Novel Framework for Anomaly Detection Using Machine Learning Technique on CIC-IDS2017 Dataset

Richa Singh
AIT CSE of Chandigarh University
Mohali, India
richa.e10605@cumail.in

Gaurav Srivastav
AIT CSE of Chandigarh University
Mohali, India
gaurav.e9988@cumail.in

2021 International Conference on Technological Advancements and Innovations (ICTAI) | 978-1-6654-2087-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICTAI53825.2021.9673238

***Abstract*— There are various deep learning-based IDS techniques are implemented in large scale. Intrusion detection systems are critical components for protecting ICT infrastructure (IDSs). Keeping this in mind, solid solution is required for different types of new attacks and complexity control. Deep learning and machine learning is widely used to handle high dimensional, complex type data. The IDS detects and attracts various attack types such as known, unknown, and zero-day attacks using unsupervised machine learning techniques. To detect threats without prior knowledge, a framework has been designed that uses the concept of One Class SVM (OCSVM) and active learning. The CIC-IDS2017 dataset was used to test the performance of the framework and compare the result with UNSW-NB15 and KDD cup 99 dataset. The final output shows that this framework gives better performance than other.**

Keywords—Machine learning, Intrusion Detection Systems, UAI layer

I. INTRODUCTION

Nowadays people are totally dependent on internet and as a result, it is used in vast areas and threats have rapidly enhanced. IDS are used in devices to prevent these attacks. IDS are becoming more important in preventing network attacks; however, IDS based on payload have a scalability problem due to the high speed and traffic of today's networks. In Flow-based IDS, Deep packet inspection mechanisms has more priority as compared to traditional IDS for: firstly, small amount of data is processed and secondly: Easily gets the data coming from those forwarding devices that use standard protocols.

Deep learning is a novel technique that is used in different sectors. It can be used in networking also. Deep learning consists of collection of algorithms through which progress can be achieved in any field based on knowledge. It has a great potential in field of networking and give major contribution in today's era.

In comparison of supervised and unsupervised learning, supervised learning used labelled data, train it and on the basis of feature vector, test data is categorized as anomaly as well as in unsupervised technique, unlabeled data are used for learning. There are different methods of unsupervised and clustering [1] technique is one of them, which search the dataset for similar instances to create cluster. Those instances which have similar characteristic have placed in same kind of cluster. Instances with same behaviors are treated in same kind of cluster. The semi-supervised method contains both the data, labelled as well as unlabeled [2]. This method learns feature label associations from labelled data and on the basis of this assign's labels to unlabeled data.

The structure of the paper is as follows. Section II discussed the survey regarding anomaly detection. Section III discussed the proposed solution; Section IV gives the result analysis after evaluation. Finally, Section V gives conclusion of the methodology presented in this paper.

II. RELATED WORK

Mostly, detection of anomaly models is categorized into three types: machine learning models, general probability (statistical) models and neural network models. Among all the three models, neural network model is well suited for network traffic characteristics and it is very different from the conventional methods, which shows only the shallow features [3].

Anomaly detection is a major concern nowadays. Many researchers give their contribution in this field. They mainly focus on the active anomaly which is presented in network any many other application areas. Active anomalies are detected by applying the active learning to the distribution dataset [4]. In [5] Using active learning techniques, outliers are reduced using artificially generated instances.

Anomaly is detected by supervised as well as unsupervised techniques. This paper used the unsupervised technique and detect active anomaly using SVM [6] [7]. In [8], this paper has the hybrid concept of both supervised and unsupervised techniques in which there are labelled data instances through which the detection is possible. The required anomalies can be identified using active learning models and on the basis of expert's recommendations new features are designed to improve the functionalities of the model. [9] In this paper algorithms are proposed using ensemble techniques.[10] shows how isolation forest technique gives their contribution in an active environment.

