International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2020): 7.803

MLP and RNN Based Intrusion Detection System Using Machine Learning with Stochastic Optimization

Mithlesh Kumar[1], Gargishankar Verma[2]

[1]M. Tech, Computer Technology, Columbia Institute of Engineering & Technology, Raipur, Chhattisgarh, India

[2]Associate Professor, Computer Science & Engineering Department, Columbia Institute of Engineering & Technology, Raipur, Chhattisgarh, India

Abstract: ***With common innovations like Internet of Things, Cloud Computing and Social Networking, a lot of traffic from these networks are produced. Thus, there is a requirement for Intrusion Detection Systems that screens the traffic and breaks down them progressively. In this paper, NSL - KDD is utilized to assess the AI calculations for Intrusion Detection (ID). The dataset are taken from publicly available data of different types of attacks in the network. Consequently, lessening and choosing a specific arrangement of components work on the speed and precision. Along these lines, features are chosen by utilizing Feature scaling and other ML approaches. We have directed a thorough trial on Intrusion Detection System (IDS) that utilizations AI calculations, in particular, MLP and RNN. We have utilized the previous model RNN with the MLP combined with some Stochastic Optimization. The proposed system architecture performs well in the commodity hardware.***

Keywords: Intrusion Detection, IDS, Network Infiltration. Multi Layer Perceptron, Recurrent Neural Network, Machine Learning

1. Introduction

With the quick advancement of data innovation in the beyond twenty years. PC networks are generally utilized by industry, business and different fields of the human existence. Along these lines, building dependable organizations is a vital assignment for IT overseers. Then again, the fast improvement of data innovation created a few difficulties to assemble dependable organizations which is a truly challenging assignment. There are many sorts of assaults undermining the accessibility, respectability and classification of PC organizations. The Denial of Service attack (DOS) considered as quite possibly the most well - known destructive attack.

The point of DOS attack is to briefly prevent a few administrations from getting the end clients. By and large, it for the most part devours network assets and over - burdens the framework with undesired solicitations. Consequently DOS goes about as a huge umbrella for a wide range of assaults which expect to devour PC and organization assets. [1] In 2000 Yahoo was the main survivor of a DOS assault and in a similar date likewise DOS recorded its very first assault openly. Right now, web administrations and social sites are focus of DOS assaults [2]. According to another point of view, the remote to local (R2L) assaults are one more umbrella for a wide range of assaults which are intended to have neighborhood right authorizations on the grounds that the accessibility of some organization assets is just exceptional for the nearby clients for example document server. There a few are sorts of R2L assaults for example SPY and PHF, these sorts of attacks aim to get ready unlawful admittance to the organization assets [3].

As a rule, there are two sorts of IDS. Irregularity interruption recognition framework executed to recognize assaults dependent on recorded typical conduct. Consequently, it contrasts the current ongoing deals and past recorded ordinary continuous deals, this sort of interruption recognition framework is generally utilized in light of the fact that it can distinguish the new kind of interruptions. However, according to another viewpoint, it enlists the biggest upsides of bogus positive alert, which implies there is an enormous number of typical bundles considered as assaults parcels. In any case, abuse interruption location framework is carried out to recognize assaults dependent on vault of assaults marks. It has no bogus caution and yet, the new kind of assault (new signature) can prevail to go through it.

With respect to writing [5] attacks recognition considered as arrangement issue on the grounds that the objective is to explain whether the bundle either typical or assault parcel. In this manner, the model of acknowledged interruption recognition framework can be carried out dependent on critical AI calculations.

2. Literature Review

Zaman and Karray [2], In this paper, highlights are positioned dependent on loads. Two calculations have been proposed: Forward Selection positioning and Backward end positioning.

Jha and Ragha [3], This paper clarifies the restrictions of SVM. Then, NSL KDD is preprocessed and provisions are positioned situated in Information Gain Ratio. Then the model is prepared utilizing SVM.

Revathi and Malathi [4], AI calculations applied on the NSL KDD dataset taking every one of the provisions. Rehashed something similar subsequent to choosing highlights utilizing Correlation based Feature Selection Technique.

Examination of the not set in stone that Random Forest has the most elevated precision

Revathi and Malathi [4], This paper determined and thought about measures like exactness, accuracy, RMS mistake, and so on of AI calculations. It additionally analyzes the genuine positive and bogus positive rates.

