ISSN: 2582 - 6379
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2020, Volume 1 Issue 2
www.ijisea.org

Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection

[1] Tadikonda Mounishanjali, [2] Kumpati Sandhya, [3] Sala Rajesh,
[4] Natta Muniyya ,[5]K. Naga Prakash

[1234] Student(s),Department of ECE
[5]Professor, Department of ECE,
Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh, India.

ABSTRACT

This Paper includes, a supervised machine learning system is developed to classify network traffic whether it is malicious or benign. To find the simplest model considering detection success rate, combination of supervised learning algorithm and have selection method are used. Through this study, it's found that Artificial Neural Network (ANN) based machine learning with wrapper feature selection outperform support vector machine (SVM) technique while classifying network traffic. To evaluate the performance, NSL-KDD dataset is employed to classify network traffic using SVM and ANN supervised machine learning techniques. Comparative study shows that the proposed model is efficient than other existing models with reference to intrusion detection success rate.

Keywords: Accuracy, ANN, Network Intrusion Systems, NSL-KDD dataset, Supervised SVM.

Abbreviations: ANN, Artificial Neural Networks; SVM, Support Vector Machine; NSL, Network security Laboratory.

I.INTRODUCTION

As in modern world computers have become the part of life, we store the memory in smart appliances like computers, smart phones, cars and locks. The numbers of unidentified attacks also increase day by day. So, it is our responsible to safeguard the applications and personal data from the intruders' or accessing the data without proper permission. The security of the system is at risk with the intrusion actions. An intrusion can be detailed as the action of force causing damage. There are only some intrusion detection techniques that can detect the threat and caution the system memory from the intrusion. As the systems became more complex, there are always available weakness in the systems due to the design and programming errors. Therefore Intrusion detection with feature selection is used to protect our smart appliances from such type of unprotected attacks.

II.RELATED WORK

Several studies are carried out on Intrusion Detection System since last 26 year. It is one of the imperative research area where more than 300 papers already published. M.Tavallaee surveyed on anomaly based intrusion detection and published his research work during the amount of 2000-08. In his research paper he had mentioned that, researchers uses their self-created dataset or they uses various publically available dataset like DARPA data , KDD cup'99 and NSL KDD dataset to spot attack or normal supported their classification accuracy, false positive rate or detection rate. Some of the researcher uses feature selection

ISSN: 2582 - 6379
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2020, Volume 1 Issue 2
www.ijisea.org

and reduction to scale back the dimensionality of dataset and it also improves the performance. Muhammad Imran et al applied K- resampling methods on 20% of NSL KDD training dataset for training and testing. Ibrahim et al. applied SOM on KDD 99 and NSL dataset and show the higher result of binary classification on KDD 99 dataset then that of NSL dataset. Bhoria et al. uses cart 4.5 to detect DOS attack. She applied resampling procedure used to evaluate machine learning techniques on 20% NSL KDD dataset for training and testing. The dataset of NSL KDD contains of 22,495 records with normal and DOS attack. Bajaj et al. applied information gain model for feature selection then applied J48, Naïve Bayes, NB tree, SVM and straightforward cart methods for binary classification. R.Patil et al. uses Adaboost machine learning on NSL KDD dataset.

III. DATASET DISCRIPTION

To prove the effectiveness of proposed system the dataset can be constructed by self or by extracting it from other sources:

- NSL-KDD dataset is an offline network dataset based on KDD 99 dataset

The proposed system is applied on NSL-KDD dataset which having 41 attribute and one class attribute. The NSL-KDD training set does not include redundant record and hence reduce the complexity level. The size of KDD99 which contain redundant records is less than that of NSL KDD. The advantages of NSL-KDD data set can be differed by the original KDD dataset by. The training is performed on KDD Train data which contain 22 attack types and testing is performed on KDD Test data which contains additional 17 attack type. These attacks can be in four different types with some common properties as shown in table I for training and testing. The four categories of attacks are:

- Denial of Service (DoS) – A malicious attempt to block system or network resources and services.
- Probe – This attack collects the information about potential vulnerabilities of the target system that can later be used to launch attacks on those systems.
- Remote to Local (R2L) – Unauthorized ability to dump data packets to remote system over network and gain access either as a user or root to do their unauthorized activity.
- User to Root (U2R) – In this, attackers access the system as a normal user and break the vulnerabilities to gain administrative privileges.