Support vector machine is a supervised technique, which is used for classification, regression and outlier detection. It is effective in high dimensional spaces. It uses a technique known as kernel to transform data and on the basis of this transformation, it's very easy to find the optimal boundary. SVM is one of the techniques which gives approximately correct result in case of outlier detection. So, in most of the research, SVM techniques are used. [11] uses SVM and gives better accuracy and low false positive rate. OSCVM methods is used in an intruder network to propose IDS [12], and analyze the functions on network flow. In [13], large volume of NetFlow data records is processed using SVM and serve as input to the kernel function and then find the result using OCSVM.

CNN are applied on kernel level and try to find out that which device is compromised and which device is not compromised and on the basis of calculate the accuracy and f score and then detect the anomaly [14]. [15][16] This paper discussed about the distributed environment, different algorithms its pros and cons.

III. PROPOSED METHODOLOGY

In this paper, unsupervised anomaly detection method is used in which SVM and isolation based active learning techniques are used for anomaly detection based on clusters. In this methodology, the groups of data are separated on the basis of high and low clustered data. Similar data is in one cluster and all the dissimilar data is in the one cluster and this can be done with the help of clustering methods that is DBSCAN and K mean clustering.

K mean clustering algorithm is much more efficient for large as well as medium dataset. It is used to minimize intra-cluster distance and maximizing inter-cluster data. The limitations of the K-mean cluster are that the clusters are predefined. But the value of k is not fixed and mostly based on the structure and scale of the different points in the dataset.

The Density based spatial clustering of applications with noise i.e., DBSCAN algorithm, is a base algorithm for density-based clustering. Density means the radius specified within the specific points. It is very effective for noisy dataset or spatial clusters. The different parameters and the radius of DBSCAN are same. It is better as compared to k mean clustering algorithm because it is not required the predetermined clusters.

A. Support Vector Machine based Isolation forest method

The paper proposes an intrusion detection algorithm, Isolation forest SVM (OCSVM), to detect unsupervised anomalies [17, 18, 19]. Sub space clustering is similar to isolation forest clustering (SSC). SSC is a traditional clustering technique. The clusters are derived from various sub clusters of the dataset. It is generated from small subspaces of various original datasets A Rabc, where a stand for number of records, b is number of attributes. The N sub clusters is equivalent to m and is generated by b.

If the value b is smaller, then it is more efficient and faster [17]. DBSCAN [20] algorithm works on high dimensional data and gives the improved result on low dimensional data. SVM is a supervised technique that analyses and recognizes patterns. OCSVM is an SVM extension method that is suitable for unlabeled data [19]. In OCSVM, SVM is trained with the help of data and is mapped to the feature space of the kernel that are used to separate the margin from its original [19]. The isolation forest OCSVM algorithm technique consists of the following steps:

- *Initialization*: Set X to a null vector and divide A into N different clusters Ai A(i 1,2..,N).
- *Learning and Clustering*: Each Ai subspace was used with OCSVM, and Pi partitions were created.
- *Gathering of evidence*: X dissimilarity vector should be updated based on each Pi partition. The distance between various outliers discovered in subspace Ai accumulated in Vector X (EA clustering) [21].
- Detection of anomalies: Rank and obtain the ranked Vector X. In case the dissimilar data value is more as compared to the fixed threshold point, then it is anomaly.

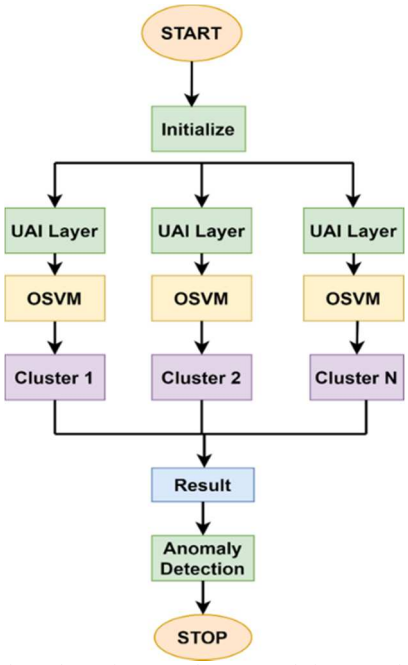

Figure 1. Flow Chart of Proposed Framework for anomaly Detection

B. Measure

For different attacks, how deep learning and machine learning techniques are efficient, is compared by various

metrics. There are few matrices and parameters used for evaluating the method are discussed below: -