Thanthrige et. al. [6], This paper determined and thought about measures like exactness, accuracy, RMS mistake, and so on of AI calculations. It additionally analyzes the genuine positive and bogus positive rates.

Anwer et al. [7], Presents a structure that utilizations channel and covering highlights determination strategies, to choose the most un - number of elements that accomplish the best precision. UNSWNB15 dataset is utilized and J48 choice tree classifier is applied.

3. Methodology

Intrusion detection system is a software application that monitors networks for malicious activities or unauthorized access. For the real - time monitoring of these malicious activities, machine learning approaches are utilized to train the model, so that a packet is dropped when it is found as a malicious packet. For training the model, various steps are shown in Figure 1 which includes data collection, pre processing, feature selection, model for training and simulation are presented.

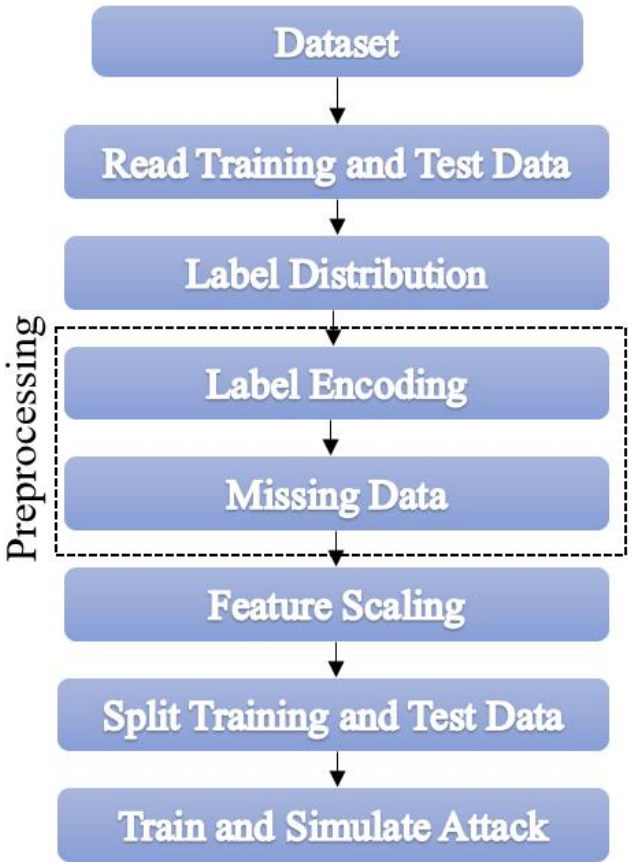


Figure 1: System architecture

a) Dataset

For Intrusion Detection, NSL - KDD [1] dataset is utilized which comprises of 41 components in each preparation and testing dataset. NSL - KDD dataset has excess information and absolute provisions that should be pre handled.

b) Data Pre - processing

Information pre - handling changes crude information into a more reliable configuration. Here, we utilize pre - handling to standardize and normalize the information.

1) Identifying straight out Features: List of clear cut components has been recognized in preparing and testing datasets with number of classes as shown in result section.
2) One Hot Encoding: Used Machine learning calculations can not deal with downright elements. Along these lines, all unmitigated components are being changed over to twofold vectors for preparing and testing reason. In the first place, straight out esteem is planned to a whole number worth, and afterward, every whole number is addressed in twofold vector which has each of the 0 qualities aside from file of number which is stamped 1.
3) Adding missing classification in testing information: Six classes have been found missing in help highlight in testing information. These classifications are loaded down with 0's.
4) Splitting Dataset: After performing one - hot encoding, diverse attacks types found in preparing and testing dataset are planned tasks, Probe, R2L, U2R. Information is parted into 4 informational collections dependent on attacks types (DoS, Probe, R2L, U2R) to prepare the model for a wide range of attacks and anticipate results for these attacks.

c) Feature Selection

Provisions are critical in AI since they are the main quantifiable properties of the peculiarity being noticed. Picking instructive and autonomous elements is a urgent advance. Element choice or trait determination is a course of choosing a subset of instructive provisions from the whole set. Component choice ways are utilized to find impacted factors and eliminate pointless, unnecessary traits from information which don't influence the exactness of prescient models, in case they are incorporated or not, or may truth be told decline the precision of the model. Element choice and Feature extraction are unique. The vital contrast between include choice and extraction is that highlight determination attempts to discover the best subset of provisions among unique components while highlight extraction makes a bunch of new elements.

