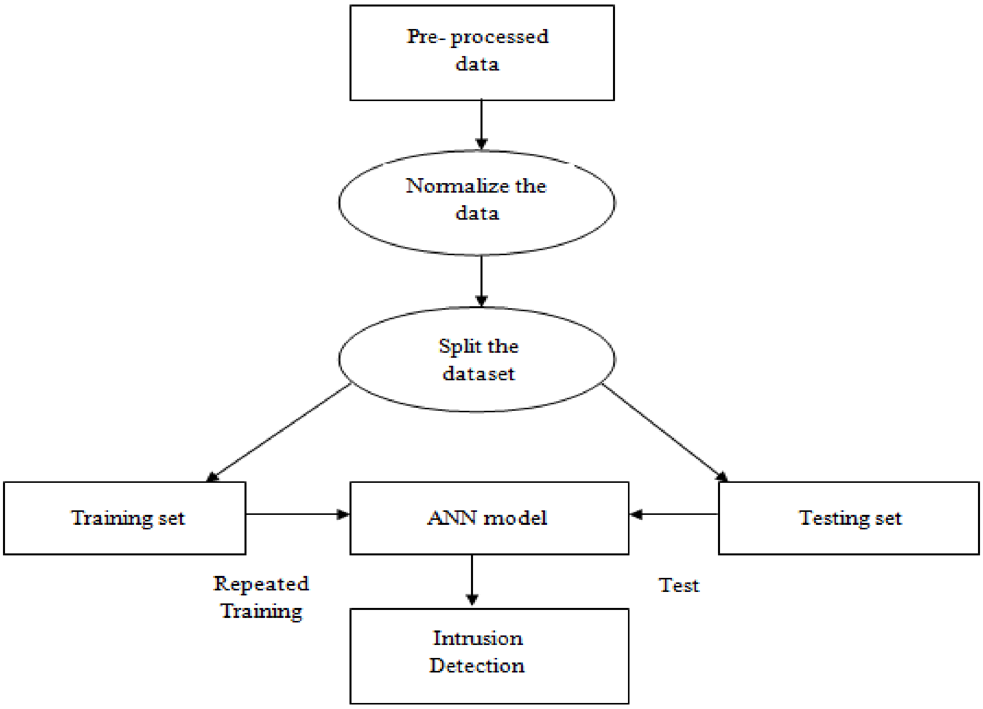


Fig. 2 Dataflow diagram of system

a. Pre-processed data

More than 50% of the instances are attacks, thus, in order to achieve a reduced attack rate and a balanced distribution of normal and attack classes were done. So, there is a need to adjust the number of instances for normal and attack classes. Around 50% instances are selected as normal connection and remaining 50% as the attack connection in which they are categorized in to four distinct types of attacks knowingly denial of service attack, probe attack, remote to local attack and user to remote attack. These data containing attacks are separated into two portions with four types of attack and normal types.

b. Normalized data

Neural network utilize just numeric data and in a similar range for ANN to give a precise outcome. Standardization process is done based on numeric type value within similar scale.
There are three features in every packets where it is having character type values such as protocol_type, service and flag which is need to be changed over to numeric type value for every time this features are keeps on repeating, then according to number of time it has been repeated the features are assigned with ascending manner like the feature given 1 whose repeated time is greater. Similarly, the feature is given 2 for having less repetition and so on.
Next, neural network utilizing numeric data must be in the similar range to give a precise outcome, normalization stage must be done for those features in each packet on NSL-KDD dataset.

c. Splitting of dataset

The pre-processed data must be split into training and a test set. There are two competing concerns here: with less training data, the parameter estimates have higher variance (estimates how close are the learned neural network model with weights from the true one). With less testing data, the performance statistic will have greater variance. Care must be taken that neither variance is too high.

Here, the datasets are splits into 70% for training data and rest 30% for test data for all datasets i.e. for Dataset (23,000 samples), Dataset (50,000 samples) and Dataset (75,000 samples) with epoch of 50.