- *Confusion matrix*

  This matrix is used to compare the actual values with the predicted values. It is used to predict the accuracy on the basis comparisons between the actual one's vs predicted ones. Table 1 shows the confusion matrix; Row represent by the actual values and column represent the predicted values. The values which is correctly classified as X is true positive and for incorrectly classified, represented by False positive. False negative belongs to the actual class X i.e. incorrectly classified as Y', and True Negative (TN) refers to the Y' that was correctly classified as Y'.

- *TPR*

  TPR is known as recall or detection rate or detection probability. It is calculated by true positive divided by true positive plus false negative as below.

  TPR=True positive/(True Positive+False Negative)

- FPR

  It is also known as false alarm rate and is calculated by the formula i.e. False positive divided by false positive plus true negative. With the help of FPR, easily identify the incorrect samples that are treated as anomaly.

  FPR=False Positive/ (false Positive + True Negative)

- There is a ROC curve which is used to represent the curve of TPR vs FPR on different criteria. [22]

TABLE I. CM OF ACTUAL VS PREDICTED CLASS

Actual Class	Predicted Class	
	X	X^1
X	True Positive	False Negative
X^1	False Positive	True Negative

C. Active-Learning Technique

In this technique, a model r (x) is trained to capture r full (x), i.e., the entire dataset distribution. The one way to find anomaly is by using the formula [23].

$$t(x)* 1/r(x)$$

It means the low value of r full(x) is anomaly. But this method has several disadvantages:

- If in case anomalies occurs at any interval of time than the expected time, then it means r full will gives poor result.
- If it is tightly forming a group, high probability region can be easily finding out using high-end techniques.

If the dataset is not balanced then in that case, active learning is used (0) [24],[25]. It has the ability to give better result from unlabeled data. [26-28]. The majority of unsupervised anomaly detection practical states show marginal the accuracy, instances are evaluated later by human professionals on the basis of its rank. There is a dataset X=x|x pfull (x), in which the uncompromised data are ranked and sent for auditing purpose. Rather than selecting and ranking all the instances are possible to iterate in a small cluster to professionals. In labelled instances, the number of anomalies may be enhanced. It can be iterated with an expert in the active learning technique.

Every step, in labelled instances, the dissimilar data are sent to the strong review, and training process is started after getting the feedback. In this strategy, most likely positive selection occurs at each step of the top selected k elements. This is one method for selecting meaningful instances among highly imbalanced datasets [29],[30], the current research of detecting anomaly is mentioned in this paper [31]-[33]. Considering all these, UAI i.e., Unsupervised Active Inference layer was introduced.

The technique used in this model is to add UAI layer on top of any unsupervised learning models that provides an uncompromised value for ranking purpose. To find out the anomaly score, layer latent representation (l(x)) act as input along with output anomaly score (s(x)) that is the output of the model and move it to the classifier. It is formalized as follows:

$$p'(y|x)\alpha\ suai(x)=classifier([l(x);\ s(x)]) \quad (1)$$

where p'(y|x) is an empirical estimate of the probability that x is an anomalous point. According to the research, there is a simple statistical structure [34], which simplifies modelling work and allows for the detection of unnatural points [35].

The model UAI layer in this work is denoted as:

$$p'(y|x)\ \alpha\ s_{uai}(x)= \sigma\ (Wact[l(x);s(x)] + b_{ac}t)$$

where Wact is linear transformation, bact is the bias, and (.) is activation function i.e. sigmoid function Back propagation with the cross entropy loss function is used to get the exact score of W and b, with the actively labelled instances serving as targets. Gradients can pass via l, s is not able to differentiate. UAI are network that is used make the UAI layer at the top.