Table 1 : Shows Attack Categorization For Training And Testing Datasets

Attack Types	Category
Normal	Normal
apacha2	**DOS**
back	
land	
mailbomb	
netpune	
pod	
processtable	
smurf	
teardrop	
udpstorm	
buffer_overflow	
httprunnel	
loadmodule	

ISSN: 2582 - 6379
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2020, Volume 1 Issue 2
www.ijisea.org

perl	**U2R**
ps	
rootkit	
sqlattack	
xterm	
ftp-write	**R2L**
guess_password	
imap	
multihop	
named	
phf	
sendmail	
snmpgetattack	
snmpguess	
spy	
warezclient	
warezmaster	
worm	
xlock	
xsnoop	

IV. PROPOSED METHOD OF ANN

Artificial Neural Network (ANN) used for supervised classification learning. ANN is a computational model consists of a number of simple, highly interconnected neurons.

A. DATASET SELECTION: Number of dataset available in NSL-KDD dataset repository for training and testing, from those available dataset KDDTrain+.ARFF (dataset with binary labels) is used for training and KDDTest+.ARFF is used for testing the binary category problem. And for five category KDDTrain+.TXT and KDDTest+.TXT is used for training and testing dataset respectively. The test set included additional 17 attacks which are not in the train dataset.

B. Data Pre-processing: NSL-KDD dataset contain numeric and some nonnumeric attribute. Non numeric attribute like protocol type, service and flag attribute need to convert as numeric attribute because the training input and testing input is given to ANN should be numeric matrix. And the class attribute also labelled as numeric type, Here position of one, corresponds with row, denote the targeted class.

Normal	10000
DOS	01000
PROBE	00100
R2L	00010
U2R	00001

C. FEATURE REDUCTION: The NSL-KDD data set has 41 attribute and one class attribute. From those 41 attribute some attribute have no role and some have minimum role in detection of attack. Bajaj et al. [10] uses information gain attribute evaluation, gain ratio attribute evaluation and correlation attribute evaluation algorithm. The researcher [10] shows that attribute 9, 20 and 21 have no role and attribute 15, 17, 19, 32, 40 have minimum role in detection of attack. Observation of NSL-KDD dataset shows that features 7,8,11 and 14 have almost all zero values in dataset. Removing these entire least usable features from training and testing set of the dataset we left with 29 features, this reduces the size of the dataset. Now the reduced dataset is passed for training and testing. Training and testing also performed with 41 features, in this case feature reduction is not performed on dataset.

ISSN: 2582 - 6379
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2020, Volume 1 Issue 2
www.ijisea.org

