IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Low-High Split Clustering for Anomaly based Detection of Network Attacks

[1]Avinash R. Sonule, [2]Mukesh Kalla, [3]Amit Jain, [4]D S Chouhan,

[1]Research Scholar, [2]Major Advisor, [3]Co-Advisor, [4]Co-Advisor

[1]Department of Computer Science and Engineering,

[1] Sir Padampat Singhania University (SPSU), Udaipur-313601, Rajasthan, India

Abstract: The network attacks detection become the prime security problems in the day today life. As the use of computing resources is increased, cyberpunks are planning new tactics of network attacks. Many techniques have been invented to detect these attacks which are based on data mining and machine learning approaches. Many clustering methods have been used to detect network intrusions. These intrusions detection methods have been applied on various IDS datasets. UNSW-NB15 is the newest dataset which contains different modern attack types and normal activities. In this paper, we have proposed unsupervised machine learning algorithms for anomaly based detection on reduced UNSW NB15 dataset.

Index Terms **- UNSW NB-15, Machine Learning, Low-High Distance, Low-High Split Clustering Algorithm.**

I. INTRODUCTION

The use of computing devices have increased in almost all fields to solve problems of societies. Most of these computing devices are connected to the Internet. This enormous demand for connectivity has challenged the traditional network architectures. These computing devices can be accessed using a number of ways and this becomes a threat to the network. Our system can predict attacks even before they happen in order to warn the users before they cause any harm. Intrusion Detection Systems (IDS)(Stefan A.,2000) is a device or software application that monitors network and the system for suspicious activities and warns the system or network administrator. There are Host based IDS and Network based IDS. A Host based Intrusion Detection System keeps track of individual host machine and gives notice to the user if suspicious activities are found. The Network based Intrusion Detection System (NIDS)(Anwer et al, 2018) is kept at a gateway or routers to detect the intrusions over the network. A NIDS keep track of network-attack patterns and protect computing resources. IDS can be categorized by the detection mechanism used by it. These IDSes are: i) misuse detection, ii) anomaly detection and iii) hybrid detection. Misuse detection techniques have been used to detect known attacks while the Anomaly detection techniques have been used to detect unknown attacks. Machine Learning (ML) can be used for all the three types of detection techniques. A machine learning models have two parts: training and testing. By using training data samples as a input, learning algorithm learn the features in the training. In the testing, the learning algorithm predicts the unknown data.

Machine learning algorithms are tested on different network attack datasets with or without feature selection approaches. Unsupervised learning algorithms take a set of data that contains only inputs, and find pattern in the data, such as grouping or clustering of data points. The algorithms therefore learn from test data that has not been labelled, classified or categorized. Instead of responding to feedback, unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data.

The rest of the paper is organized as follows: section II gives related work for clustering based algorithms applied on UNSW NB 15. Section III gives in detail of the Cluster Analysis and UNSW NB 15 dataset(Moustafa et al, 2015) (Moustafa et al, 2016). In section IV, proposed methods with Low High Split(LHS) Algorithm is given. Section V gives results and discussion. Final section gives future direction and concludes the work.

II. RELATED WORK

Many researchers have used different Machine Learning algorithms on different Datasets. Priya et al.(Priya et al, 2018) have done survey on different machine learning algorithms applied on various datasets The different machine learning algorithms have applied on UNSW NB 15 dataset. The following some researchers have used the following algorithms on UNSW NB 15 dataset.

Nahiyan et al.(Nahiyan et al, 2017) proposed an automated, agent-based, unsupervised, relatively less complicated cognitive approach. The proposed algorithm collects features from statistical analysis of the observed attributes over each time-step and uses machine learning to isolate the attack events from normal attack using an unsupervised kmeans clustering algorithm over the reduced dataset. The agent based architecture is used to optimize the computational load for central processing so that the agent based architecture deploys agents in hosts, and some processing is done at the host and the rest is performed by the node that performs

the classification. They achieved total recall, precision and f1 -score 92%, 91% and 91% respectively for time 8 seconds using UNSW-NB15 dataset.