It must be noted that while choosing the training set it must cover a wide range of input patterns so that the model can be trained well.

d. Training set

ANN model is trained by back-propagation which is a supervised learning algorithm. The model is trained a number of times with the same training set. Each time the model sees the entire training dataset in the training process is called as an **Epoch**. One epoch is equivalent to one forward pass and one backward pass of all the training illustration. Here model is trained using batch processing method where, we do not take one row of data at a time rather we send data in batches of size 30. And here we take 70% of the instances to train.

e. Test set

Here, the test data is tested which contains only 30% of the total instances and tells the result whether the particular packet belongs to normal or one of the attack classes.

f. ANN model

An ANN model is implemented as shown in the fig. 3

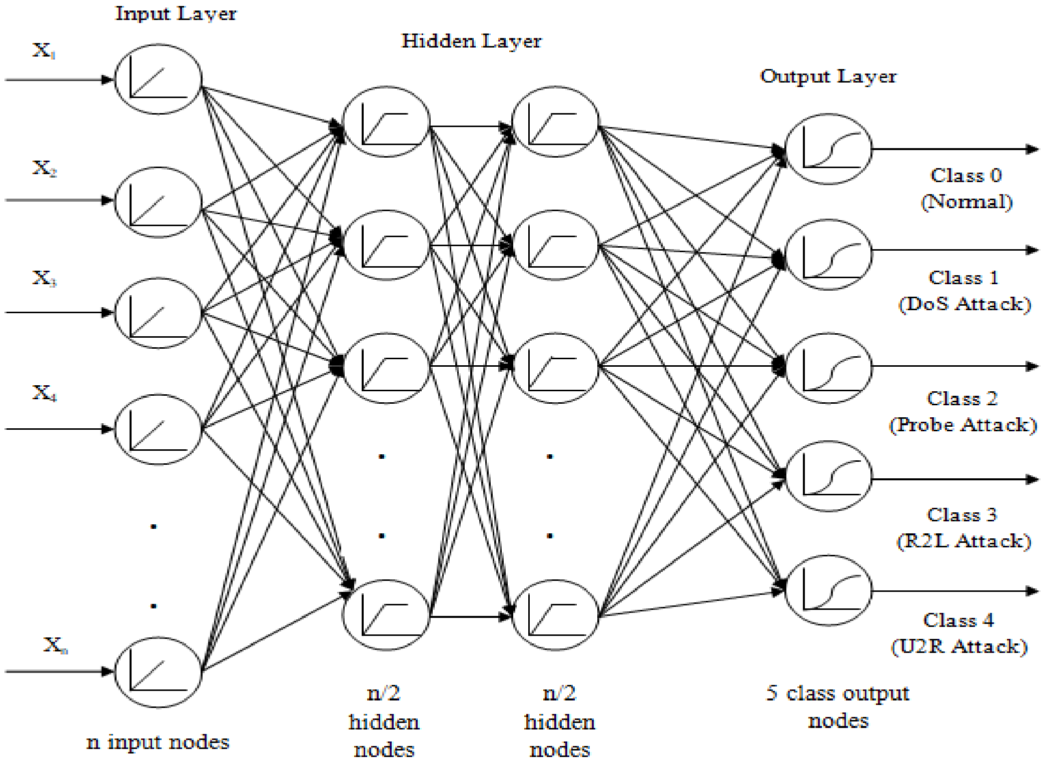


Fig. 3 Implementation of ANN model for intrusion detection classifier

As it is been seen from the above figure there consists of 42 inputs including 41 network features and last belongs to the class of attack which is then feed into the ANN model with two layers of hidden layers and finally it is been categorized from the output layer into 5 classes.
Hidden layers are considered with each having half of the input nodes. There are five output nodes as the model indicating whether packet contains either the type of attacks or the normal class.