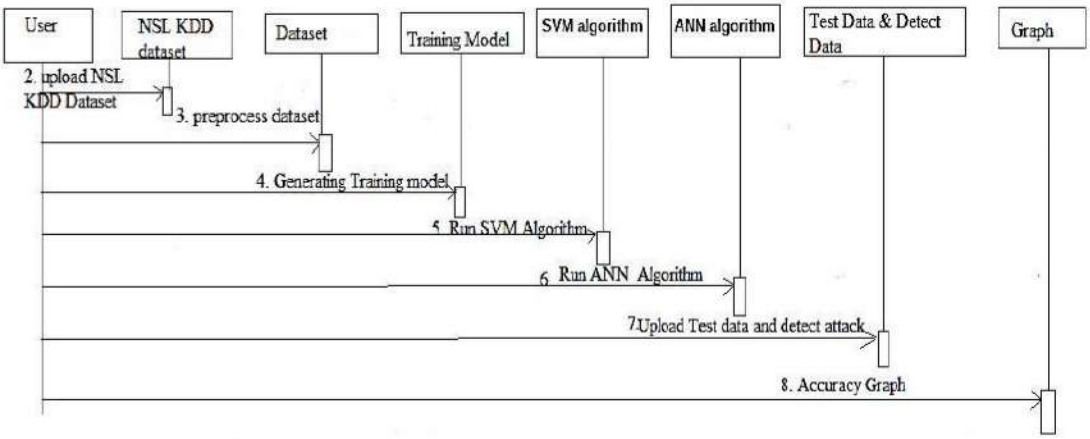

Figure 1: Shows Training and Testing

D. NORMALIZATION: Z-score normalization is used to normalize the attribute values. It normalizes the attribute value such that the mean and standard deviation after normalization become zero and one respectively. For this property of normalization z-score is also called as zero mean normalization. Its mathematical equation is given below.

$$a(i) = \frac{a(i) - mean(A)}{\text{Std(A)}}$$

Here A is the attribute and a(i) is the value of A that is going to be updated by above equation.

E. SELECT NEURAL NETWORK ARCHITECTURE: There is some field in neural network which have to be selected before training process. These fields are number of neuron, number of layer in case of multilayer, algorithm and transfer function for training the neural network.

F. TRAINING NEURAL NETWORK: The NEL-KDD train dataset have 125973 records. R2L and U2R have few patterns in their classes. For maintaining equality and for speed up the training 18718 patterns were selected. From the selected patterns, 17672 are chosen randomly from normal, DOS and probe class and all other from remaining classes. Training is performed on full featured dataset and feature reduced dataset.

G. TESTING NEURAL NETWORK: Test set of NSL-KDD dataset having some unknown attacks, which is not presenting in training set. The main task is to classify those attacks accurately. Neural network is tested by so many records nearly 22544 of full test dataset with or without features reduction as required.

H. RESULT EVALUATION: The performance of neural network can be evaluated using various parameters like accuracy, detection rate and false positive rate, the given parameter calculated using true positive (TP), false negative (FN), false positive (FP), true negative (TN).

- Detection Rate (DR) = TP/ (TP+FN)
- False Positive Rate (FPR) = FP/ (FP+TN)
- Accuracy (ACC) = (TP+TN)/(TP+TN+FP+FN)

ISSN: 2582 - 6379
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2020, Volume 1 Issue 2
www.ijisea.org

V. PROPOSED METHOD FOR SVM

1.DATA SET COLLECTION:

To verify effectiveness the feasibility of proposed IDS system, we used NSL-KDD dataset. It is a new version of KDDcup99 dataset NSL-KDD has more advantages as compared to KDDcup99 dataset. Some of the inherent problems are involved in KDDcup99. It consider as standard benchmark for intrusion detection evaluation. The training dataset of NSL-KDD and KDDcup99 is similar that consists of approximately 4,900,000 single connections vectors and that each one contains 41 features and is labelled as either normal or attack type.

Following reasons are helped to NSL-KDD that became more popular dataset than KDDcup99 dataset for intrusion detection purpose

- Redundant records from training set are eliminated
- Duplicate records from the test set are removed to improve the intrusion detection performance.
- Use of NSL-KDD dataset for experiment it consists of reasonable numbers of instance both in the training set and also in testing set.

A. DATA SET PRE-PROCESSING

- Pre-processing of original NSL-KDD dataset is mandatory for making input to SVM. Dataset pre-processing can achieved by applying following procedure
    - Data set transformation
    - Data set normalization
    - Data set discretization

2. DATA SET TRANSFORMATION:

The training dataset of NSL-KDD consist of 4,900,000 single connections instances and each connection instance contain 42 features to numeric values that helps to provide suitable input for classification using SVM. This transformation is understood by table2 and also we have to assign numeric value to the last feature in the connection instance doing this we assigned a class “zero” for normal connections and a one for any deviation from that as from transformation table 2 some useless data will filtered and also modified some text items needed to converted into numeric values every instance in the dataset has 42 feature or attribute including target class can observed in Table 2.