IV. EVALUATION

The experiment is performed on CIC-IDS2017 dataset [36],[37]. The performance of various deep learning-based framework is analyzed and compared with other frameworks.

A. Benchmark Dataset

Finally, the result of anomaly detection network traffic is much closer to the benchmark dataset [38][39].

The CIC-IDS2017 dataset are too much efficient to solves the problem of the NSLKDD and KDDCup99 dataset. It has a large number of records in training as well as testing datasets. The information shown in table 2, clearly indicate that there are normal records but apart from those four abnormal records are also available there.

At regular interval of time, Data is collected during the network traffic which is made up of several data packets. These packets are a series of traffic bytes. Every data packet has 80 features, and belongs to one class. The representation is in this form A=(x0,xi,....xn), where A is data packets having continuous features and xi is the ith feature in the data packet The dataset contains basic features (numbered 1–10), content features (numbered 11–22), and traffic features[40]. There are four attack types in the dataset which is mentioned in the table 2 & 3 with respect to NSLKDD dataset & CIC-IDS2017 dataset.

TABLE II. CLASSIFICATION OF DATA IN KDDCUP99 DATASET

	Table Column Head					
	Total	Normal	Dos	Probe	R2L	U2L
CIC-IDS2017 Train	125973	68721	44835	10676	996	53
CIC-IDS2017 Test	22533	9812	7548	2134	2690	203

TABLE II. CLASSIFICATION OF DATA IN CIC-IDS2017 DATASET

	Table Column Head					
	Total	Normal	Dos	Probe	R2L	U2L
CIC-IDS2017 Train	125973	68721	44835	10676	996	53
CIC-IDS2017 Test	22533	9812	7548	2134	2690	203

The following graph represent the performance analysis of previous framework and the comparison result graph with other frameworks by using NSLKDD dataset.

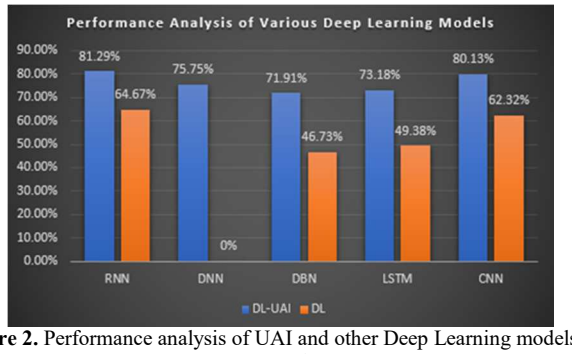

Figure 2. Performance analysis of UAI and other Deep Learning models on NSLKDD dataset

The following graph represent the performance analysis of UAI and other deep learning models on CIC-IDS2017 dataset.

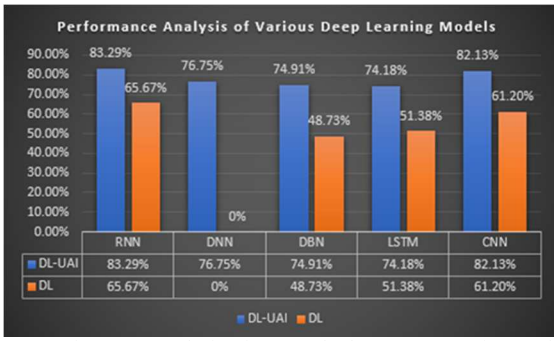

Figure 3. Performance analysis of UAI and other Deep Learning models on CIC-IDS2017 dataset

V. CONCLUSION

This paper uses the concept of support vector machine and isolation forest with active learning, through which unsupervised anomaly detection detected very efficiently. Using isolation forest, anonymous data can easily separate from the normal data. For testing of various machine learning and deep learning techniques, CIC-IDS2017 dataset was used. If we compare this framework with other frameworks, then in case of deep learning techniques, it gives better result in terms of performance. The future work is this framework is to use other deep learning techniques like CNN and check the performance.