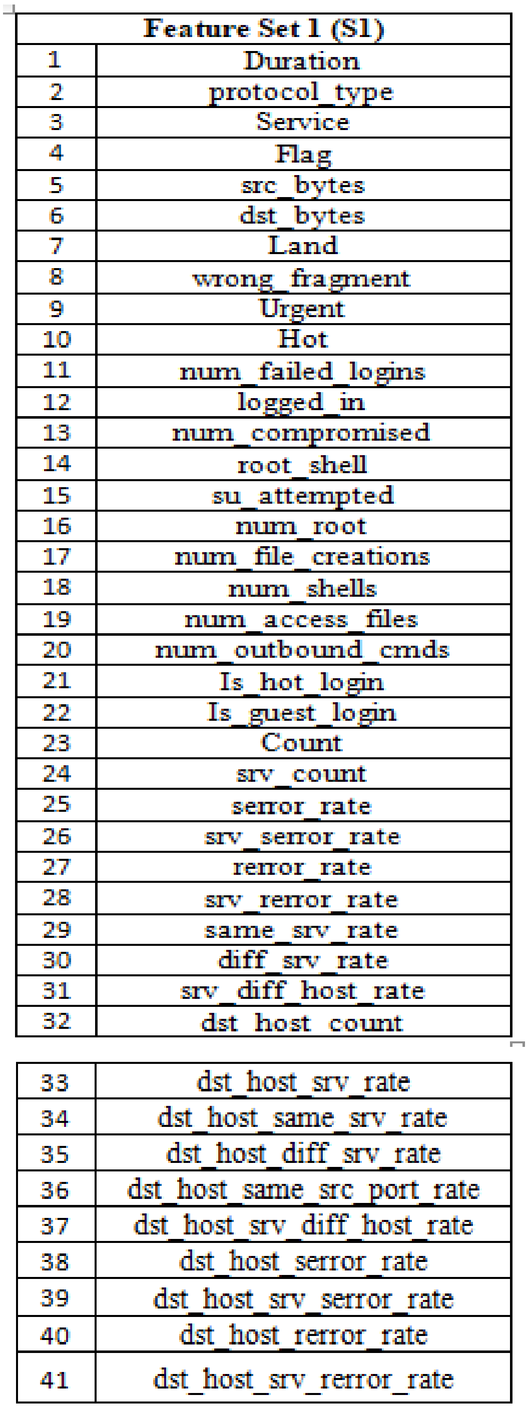
g. Intrusion detection

Detection module is the most vital phase that was proposed and whose role is to examine and identify intrusion utilizing back propagation network. And this model is used to detect intruders and network flexibility if so provided. So, it analyses and detects each and every packet either as normal or attacked classes.

VII. RESULTS AND SNAPSHOTS

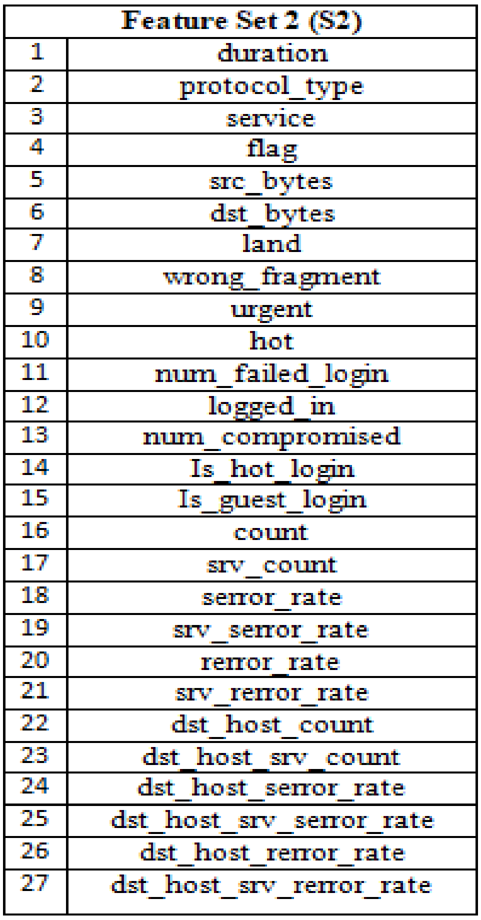
This section consists of snapshots of simulation carried out of NSL data set and comparison of precision by changing some parameter or attributes.