Moustafa et al. (Moustafa et al, 2018) proposed threat intelligence scheme which models the dynamic interactions of industry 4.0 component including physical and network systems. Industry4.0 includes the integration of Cyber-Physical systems (CPS), Internet of Things (IoT), Cloud and Fog computing paradigms for developing smart systems, smart homes, and smart cities. The smart data management module handles heterogeneous data sources. This includes data to and from sensors, actuators, in addition to other forms of network traffic. The proposed threat intelligence technique is designed based on Beta Mixture-Hidden Markov Models (MHMM) for discovering anomalous activities against both physical and network systems. The scheme is evaluated on the UNSW-NB15 dataset of network traffic. The results shows that the proposed technique outperforms five peer mechanisms: Cart, KNN, SVM, RF and OGM. Using the UNSW-NB15 dataset, the proposed MHMM mechanism gives95.89% DR, 96.32% accuracy and 3.82% FPR which is better than others.

Tian et al. (Tian et al, 2018) developed a methodology for anomaly detection by introducing Ramp loss function to the original One-class SVM, called "Ramp-OCSVM". The Concave– Convex Procedure (CCCP) is utilized to solve the obtained model that is a non-differentiable non-convex optimization problem. They used comprehensive experiments and parameters sensitivity analysis on UNSW-NB15 data sets. Ramp-OCSVM outperforms the OC-SVM, ROCSVM and eta OCSVM on UNSW-NB15 data sets. Using RampOCSVM, they achieved values of 97.24%, 93.07% and 2.25% for the total accuracy, detection rate, false alarm rate respectively.

Moustafa et al. (Moustafa et al, 2018) [9] proposed an architectural scheme for designing a threat intelligence technique for web attacks through a step methodology: First by collecting web attack data by crawling websites and accumulating network traffic for representing this data as feature vectors; second by dynamically extracting important features using the Association Rule Mining (ARM) algorithm; third by using these extracted features to simulate web attack data; and last by using a new Outlier Gaussian Mixture (OGM) technique for detecting known as well as zero-day attacks based on the anomaly detection methodology. The OGM technique compared with four competing techniques, namely Cart, KNN, SVM and RF .The Receiver Operating Characteristics (ROC) curves signify the relationship between the DR and FAR in order to effectively show the potential process of running these techniques using the original data in the UNSW-NB15 dataset. Empirical results show that the OGM outperforms others, producing a 95.68% DR and 4.32% FAR, while the others achieve in an average of 89%- 93% DR and 6.4%-10.5% FAR.

Muhammad et al. (Muhammad et al, 2021) proposed multi-class classification and performed experiments on reduced data-set with full features and cluster-based features and used all imputation techniques used in binary classification. The three ML algorithms are used for evaluation. Multi-class classification results using reduced data-set. They achieved highest accuracy of 97.37% by applying RF on regression imputed data-set. With SVM, they achieved 95.67% accuracy, and with ANN 91.67% accuracy.

III. CLUSTER ANALYSIS AND UNSW NB-15 DATASET

Cluster Analysis is a procedure of assembling the objects in Clusters whose member exhibits similar features. Cluster is a group of data objects that are similar to each other and are different from the objects belonging to other clusters. In cluster analysis, the set of data object partition into several different groups based on data similarity and then the labels are assigned to each and every group. It is a vital task in the detection of network attacks.

Cluster analysis is repetitive process of acquiring knowledge which includes trial and errors. Often it is required to edit the data pre-processing and configuration variables to get desired results.. The major advantage of clustering includes, it is flexible to the changes and helps pick out useful features that differentiate different groups. The Clustering methods can be classified into different categories. These Clustering methods can be classified into the following categories: Partitioning Method, Hierarchical Clustering, Centroid based Clustering, Density-based Method, Distribution-based Clustering, Grid-based Method and Constraint-based Methods etc.

The some of the drawbacks of Clustering Algorithms include 1) Need to start with random centroid/medoids.2) Wrong selection of centroid leads to more iterations or wrong clusters. 3) Large number of features and data objects increases the time complexity. High frequency of shifting of data objects from one cluster to another degrades the performance. 4} the efficiency of algorithm depends on the definition of distance in distance-based clustering. 5) Difficulty in defining the distance measure in multi-dimensional spaces when it doesn't exist. 6} The clustering algorithm results in many cases may be arbitrary in itself and can be read in many ways.

UNSW-NB15 Dataset: The DARPA98, KDD, NSL-KDD, ISCX-2012 and other datasets do not represent the modern network traffic with different attack scenarios. The cyber security research group at the Australian Centre for Cyber Security (ACCS) have developed Network Intrusion Dataset, UNSW-NB15 dataset. The raw network packets of the UNSW-NB15 dataset [6] was created by the IXIA PerfectStorm tool in the Cyber Range Lab of ACCS for generating a hybrid of real modern normal activities and synthetic contemporary attack behaviors. Twelve algorithms were developed using a C# language to analyze in-depth the flows of the connection packets. The Argus, Bro-IDS tools are used with twelve algorithms to generate total 49 features with the 9 class label. For this paper we have used reduced (few tuples) UNSW-NB15 dataset with few attributes. The column having discrete values are used. We can use different feature selection methods to get reduced dataset.