1) Filter Method: Filter technique applies static measures to ascertain score for each component. An element is either chosen or disposed of from dataset dependent on the score of each element. For instance, data gain, Chi squared test and relationship coefficient score.
2) Wrapper Method: Wrapper techniques are like an inquiry issue, where components are ready in various mix, assessed and contrasted with different blends. A prescient model is utilized which relegates a score dependent on model exactness for assessment of

provisions. Looking can be stochastic, for example, irregular slope climbing calculation, or heuristics, as forward and in reverse passed to add and eliminate highlights. For example, a recursive component disposal calculation.

3) Embedded Method: Embedded technique checks each provisions that builds exactness of model while model is being made.

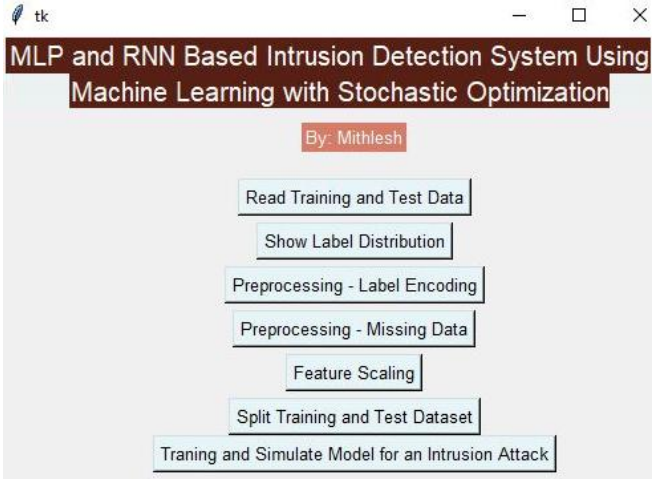
d) Train and Simulate Attack

After feature scaling the most important step is to train the model. We have utilized MLP and RNN based training with Stochastic optimization approach. This approach can guaranteed produced efficient algorithm outputs.

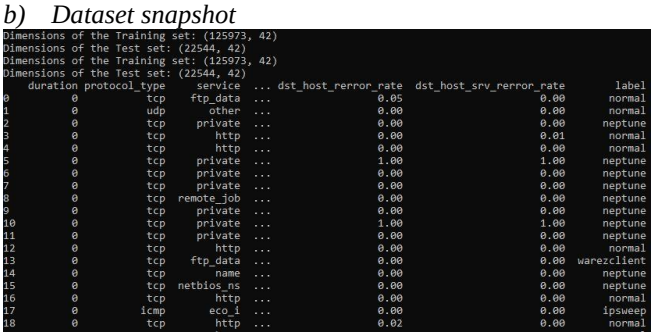
4. Result

In this section various stages output are presented.