Table 1: List of 41 Features Set



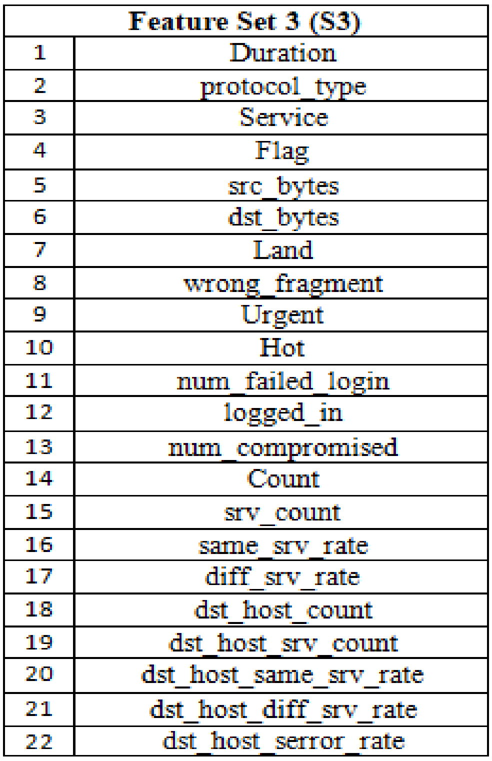
Feature Set 1 (S1)	
1	Duration
2	protocol_type
3	Service
4	Flag
5	src_bytes
6	dst_bytes
7	Land
8	wrong_fragment
9	Urgent
10	Hot
11	num_failed_logins
12	logged_in
13	num_compromised
14	root_shell
15	su_attempted
16	num_root
17	num_file_creations
18	num_shells
19	num_access_files
20	num_outbound_cmds
21	Is_hot_login
22	Is_guest_login
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
32	dst_host_count
33	dst_host_srv_rate
34	dst_host_same_srv_rate
35	dst_host_diff_srv_rate
36	dst_host_same_src_port_rate
37	dst_host_srv_diff_host_rate
38	dst_host_serror_rate
39	dst_host_srv_serror_rate
40	dst_host_rerror_rate
41	dst_host_srv_rerror_rate

Table 2: List of 27 Features Set



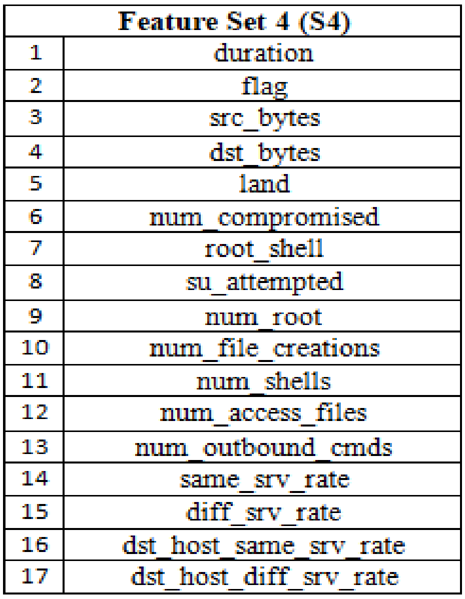
Feature Set 2 (S2)	
1	duration
2	protocol_type
3	service
4	flag
5	src_bytes
6	dst_bytes
7	land
8	wrong_fragment
9	urgent
10	hot
11	num_failed_login
12	logged_in
13	num_compromised
14	Is_hot_login
15	Is_guest_login
16	count
17	srv_count
18	serror_rate
19	srv_serror_rate
20	rerror_rate
21	srv_rerror_rate
22	dst_host_count
23	dst_host_srv_count
24	dst_host_serror_rate
25	dst_host_srv_serror_rate
26	dst_host_rerror_rate
27	dst_host_srv_rerror_rate