IV. PROPOSED METHOD

The proposed framework for Network attacks detection is shown in Figure 1. We can use different pre -processing and feature selection methods on UNSW NB 15 Network Intrusion Dataset. The reduced UNSW NB 15 dataset as shown in Table 1 with 2 attributes/features, Mean of the flow packet size transmitted by the source (smeansz), Mean of the flow packet size transmitted by the destination (dmeansz)(Content Feature).

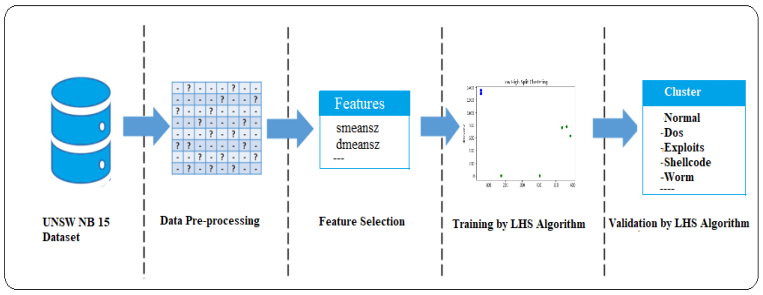

Fig. 1 : Proposed framework for network attacks detection

Preparation of Training/Testing Dataset: **Twin sample validation** can be used to validate results of Low-High Split unsupervised learning. Twin sample can be created by inter changing the features/attributes order of data objects/tuples.

Twin-Sample Validation

1. Creating a twin-sample of training data(one sample for training and other for testing).
2. Applying unsupervised learning on twin-sample.
3. Applying unsupervised learning on twin-sample
4. Getting results for twin-sample of datasets.
5. Calculating similarity between two sets of results.

Low High Split(LHS) Algorithm

1] Find the Smallest/Lowest and Largest/Highest values from the dataset/cluster.
2] Form the new clusters by assigning other data objects to Low or High object by their closest distance using distance formulae.
3] Find the Lowest and Highest values from the new clusters.
4] Find the difference between Lowest and Highest of respective clusters.
5] Split the cluster having higher absolute Low- High difference. If differences are same split the cluster having less number of data object. If same number of objects, split any one cluster.
6] Repeat Step 1 to 5 till we get required number of clusters or till single object in cluster (Low-High difference becomes zero).

Note: Lowest value of the tuple/data object is the smallest value in first column and then in second column and so on till last column. Highest value of the tuple is the highest value in first column and then in second column and so on till last column. We can calculate Low and High values by the same way by interchanging the columns(features).

$$\text{Euclidian Distance(ds) Formula} = \text{SQRT}((X_1 - X_2)^2 + (Y_1 - Y_2)^2 + \text{----+} (Z_1 - Z_2)^2) \quad (1)$$

Where $X_1 - X_2$ is difference in first feature of two different data objects.
$Y_1 - Y_2$ is difference in second feature of two different data objects.
$Z_1 - Z_2$ is difference in last feature of two different data objects.

V. RESULTS AND DISCUSSION

We have applied the proposed algorithm on the reduced UNSW NB 15 dataset with two features as shown in Table 1. In unsupervised machine algorithms/clustering algorithms class label are not required.

Table 1: Reduced Dataset with two features

Tuple No/ Data Objects	Mean of the flow packet size transmitted by the source (smeansz),	Mean of the flow packet size transmitted by the dst (dmeansz)
A	537	760
B	54	1352
C	564	774
D	54	1298
E	175	0
F	404	0
G	587	629
H	55	1308

We have to calculate distance between Lowest Object and all other objects of cluster except Highest Object. Similarly we have to calculate the distance between Highest Object and all other object of the cluster except Lowest Object. The distance between Lowest Object and Highest Object need not required to calculate. The distance between each and every object also need not required to calculate which the advantage over other algorithms.

Lowest value/Object D = (54, 1298) and Highest values/Object G = (587, 629).

The Euclidean distance(ds) between the other tuples/ data object from Low and High Object is shown in Table 2.

