Journal of Ubiquitous Computing and Communication Technologies (UCCT) (2020)
Vol.02/ No.04
Pages: 203-213
https://www.irojournals.com/jucct/
DOI: https://doi.org/10.36548/jucct.2020.4.003

Deep Learning Approach to DGA Classification for Effective Cyber Security

Karunakaran P
Senior Instructor,
Bahrain Training Institute,
Bahrain.
karunakaran.patcha@bti.moe.bh

ABSTRACT- In recent years, invaders are increasing rapidly in an internet world. Generally, in order to detect the anonymous attackers algorithm needs more number of features. Many algorithms fail in the efficiency of detection malicious code. Immediately this codes will not infect the system; it will attack server after communicate later. Our research focuses on analyzing the traffic of botnets for the domain name determination to the IP address of the server. This botnet creates the domain name differently. Many domains are generated by attackers and create the huge Domain Name System (DNS) traffic. In this research paper, uses both public and real time environments datasets to detect the text features as well as knowledge based feature extraction. The classifying of Domain Generation Algorithm (DGA) generated malicious domains randomly making the efficiency down in many algorithms which were used preprocessing without proper feature extraction. Effectively, our proposed algorithm is used to detect DGA which generates malicious domains randomly. This effective detection of our proposed algorithm performs with text based label prediction and additional features for extraction to improve the efficiency of the model. Our proposed model achieved 94.9% accuracy for DGA classification with help of additional feature extraction and knowledge based extraction in the deep learning architecture.

Keywords*: deep learning, Cyber security, Domain Generation Algorithm*

1. INTRODUCTION

The increases of invaders in the internet world are making hurtful matter day by day. The server will corrupt by malwares are program which intrudes the data and files. Obviously, all the system performance and speed will go down day by day. This malwares are program attacks in android devices too. The droid detector is used to detect the android attacks [1, 2]. Many intrusion detection systems have been established to detect the network attacks surroundings [3]. There must be an essential of high accuracy to detect those invaders for cyber security. One of the destructive hazards in the internet is Botnets which is useful for attacking by malware [4, 5]. Previously the IP address is static while they are in attacking mode. Now this Botnets are using domain fluxing concepts which will adapt the DGA easily. This DGA is generating large number of pseudo domain names periodically which is useful to shift one domain to another domain at constant intervals. This dynamic change in the domain to domain for the server is big challenging research task of this domain. Generally the Botnets works through Command and Control (C&C) server. Generally, domain length is concentrated at the intervals of 10-20 [5, 6, 7]. Many algorithms are particularly LSTM that achieving high accuracy with many multi class classifications. This figure 1 shows that some cyber security issues.

ISSN: 2582-337X
Submitted: 04.12.2020
Accepted: 30.12.2020
Published: 06.01.2021

Figure 1 Cyber security Issues

Zeus DGA contains MD5 of the details (year, month, day and sequence number) and domain name having random text combination (for example trwidizlnicwtut). Existing Torpig DGA techniques is generating domain names with random number generator between the texts. Kwyjibo is generating a dictionary based random words that are easy. Another DGA srizbi is used data transformation using exclusive or (EX-OR) operation [8, 9, 10].

2. ORGANIZATION OF THE RESEARCH

The structure of the research article organized as follows; Section 2 gives literature survey of recent cyber security approach. Section 3 provides the description of theoretical analysis. Section 4 delivers description of results and discussion finally the conclusion and further improvement is in section 5.

3. RELATED WORKS

Tran et al proposes the malicious domain detected by DGA binary classification model that LSTM.MI model pools both class classification named binary and multiclass. This study investigates that binary classification and multiclass classification in single model [11]. The Recurrent Neural Network (RNN) application is appropriate when there is back propagation algorithm used. These applications are dealt with sequential inputs with feed forward to the training and testing data. The many forecasting applications of the current output depend on the several previous samples which is shown in the figure 3 [5].

Liou et al introduces auto encoder for resemblance investigation. This proposes study is dealing with feature refinement technology. They didn’t use labels for classification. Generally, the feature extraction using principal component analysis (PCA) and auto encoder is shown in figure 2. In the auto encoder, this technology is not suitable to extract the features when both classifications are used at the same time [12].

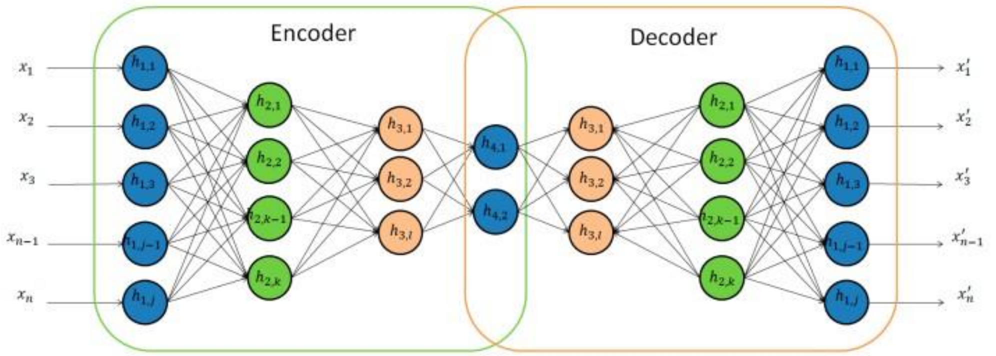

Figure 2 Structure of auto encoder

The classification of DGA is constructed by two processes named machine learning classification algorithm and clustering based DNS. Here this machine learning algorithm applies DNS blacklist guidelines and try to detect DGA through features. This study was suggested by Chin et al [13]. Woodbridge et al introduces a DGA classifier for a real time prediction by manually created features. They constructed LSTM network to better deep learning that sequential data is too long distance. They achieved 90% of detection rate of DGA generated domain names [14]. Qiao et al set the fixed length of DGA domain name. Because of the deep learning models require static generation of domain names to learn accurately[15].