Table 3: List of 22 Features Set



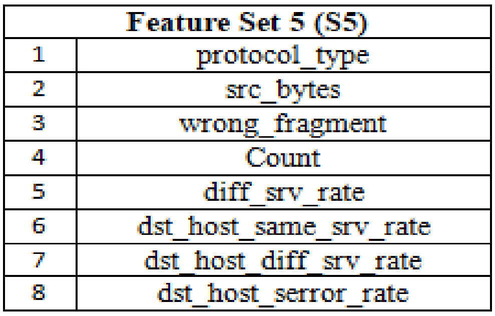
Feature Set 3 (S3)	
1	Duration
2	protocol_type
3	Service
4	Flag
5	src_bytes
6	dst_bytes
7	Land
8	wrong_fragment
9	Urgent
10	Hot
11	num_failed_login
12	logged_in
13	num_compromised
14	Count
15	srv_count
16	same_srv_rate
17	diff_srv_rate
18	dst_host_count
19	dst_host_srv_count
20	dst_host_same_srv_rate
21	dst_host_diff_srv_rate
22	dst_host_serror_rate

Table 4: List of 17 Features Set



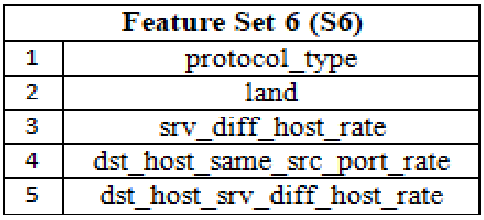
Feature Set 4 (S4)	
1	duration
2	flag
3	src_bytes
4	dst_bytes
5	land
6	num_compromised
7	root_shell
8	su_attempted
9	num_root
10	num_file_creations
11	num_shells
12	num_access_files
13	num_outbound_cmds
14	same_srv_rate
15	diff_srv_rate
16	dst_host_same_srv_rate
17	dst_host_diff_srv_rate

Table 5: List of 8 Features Set



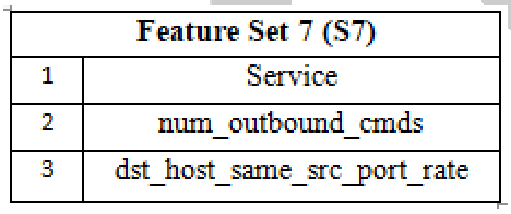
Feature Set 5 (S5)	
1	protocol_type
2	src_bytes
3	wrong_fragment
4	Count
5	diff_srv_rate
6	dst_host_same_srv_rate
7	dst_host_diff_srv_rate
8	dst_host_serror_rate

Table 6: List of 5 Features Set



Feature Set 6 (S6)	
1	protocol_type
2	land
3	srv_diff_host_rate
4	dst_host_same_src_port_rate
5	dst_host_srv_diff_host_rate

Table 7: List of 3 Features Set



Feature Set 7 (S7)	
1	Service
2	num_outbound_cmds
3	dst_host_same_src_port_rate

To enhance the performance of IDS, these above feature sets are chosen for analysis. From the above Table 7.1 the first feature set incorporates all the features of NSL dataset. Then after, it is performed with combinations of basic features and time based features to determine the patterns.