The Data objects belong to the following clusters by considering their closest distance.

Cluster C1= {D, B, H} Cluster C2= {G, A, C, E, F}

Lowest and Highest Objects from the new clusters:

C1 {Low D= (54, 1298) & High H = (55, 1308)}, C2 {Low E = (175, 0) & High G= (587, 629) }

Table 2 : Distance of objects from Low and High Object

Data Objects	Distance from Lowest Data Objects D to	Cluster
A	DA = 723	
B	DB = 54	C1
C	DC = 731.21	
E	DE = 1303.62	
F	DF = 1344.36	
G*	DG = 855.36	
H	DH = 10.04	C1
	Distance from Highest Data Objects G to	
A	GA= 140.21	C2
B	GB= 898.23	
C	GC= 117.27	C2
D*	GD= 855.36	
E	GE= 751.92	C2
F	GF= 655.08	C2
H	GH= 862.59	

The difference between Low and High of cluster C1 and C2:

C1= ds(Low=D, High=H)=10.04 C2= ds(Low=E, High=G)= 751.92

Split the cluster C2 which have higher Low- High difference. The distances from Low and High data objects of cluster C2 is shown in Table 3.

Table 3: Distances from Low and High objects of cluster C2

Data Objects	Distance from Lowest Data Objects E to other objects of cluster	Cluster
A	EA= 842.99	
C	EC= 866.25	
F	EF= 229	C3
	Distance from Highest Data Objects G to other objects of cluster	
A	GA= 140.21	C4
C	GC= 117.27	C4
F	GF= 655.08	

The Data objects of Cluster C2 now form the following new clusters by considering their closest distance.

Cluster C3= {E, F}, Cluster C4= {G, A, C}

Lowest and Highest Objects from the new clusters:

C3 {Low E= (175, 0) & High F = (404, 0)}, C4 {Low A = (537, 760) & High G= (587, 629) }

The difference between Low and High of respective clusters:

C1= ds(Low=D, High=H)=10.04 C3= ds(Low=E, High=F)= 229, C4= ds(Low=A, High=G)= 140.21

Split the cluster C3 which have higher Low- High difference. Find the distances from Low and High objects.

The Data objects of Cluster C3 now form the following new clusters with single data object.

Cluster C5 = {E} Cluster C6 = {F}

The difference between Low and High of respective clusters:

C1= ds(Low=D, High=H)=10.04, C4= ds(Low=A, High=G)= 140.21

C5 and C6 are with now single Data object.

Split the cluster C4 which have higher Low- High difference. The distances from Low and High objects for cluster C4 is shown in Table 4.

Table 4: Distances from Low and High objects for cluster C4

Data Objects	Distance from Lowest Data Objects A to other objects of cluster	Cluster
C	AC = 45.35	C7
	Distance from Highest Data Objects G to other objects of cluster	
C	AC = 117.27	

The Data objects of Cluster C4 now form the following new clusters by considering their closest distance.

Cluster C7= {A, C} Cluster C8= {G}

We can repeat these clusters formation till we get required number of clusters or there is a single object in each cluster. Initially we get two clusters C1with normal flow data objects and C2 with Attacks data objects. In second t step we get clusters C3 having Low frequency attacks data objects and C4 with High frequency data objects. In third step we get clusters C5 with Shellcode data object and C6 with worm data object. After fourth step we get clusters C7 with DoS attacks and C8 with Exploits.

For the above Dataset, if we check class label Attack Type from UNSW NB 15 Dataset we found that network dataset initially categorized into Normal flow and Attack. The attacks are divided into high frequency attack and low frequency attacks. Finally high frequency attacks are clustered into DoS and Exploits etc. The low frequency attacks are clustered into Shellcode and Worm. The given network dataset is grouped into 5 clusters: Normal Flow, Shellcode, Worm, DoS, Exploits as shown in Figure 2. The formation of clusters can be in any sequence but final clusters will be the same. Figure 3 shows plotting of data points and clusters.