Table 2 : Shows Features of NSD-KDDCUP99 Dataset

S.No	Feature Name
1.	Duration
2.	Protocol_type
3.	Service
4.	Flag
5.	Src_bytes
6.	Dst_bytes
7.	Land

ISSN: 2582 - 6379
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2020, Volume 1 Issue 2
www.ijisea.org

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_creation
18.	Num_shells
19.	Num_access_files
20.	Num_outbound_cmds
21.	Is_host_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_count
34.	Dst_host_same_srv_rate
35.	Dst_host_diff_srv_rate
36.	Dst_host_same_srv_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
42.	Normal or Attack

An example for original NSL-KDD data set record is shown in Table 3.

ISSN: 2582 - 6379
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2020, Volume 1 Issue 2
www.ijisea.org

Table 3 : Shows The Original NSL KDD cup 99 dataset

0 tcp ftp_data SF491 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 2 0 0 0 0 1 0 0 150 25 0.17 0.03 0.17 0 0 0 0.05 0 normal
0 udp other SF 146 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 13 1 0 0 0 0 0.08 0.15 0 255 1 0 0.6 0.88 0 0 0 0 0 normal

Table 4 : Transformation Table

Type	Feature Name	Numeric Value
Attack or Normal	Normal	0
Attack or Normal	Attack	1
Protocol_type	TCP	2
	UDP	3
	ICMP	4
Flag	OTH	5
	REJ	6
	RSTO	7
	RSTOS0	8
	RSTR	9
	S0	10
	S1	11
	S2	12
	S3	13
	SF	14
	SH	15
Service	All services	16 to 81

The nominal values like http, tcp, SF in dataset. If we want to transform these nominal values to numeric values in advance. The service type of 'tcp' is mapped to 2,'udp' is mapped to 3, 'icmp' is mapped to 4 and the Table 5 will helps to transform the nominal values of dataset features into the numeric values. After transformation, the original NSLKDD cup99 dataset will become as shown in Table 5.

Table 5 : Shows Original NSL- KDD cup'99 dataset after transformation

0,2,32,14,491,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0,0,0,0,1,0, 0,150,25,0.17,0.03,0.17,0,0,0,0.05,0,0
0,3,16,14,146,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,13,1,0,0,0,0,0. 08,0.15,0,255,1,0,0.6,0.88,0,0,0,0,0,0

ISSN: 2582 - 6379
International Journal for Interdisciplinary Sciences and Engineering Applications
IJISEA - An International Peer- Reviewed Journal
2020, Volume 1 Issue 2
www.ijisea.org

3. DATA SET NORMALIZATION: Dataset normalization is very necessary to enhance the performance of intrusion detection system when datasets are too large. In this we used the min-Max method of normalization.

4. DATA SET DISCRETIZATION: Dataset discretization technique helps to selecting features continuously for intrusion detection and to create some homogeneity between the given values, which have different data types. Here, we have used range discretization technique for this purpose.