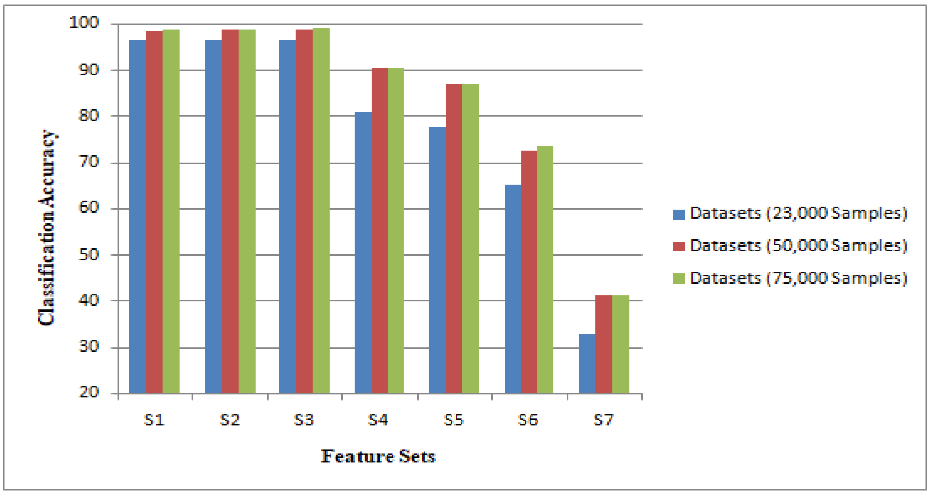


Fig. 4 Comparison bar graph for no. of samples

The above figure depicts the comparison graph among the three datasets having different number of samples i.e. 23,000 samples, 50,000 samples and 75,000 samples. From comparison bar graph, it summaries that feature sets having 22 attributes seems to have highest classification accuracy from all the above datasets.

Now we take the feature set with 22 attributes for further in to experiment to find out which attribute gives more and less significant value for classification accuracy among the 22 attributes.