VI. ACKNOWLEDGEMENT

This experiment is conducted on Chandigarh University Networks labs. We sincerely express my gratitude to Chandigarh University and its lab staff.

REFERENCES

[1] M. Luo, L. Wang, H. Zhang, and J. Chen, ``A research on intrusion detection based on unsupervised clustering and support vector machine," in Proc. 5th Int. Conf. Inf. Commun. Secur. (ICICS), Hohhot, China, S. Qing, D. Gollmann, and J. Zhou, Eds. Berlin, pp. 325_336. Springer, China (2003).

[2] X. Zhu and A. B. Goldberg, ``Introduction to semi- supervised learning," Synth. Lect. Artif. Intell. Mach. Learn., vol. 3, no. 1, pp. 1_130, (2009).

[3] N. Moustafa, J. Hu, and J. Slay, ``A holistic review of network anomaly detection systems: A comprehensive survey," J. Netw. Comput. Appl., vol. 128, pp. 33_55, (2019).

[4] D. Pelleg and A. W. Moore, "Active learning for anomaly and rarecategory detection," in NeurIPS, pp. 1073–1080, (2005).

[5] N. Abe, B. Zadrozny, and J. Langford, "Outlier detection by active learning," in ACM SIGKDD Conf. ACM, pp. 504–509, (2006).

[6] N. G¨ornitz, M. Kloft, K. Rieck, and U. Brefeld, "Toward supervised anomaly detection," Journal of Artificial Intelligence Research, vol. 46,pp. 235–262, (2013).

[7] D. M. Tax and R. P. Duin, "Support vector data description," Machine learning, vol. 54, no. 1, pp. 45–66, (2004).