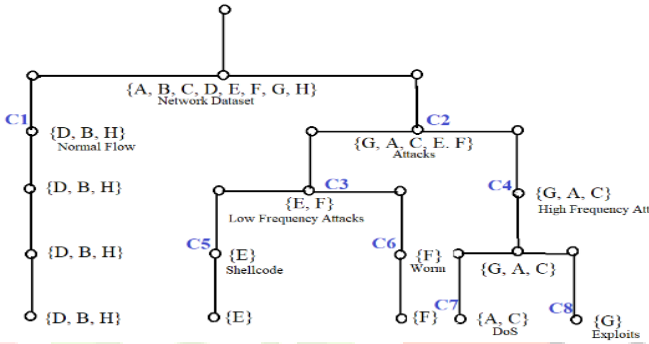

Fig 2 : Clusters Formation

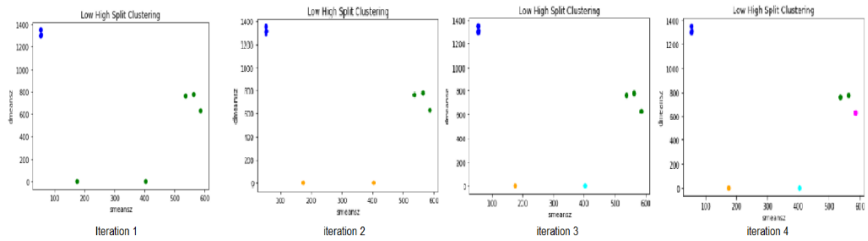

Fig 3 : Plotting of data points and clusters

Twin-Sample Validation : After interchanging the features we get one of Twin Sample as shown in Table 5. The starting with the Lowest Data Object is F (0, 175) and Highest Data Object is B (1352, 54) The formation of clusters during validation is shown in Figure 4. We get the same final clusters after iteration 4 of proposed algorithm.

For the twin samples if we want less number of clusters we may not get the same clusters. For more number of clusters we get the same results. If we want 3 or more clusters then we get the same clusters. For more number of clusters we are getting very good accuracy.

Table 5: Twin sample for validation

Tuple No/ Data Objects	Mean of the flow packet size transmitted by the dst (dmeansz)	Mean of the flow packet size transmitted by the source (smeansz)
A	760	537
B	1352	54
C	744	564
E	1298	54
F	0	175
G	0	404
H	629	587
I	1308	55

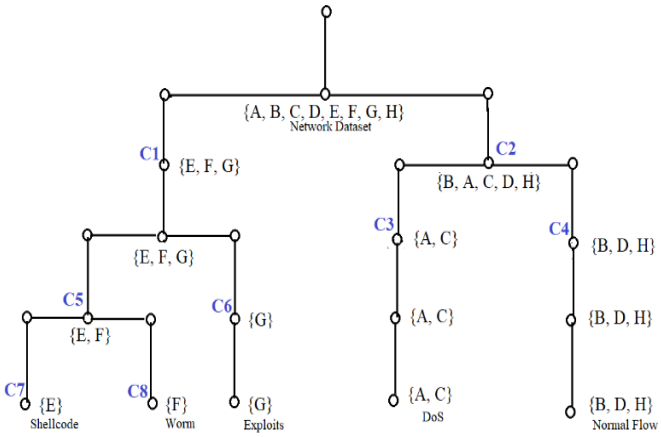

Fig 4 : Clusters formation for Validation Sample.

The confusion matrix for above Twin samples is shown in Table 6. The proposed algorithm is giving good 100 % accuracy after iteration 4.

Table 6: Confusion matrix

7	Cluster 1: Normal Predicted	Cluster 2: Shellcode Predicted	Cluster 3: Worm Predicted	Cluster 4: DoS Predicted	Cluster 5: Exploits Predicted	
Cluster 1: Normal Actual	3	0	0	0	0	3
Cluster 2: Shellcode Actual	0	1	0	0	0	1
Cluster 3: Worm Actual	0	0	1	0	0	1
Cluster 4: DoS Actual	0	0	0	2	0	2
Cluster 5: Exploits Actual	0	0	0	0	1	1
	3	1	1	2	1	7

The reduced UNSW NB 15 dataset with three attributes/features, Mean of the flow packet size transmitted by the source (smeansz), Mean of the flow packet size transmitted by the dst (dmeansz)(Content Feature), is_sm_ips_ports, If source equals to destination IP addresses and port numbers are equal, this variable takes value 1 else 0. (Additional Generated Features), is as shown in Table 7.