VI . RESULTS

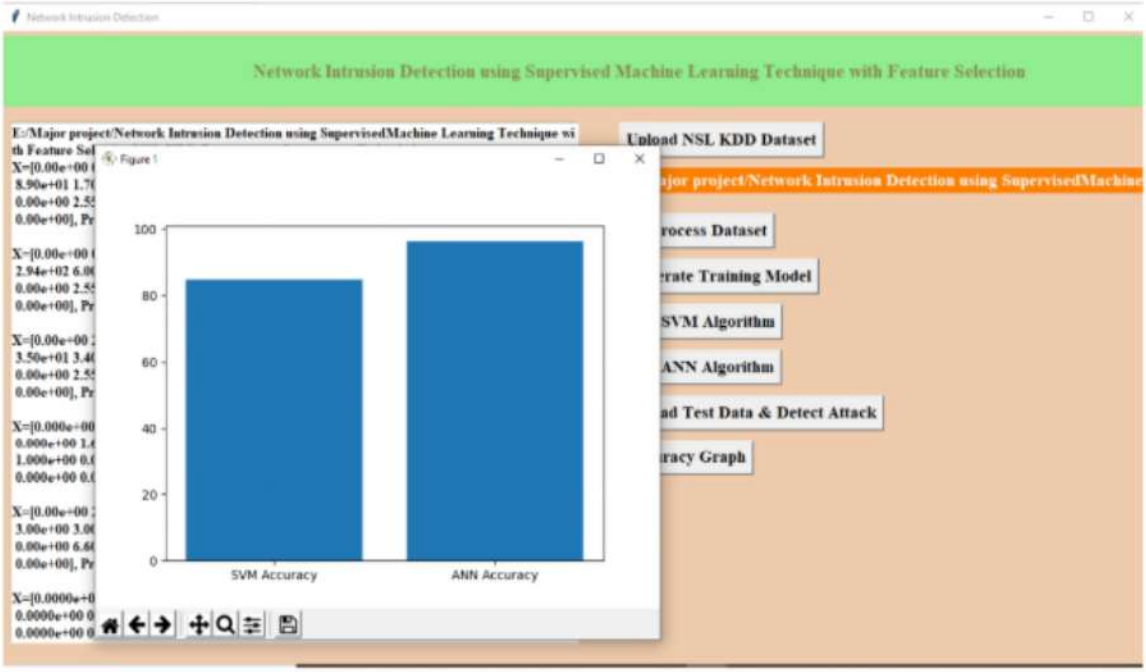

Figure 2: Shows the Experiment Result

From the above graph we noticed that, the SVM have 86.7 accuracy and ANN have the 96.8 accuracy. So, the detection capability of ANN is more as compared to SVM. In the above graph x-axis contains algorithm name and y-axis represents the accuracy of that algorithms.

VII. CONCLUSION

In this report, we have presented different machine learning models using different machine learning algorithms and different feature selection methods to find a best model. The analysis of the result shows that the model built using ANN and wrapper feature selection outperformed all other models in classifying network traffic correctly with detection rate of 94.02%. We believe that these findings will contribute to research further in the domain of building a detection system that can detect known attacks as well as novel attacks. The intrusion detection system exist today can only detect known attacks. Detecting new attacks or zero day attack still remains a research topic due to the high false positive rate of the existing systems.

REFERENCES :

[1] H. Song, M. J. Lynch, and J. K. Cochran, "A macro-social exploratory analysis of the rate of interstate cyber-victimization," American Journal of Criminal Justice, vol. 41, no. 3, pp. 583–601, 2016.

[2] P. Alaei and F. Noorbehbahani, "Incremental anomaly-based intrusion detection system using limited labeled data," in Web Research (ICWR), 2017 3th International Conference on,2017, pp. 178–184.

[3] M. Saber, S. Chadli, M. Emharraf, and I. El Farissi, "Modeling and implementation approach to evaluate the intrusion detection system," in International Conference on Networked Systems, 2015, pp. 513–517.

[4] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, "Toward credible evaluation of anomaly-based intrusion-detection methods," IEEE Transactions on Systems, Man, and Cybernetics, PartC (Applications and Reviews), vol. 40, no. 5, pp. 516–524, 2010.

[5] A. S. Ashoor and S. Gore, "Importance of intrusion detection system (IDS)," International Journal of Scientific and Engineering Research, vol. 2, no. 1, pp. 1–4, 2011.

www.ijisea.org