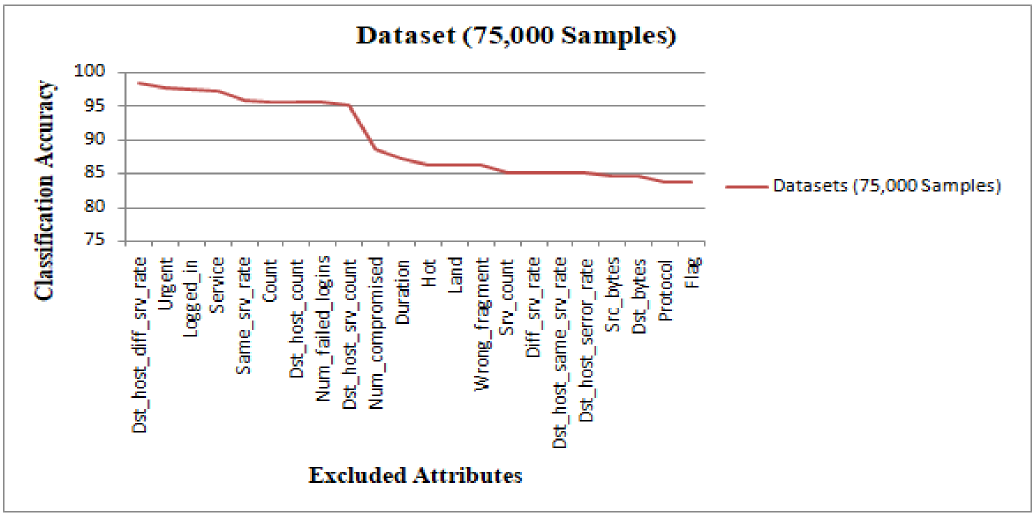
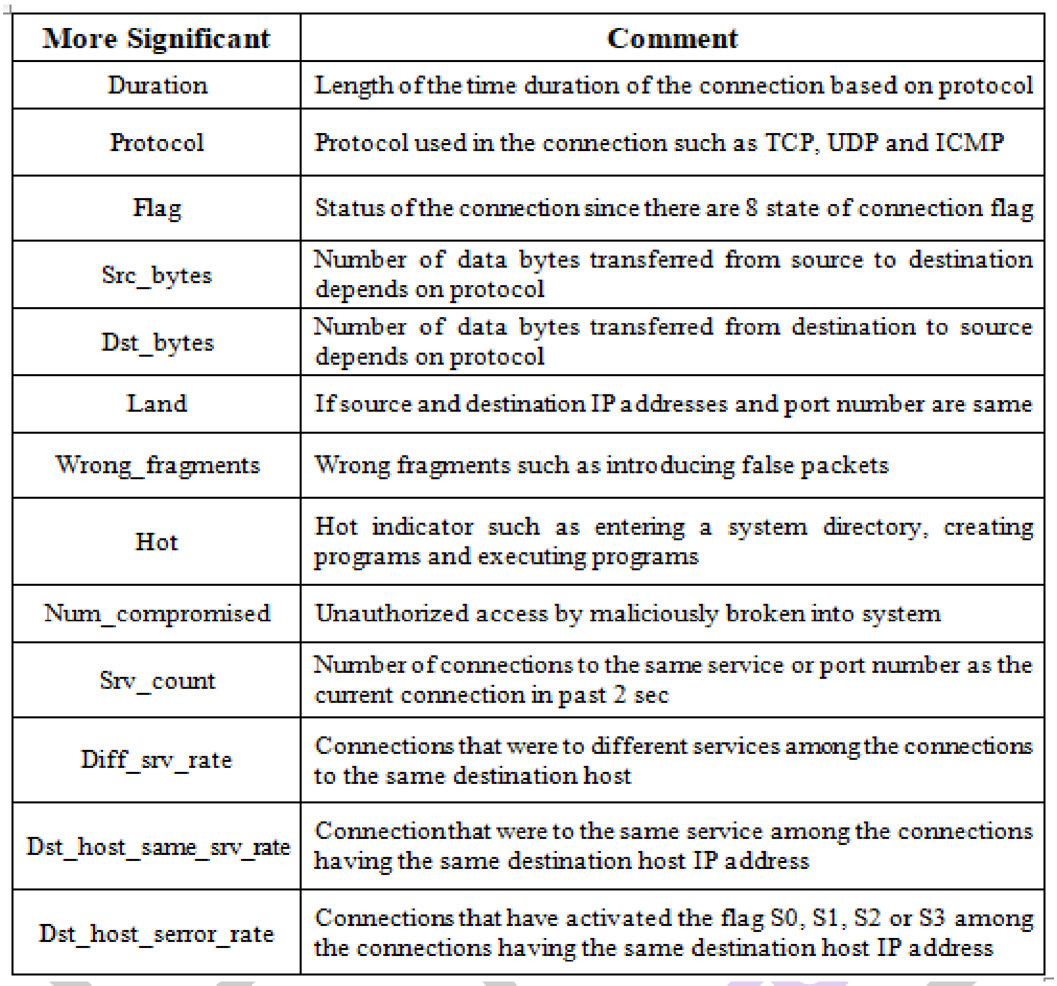
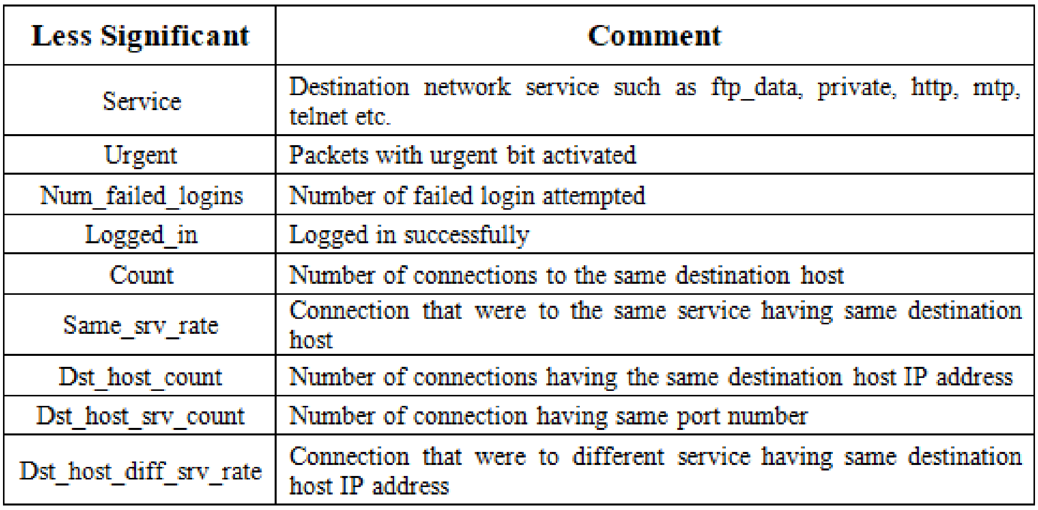


Fig. 5 Classification accuracies graph of excluded attributes of feature set S3

The above figure shows the classification accuracies graph of feature set with 22 attributes after eliminating each attribute from 22 attributes from dataset (75,000 samples). The attributes one which gives more classification accuracy shows less significant value that means it contribute less value to the overall classification accuracy. And the attributes which gives less classification accuracy shows more significant value that means it contribute more value to the overall classification accuracy. Thus, the attributes more than 90% classification accuracy from Fig. 5 is classified as Less Significant and those attributes less than 90% classification accuracy are classified as More Significant attributes.