Table 7 : Reduced dataset with three features

Tuple No/ Data Objects	Mean of the flow packet size transmitted by the source (smeansz),	Mean of the flow packet size transmitted by the dst (dmeansz)	is_sm_ips_ports
A	537	760	0
B	54	1352	1
C	564	774	0
D	54	1298	1
E	175	0	0
F	404	0	0
G	587	629	0
H	55	1308	1

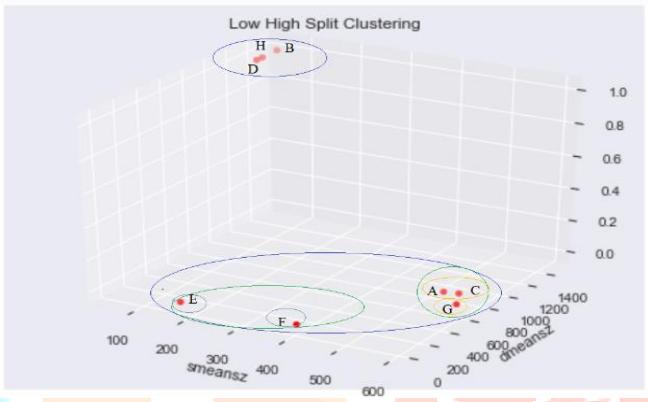

Fig 5: Cluster formation with three features

The clusters for the dataset with three features are shown in Figure 5. The cluster formation with 3 features is in the same sequence as shown in Figure 3. For the validation sample with 3 features in the order(is_sm_ips_ports, smeansz, smeansz) we get the same Clusters {A, B ,C, D, E, F, G. H} →{{G, A, C, E, F},{D, B, H}}→{{E, F},{G, A, C},{D, B, H}}→{E}, {F},{G, A, C},{D, B, H}}→{{E}, {F},{ A, C},{G},{D, B, H}}as in training.

As we increase the number of features of dataset for training ang testing, clusters will more differentiate themselves from each other.

VI. Conclusion and Future Work

In this paper we have proposed Low - High Split(LHS) unsupervised algorithm. This algorithm is applied on reduced UNSW NB15 dataset. We have used Twin Samples for training and testing. We have used small dataset to concentrate more on algorithm. The proposed algorithm is giving good accuracy. This algorithm can be applied on any dataset of any size.

REFERENCES

[1] Stefan A.2000. Intrusion detection systems: A survey and taxonomy, Technical report, Vol. 99.

[2] Anwer H.M., Farouk M. and Abdel-Hamid A. 2018. A Framework for Efficient Network Anomaly Intrusion Detection with Features Selection," 9th International Conference on Information and Communication Systems (ICICS) 2018, IEEE, p.157.

[3] Moustafa N. and Slay J. 2015.Unsw-nb15: A comprehensive data set for network intrusion detection systems (unsw-nb15 network data set), in Military Communications and Information Systems Conference (MilCIS), Canberra, Australia, pp. 1–6.

[4] Moustaf N and Slay J. 2016.The evaluation of network anomaly detection systems: Statistical analysis of the unsw-nb15 data set and the comparison with the kdd99 data set, Information Security Journal: A Global Perspective, 25(6) 18-31.

[5] Mishra P., Varadharajan V., Tupakula U.and Pilli E.S. 2018. A Detailed Investigation and Analysis of using Machine Learning Techniques for Intrusion Detection, IEEE Communications Surveys & Tutorials.

[6] Nahiyan K., Kaiser S., Ferens K. and McLeod R. 2017. A Multi-agent Based Cognitive Approach to Unsupervised Feature Extraction and Classification for Network Intrusion Detection, Int'l Conf. on Advances on Applied Cognitive Computing| ACC'17. CSERA Press, page no. 25.

[7] Moustafa N.,Adi E.,Turnbull B. and Hu J.2018. A New Threat Intelligence Scheme for Safeguarding Industry 4.0 Systems, IEEE Access Open Access Journal, vol 4 , page no. 1.

[8] Tian Y. , Mirzabagheri M., Mojtaba S., Bamakan H., Wang H. and Qu Q.2018. Ramp loss one-class support vector machine; A robust and effective approach to anomaly detection problems, Journal neorocomputing , Elsevier.

[9] Moustafa N., Misra G. and Slay J.2018. Generalized Outlier Gaussian Mixture technique based on Automated Association Features for Simulating and Detecting Web Application Attacks, Journal of IEEE Transactionson Sustainable Computing, IEEE.

[10] Muhammad A., Qaiser R., Muhammad Z.,, Hasan T., Syed A. and Muhammad S. 2021. Intrusion detection in internet of things using supervised machine learning based on application and transport layer features using UNSW-NB15 data-set, EURASIP Journal on Wireless Communications and Networking, Springer Open, p 1-23.