[8] K. Veeramachaneni, I. Arnaldo, V. Korrapati, C. Bassias, and K. Li, "Aiˆ 2: training a big data machine to defend," in Big Data Sec. Conf. IEEE, pp. 49–54, (2016).
[9] M. Sharma, K. Das, M. Bilgic, B. Matthews, D. Nielsen, and N. Oza, "Active learning with rationales for identifying operationally significant anomalies in aviation," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer,pp. 209–225, (2016).
[10] S. Das, W.-K. Wong, A. Fern, T. Dietterich, and M. Siddiqui, "Incorporating feedback into tree-based anomaly detection," Workshop on Interactive Data Exploration and Analytics (IDEA), (2017).
[11] H.-J. Liao, C.-H. R. Lin, Y.-C. Lin, and K.-Y. Tung``Intrusion detection system: A comprehensive review," J. Netw. Comput. Appl., vol. 36, no. 1, pp. 16_24, (2013).
[12] P. Winter, E. Hermann, and M. Zeilinger, ``Inductive intrusion detection in _ow-based network data using one-class support vector machines," in Proc. 4th IFIP Int. Conf. New Technol., Mobility Secur., pp. 1_5,(2011).
[13] C.Wagner, J. Franáois, R. State, and T. Engel,``Machine learning approach for IP-_ow record anomaly detection," Lecture Notes Comput. Sci., vol. 6640, no. 1, pp. 28_39, (2011).
[14] Singh, Richa, Arunendra Singh, and Pronaya Bhattacharya. "A machine learning approach for anomaly detection to secure smart grid systems." In Research Anthology on Smart Grid and Microgrid Development, pp. 911-923. IGI Global, (2022).
[15] Singh, R., & Singh, A, "Challenges of Various Load Balancing Algorithms in Distributed Environment". In (2018) Journal International Journal of Information Technology and Electrical Engineering (ITEE) (pp. 9-13). ITEE(2018).
[16] Singh, R., Singh, A,& Bhattacharya P, "Challenges of Load Balancing Techniques in Grid Environment". In (2018) Journal International Journal of Information Technology and Electrical Engineering (ITEE) (pp. 1-6). ITEE(2018).
[17] M. F. Umer, M. Sher, and Y. Bi, ``A two-stage _ow- based intrusion detection model for next-generation networks," PLoS ONE, vol. 13, no. 1, (2018).
[18] P. Casas, J. Mazel, and P. Owezarski, Unsupervised network intrusion detection systems: Detecting the unknown without knowledge, Computer Communications, vol. 35, no. 7, pp.772–783, (2012).
[19] L. Parsons, E. Haque, and H. Liu, Subspace clustering for high dimensional data: A review, SIGKDD Explor. Newsl., vol. 6, no. 1, pp. 90–105, (2004).
[20]B. Sch¨olkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, Estimating the support of a high dimensional distribution, Neural Computation, vol. 13, no. 7, pp. 1443–1471, (2001).
[21] A. K. Jain, Data clustering: 50 years beyond K- means, Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, (2010).
[22] A. L. N. Fred and A. K. Jain, Combining multiple clustering's using evidence accumulation, IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 835–850, (2005).
[23] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, (2006).
[24] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in ACM SIGKDD Conf. ACM, pp. 665–674, (2017).
[25] R. Silva, M. Gonc¸alves, and A. Veloso, "A two- stage active learning method for learning to rank," J. Assoc. Inf. Sci. Technol., vol. 65, no. 1, pp. 109–128, (2014).
[26] R. Silva, M. Goncalves, and A. Veloso, "Rule-based active sampling for learning to rank," in ECML PKDD Conf., pp. 240–255., (2011).
[27] B. Settles, "Active learning," Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 6, no. 1, pp. 1–114, (2012).
[28] A. Ferreira, R. Silva, M. Gonc¸alves, A. Veloso, and A. H. F. Laender, "Active associative sampling for author name disambiguation," in JCDL Conf., pp. 175–184, (2012).
[29] M. Moreira, J. dos Santos, and A. Veloso, "Learning to rank similar apparel styles with economically- efficient rule-based active learning," in ACM ICMR Conf., pp. 361–370, (2014).
[30] M. Sharma, K. Das, M. Bilgic, B. Matthews, D. Nielsen, and N. Oza, "Active learning with rationales for identifying operationally significant anomalies in aviation," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 209–225, (2016).
[31] M. Bilgic and P. N. Bennett, "Active query selection for learning rankers," in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM,pp. 1033–1034, (2012).
[32] K. Veeramachaneni, I. Arnaldo, V. Korrapati, C. Bassias, and K. Li, "Aiˆ 2: training a big data machine to defend," in Big Data Sec. Conf. IEEE, pp. 49–54, (2016).
[33] S. Das, W.-K. Wong, T. Dietterich, A. Fern, and A. Emmott, "Incorporating expert feedback into active anomaly discovery," in ICDM. IEEE, pp. 853– 858, (2016).
[34] S. Das, W.-K. Wong, A. Fern, T. Dietterich, and M. Siddiqui, "Incorporating feedback into tree-based anomaly detection,"Workshop on Interactive Data Exploration and Analytics (IDEA), (2017).
[35] Y. Bengio, G. Mesnil, Y. Dauphin, and S. Rifai, "Better mixing via deep representations," in ICML, pp. 552–560, (2013).
[36] V. Engen, J. Vincent, and K. Phalp, ''Exploring discrepancies in findings obtained with the KDD Cup '99 data set,''' Intell. Data Anal., vol. 15, no. 2, pp. 251–276, (2011).
[37] L. Dhanabal and S. P. Shantharajah, ''A study on NSL-KDD dataset for intrusion detection system based on classification algorithms,'' vol. 4, no. 6, pp. 446–452, (2018).
[38] S. Revathi and A. Malathi, ''A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection,'' Int. J. Eng. Res. Technol., vol. 2, no. 12, pp. 1848–1853, (2013).
[39] D. H. Deshmukh, T. Ghorpade, and P. Padiya, ''Improving classification using preprocessing and machine learning algorithms on NSL-KDD dataset,'' in Proc. Int. Conf. Commun., Inf. Comput. Technol. (ICCICT), pp. 1–6), (2015).