a) Framework Screenshot

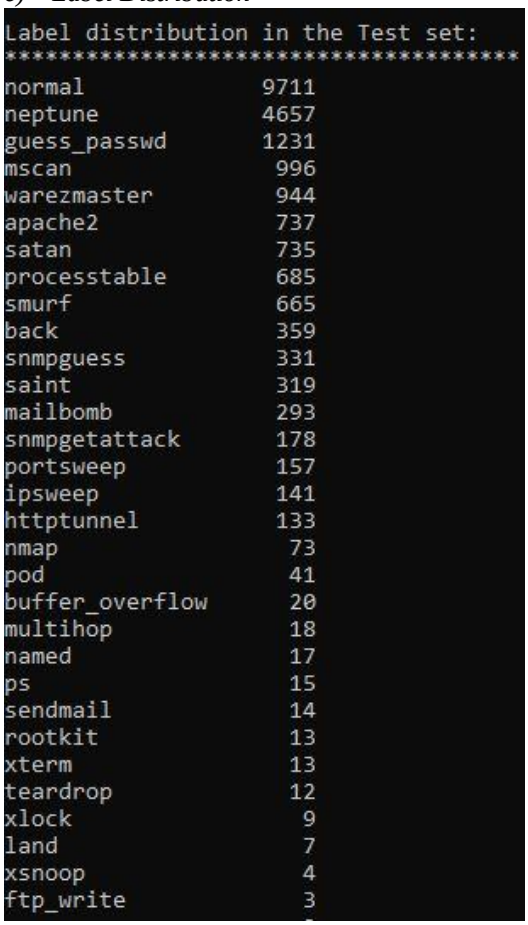


b) Dataset snapshot



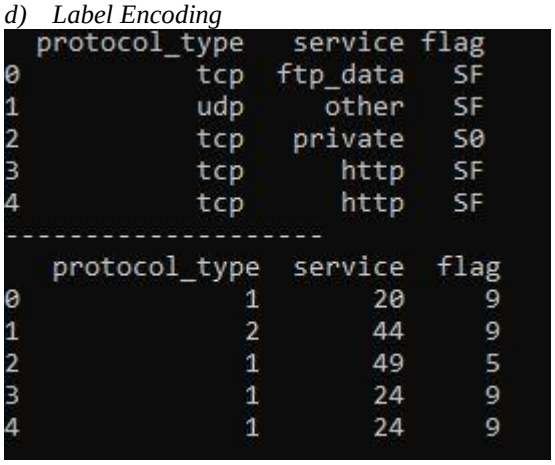
```
Dimensions of the Training set: (125973, 42)
Dimensions of the Test set: (22544, 42)
Dimensions of the Training set: (125973, 42)
Dimensions of the Test set: (22544, 42)
    duration protocol_type     service  ... dst_host_rerror_rate  dst_host_srv_rerror_rate        label
0          0           tcp    ftp_data  ...                0.05                      0.00       normal
1          0           udp       other  ...                0.00                      0.00       normal
2          0           tcp     private  ...                0.00                      0.00      neptune
3          0           tcp        http  ...                0.00                      0.01       normal
4          0           tcp        http  ...                0.00                      0.00       normal
5          0           tcp     private  ...                1.00                      1.00      neptune
6          0           tcp     private  ...                0.00                      0.00      neptune
7          0           tcp     private  ...                0.00                      0.00      neptune
8          0           tcp  remote_job  ...                0.00                      0.00      neptune
9          0           tcp     private  ...                0.00                      0.00      neptune
10         0           tcp     private  ...                1.00                      1.00      neptune
11         0           tcp     private  ...                0.00                      0.00      neptune
12         0           tcp        http  ...                0.00                      0.00       normal
13         0           tcp    ftp_data  ...                0.00                      0.00  warezclient
14         0           tcp        name  ...                0.00                      0.00      neptune
15         0           tcp  netbios_ns  ...                0.00                      0.00      neptune
16         0           tcp        http  ...                0.00                      0.00       normal
17         0          icmp       eco_i  ...                0.00                      0.00      ipsweep
18         0           tcp        http  ...                0.02                      0.00       normal
```

c) Label Distribution



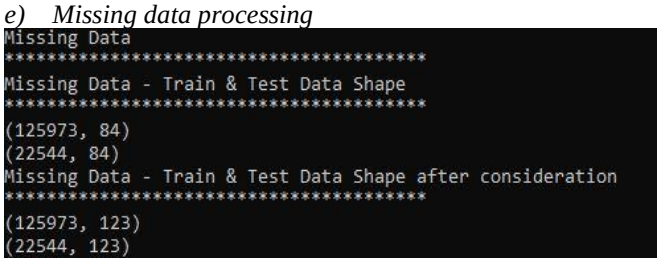
```
Label distribution in the Test set:
****************************************
normal              9711
neptune             4657
guess_passwd        1231
mscan                996
warezmaster          944
apache2              737
satan                735
processtable         685
smurf                665
back                 359
snmpguess            331
saint                319
mailbomb             293
snmpgetattack        178
portsweep            157
ipsweep              141
httptunnel           133
nmap                  73
pod                   41
buffer_overflow       20
multihop              18
named                 17
ps                    15
sendmail              14
rootkit               13
xterm                 13
teardrop              12
xlock                  9
land                   7
xsnoop                 4
ftp_write              3
```

d) Label Encoding



```
  protocol_type   service flag
0           tcp  ftp_data   SF
1           udp     other   SF
2           tcp   private   S0
3           tcp      http   SF
4           tcp      http   SF
--------------------
   protocol_type  service  flag
0              1       20     9
1              2       44     9
2              1       49     5
3              1       24     9
4              1       24     9
```