More Significant	Comment
Duration	Length of the time duration of the connection based on protocol
Protocol	Protocol used in the connection such as TCP, UDP and ICMP
Flag	Status of the connection since there are 8 state of connection flag
Src_bytes	Number of data bytes transferred from source to destination depends on protocol
Dst_bytes	Number of data bytes transferred from destination to source depends on protocol
Land	If source and destination IP addresses and port number are same
Wrong_fragments	Wrong fragments such as introducing false packets
Hot	Hot indicator such as entering a system directory, creating programs and executing programs
Num_compromised	Unauthorized access by maliciously broken into system
Srv_count	Number of connections to the same service or port number as the current connection in past 2 sec
Diff_srv_rate	Connections that were to different services among the connections to the same destination host
Dst_host_same_srv_rate	Connectionthat were to the same service among the connections having the same destination host IP address
Dst_host_serror_rate	Connections that have activated the flag S0, S1, S2 or S3 among the connections having the same destination host IP address



Less Significant	Comment
Service	Destination network service such as ftp_data, private, http, mtp, telnet etc.
Urgent	Packets with urgent bit activated
Num_failed_logins	Number of failed login attempted
Logged_in	Logged in successfully
Count	Number of connections to the same destination host
Same_srv_rate	Connection that were to the same service having same destination host
Dst_host_count	Number of connections having the same destination host IP address
Dst_host_srv_count	Number of connection having same port number
Dst_host_diff_srv_rate	Connection that were to different service having same destination host IP address

Table 8: More and Less significant attributes

VIII. CONCLUSION

The result is analyzed based on network layer data sets which demonstrate that it is one of the good competitor data set for simulation and test performance of intrusion detection system. This ANN model increases the precision and decreases the detection time. And the analysis led on network layer data sets with the combination of diagrams and list of tables helps to understand more clearly. ANN model gives the best number of attributes need for detecting the intrusion. And in that it shows the more and less significant features that contribute for overall classification accuracy. So, after trying with different number of attributes, among that the attribute having 22 features gives classification accuracy of around 99%. And from that 22 features, 13 features gives more significant while rest 9 features gives less significant to the overall classification accuracy. These more significant features give classification accuracy of 96%.

REFERENCES

[1] Gozde Karatas, Ozgur Koray Sahingoz, “Neural Network Based Intrusion Detection Systems with Different Training Functions”, IEEE 2018

[2] Krishan Kumar, Gulshan Kumar, Yogesh Kumar, “Feature Selection Approach for Intrusion Detection System”, IJATCSE, Vol. 2, 2013

[3] L. Dhanabal, Dr. S. P. Shantharajah, “A Study on NSL-KDD Dataset for Intrusion Detection System Based On Classification Algorithm”, IJARCCE, Vol. 4, 2015

[4] Mohammad Almseidin, Maen Alzubi, Szilveszter Kovacs, Mouhammd Alkasassbeh, “Evaluation of Machine Learning Algorithms for Intrusion Detection System”, IEEE 2017

[5] Takwa Omrani, Adel Dallali, Bilgacem Chibani Rhaimi, Jaouhar Fattahi, “Fusion of ANN and SVM Classifiers for Network Attack Detection”, IEEE 2018

[6] Ms Pooja Bhoria, Dr. Kanwal Garg, “Determining feature set of DOS attacks”, IJARCSSE, Vol. 3, 2013

[7] Laheeb M. Ibrahim, Dujan T. Basheer, Mahmod S. Mahmod, “A Comparison Study for Intrusion Database (KDD99, NSL-KDD) Based on SOM ANN”, Journal of Engineering Science and Technology, Vol. 8, 2013