e) Missing data processing



```
Missing Data
****************************************
Missing Data - Train & Test Data Shape
****************************************
(125973, 84)
(22544, 84)
Missing Data - Train & Test Data Shape after consideration
****************************************
(125973, 123)
(22544, 123)
```

f) Traing and Test Split

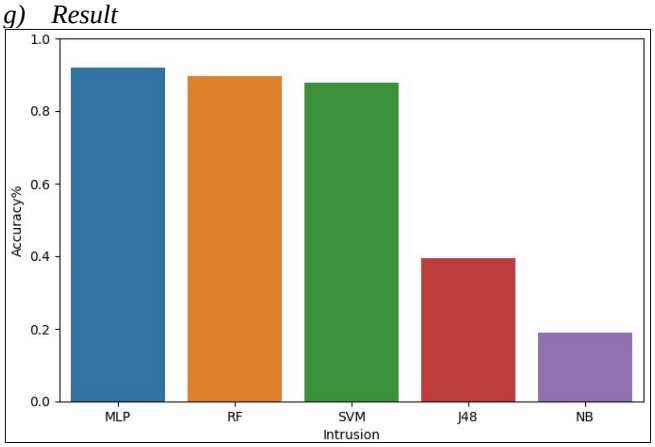
Test Split Data

	duration	src_bytes	dst_bytes	land	wrong_fragment	urgent	h
0	0	0	0	0	0	0	
1	0	0	0	0	0	0	
2	2	12983	0	0	0	0	
3	0	20	0	0	0	0	
4	1	0	15	0	0	0	
5	0	267	14515	0	0	0	
6	0	1022	387	0	0	0	
7	0	129	174	0	0	0	
8	0	327	467	0	0	0	
9	0	26	157	0	0	0	

Train Split Data

	duration	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot
0	0	491	0	0	0	0	0
1	0	146	0	0	0	0	0
2	0	0	0	0	0	0	0
3	0	232	8153	0	0	0	0
4	0	199	420	0	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0
12	0	287	2251	0	0	0	0
13	0	334	0	0	0	0	0
14	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0
16	0	300	13788	0	0	0	0

g) Result



The MLP based approach combined with stochastic optimization performs well as compared to all other ML approaches. The accuracy obtained is 91.8%.

5. Conclusion

In this paper, we presented our proposed model on intrusion detection that worked on machine learning algorithms, namely, MLP and RNN. We have conducted an extensive experimentation on the machine learning algorithms using features, and it is observed to be time consuming and performance degrading hence relevant features are selected. As per results obtained MLP with stochastic optimization attains 91.8% accuracy.

References

[1] C. H. Low, "NSL - KDD dataset, " Retrieved from https: //github. com/defcom17/NSL KDD.
[2] S. Zaman and F. Karray, "Features selection for intrusion detection systems based on support vector machines," in 2009 6th IEEE Consumer Communications and Networking Conference, Jan 2009, pp.1–8.
[3] J. Jha and L. Ragha, "Intrusion detection system using support vector machine," International Journal of Applied Information Systems (IJAIS), 2013.
[4] Revathi and Malathi, "A detailed analysis on nsl - kdd dataset using various machine learning techniques for intrusion detection," International Journal of Engineering Research and Technology, 2013.
[5] M. Almseidin, M. Alzubi, S. Kovacs, and M. Alkasassbeh, "Evaluation of machine learning algorithms for intrusion detection system, " in 2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY), Sept 2017, pp.000 277–000 282.
[6] U. S. K. P. M. Thanthrige, J. Samarabandu, and X. Wang, "Machine learning techniques for intrusion detection on public dataset, " in 2016 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), May 2016, pp.1–4.
[7] H. M. Anwer, M. Farouk, and A. Abdel - Hamid, "A framework for efficient network anomaly intrusion detection with features selection, " in 2018 9th International Conference on Information and Communication Systems (ICICS), April 2018, pp.157–162.
[8] M. J. Fadaeieslam, B. Minaei - Bidgoli, M. Fathy, and M. Soryani, "Comparison of two feature selection methods in intrusion detection systems, " in Computer and Information Technology, 2007. CIT 2007.7th IEEE International Conference on. IEEE, 2007, pp.83–86.
[9] C. Yun and J. Yang, "Experimental comparison of feature subset selection methods, " in Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on. IEEE, 2007, pp.367–372.