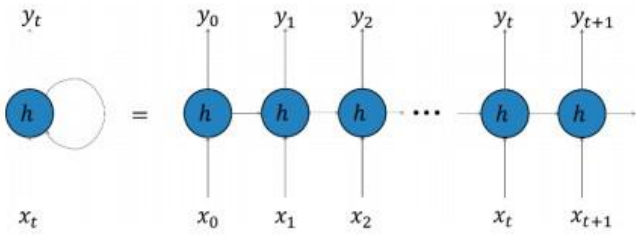

Figure 3 the RNN Classification

Another study Yu et al proposes extracted 11 features from domain to detect the DGA domain name by convolutional neural network (CNN) and LSTM. The final authentication was compared with many machine learning algorithms. The proposed algorithm is accurate with 72.89 % during their experimental analysis [16]. The figure 4 shows input, hidden, output layers structure of CNN classification.

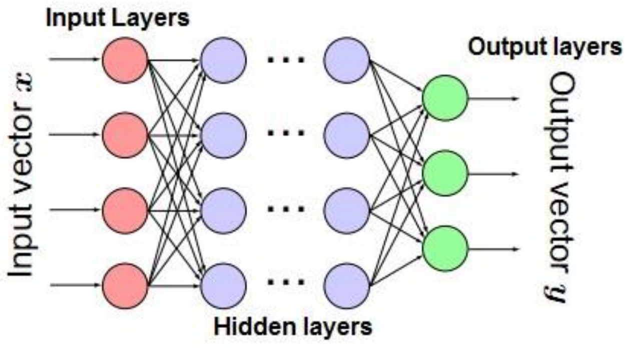

Figure 4 Layer structure of CNN Classification

The paper proposes there is lots of comparison of CNN and many neural networks based architectures. This study believes combination of CNN and LSTM model to provide good accuracy. The dataset was trained and evaluated huge number of domains. Multilayer perceptron (MLP) algorithms are comparing the performance of RF model with same features as the processed approach. They achieved 98% accuracy with the deep learning algorithms. Above mentioning many algorithms are failing to detect the new type of domain names generates by DGA which cannot be blacklist or filtering [16]. Hence, there must be more research on developing an effective deep learning algorithm for DGA analysis to fulfill the research gap. The main goal of this research article is that protect the device using effective algorithm to create the security events in contrast to violence in an internet.

Problem Statement

One of the most complexities in an internet protocol is that detection of type of malicious software which is used by the attackers. Because of the attackers are updated their algorithm periodically due to DNS queries. The detection of DGA generated domain names effectively is a big challenging task in an internet domain. During the filtering the domain requires significant time to extract the Botnet code.

Proposed Solution

Mainly, more number of features is required to detect the anonymous attackers in an effective way in cyber security. In this work, the text features with additional datasets are collected from both public sources as well as real time environments for improvement of effectiveness. Currently the knowledge based feature extraction is absent in cyber security. The intelligent or knowledge based feature is the one of the solution and it is to detect an anonymous attackers with good efficiency. The DGA is scrutinized by deep learning architecture.

4. THEORETICAL ANALYSES

The malicious code consists of domain address and IP address to access the C & C server through security guidelines itself. This section is having DGA detection method, Proposed DGA classification for effective cyber security.

4.1 DGA Detection Method

Generally, the DGA can creates millions of domain by giving predetermined seed value. This creation uses times series data with seed value which will be predicted attackers easily. The attacker calculates the domain address in advance by legimate procedure [5, 17]. The table 1 contains the description of multiclass dataset.

ISSN: 2582-337X
Submitted: 04.12.2020
Accepted: 30.12.2020
Published: 06.01.2021

4.2 Proposed DGA classification

There are two type of classification that binary and multi class classification to detect the malicious model. Here we can use both classification methods to detect malicious domain. The basic typical methods are auto encoder and PCA (Principal Component Analysis) used for feature extraction. This method is an inefficient for above mentioned both class classification. Our proposed model uses three layers name input, output, and hidden layer. Here the extracted features are labeled by output layer and trained. So this type of approach is suitable for both type of classification. This is the kind of feature refining process for the data [18, 19, 20]. The deep learning process will suffer over-fitting problems due to more features. The In order to prevent over-fitting problem in learning method, the feature reduction will be done here. The proposed model performance also increases effectively. After feature refining effectively, our proposed algorithm works decent timing for overall execution. The classification of DGA by ML algorithm and clustering based DNS is conducted that reduce the false positive (FP) in the primary results in the clustering based classification of DGA [5, 21, 22].

Table 1 Description of Multiclass dataset

Label	Domain Type	Training	Testing 1	Testing 2
0	Benign	10000	120000	40000
1	Banjori	15000	25000	10000
2	Corebot	15000	25000	10000
3	Dircrypt	15000	25000	300
4	Dnschanger	15000	25000	10000
5	Fobber	15000	25000	800
6	Murofet	15000	16667	5000
7	Necurs	15000	20445	6200
8	Newgoz	15000	20000	3000
9	Padcrypt	15000	20000	3000
10	Proslikefan	15000	20000	3000
11	Pykspa	15000	25000	2000
12	Qadars	15000	25000	2300
13	Qakbot	15000	25000	1000
14	Ramdo	15000	25000	800
15	Ranbyus	15000	25000	500
16	Simda	15000	25000	3000
17	Suppobox	15000	20000	1000
18	Symmi	15000	25000	500
19	Tempedreve	15000	25000	100
20	tinba	15000	25000	700

Our proposed approach maintains the character dictionary for character level embedding structure (CLES). The text is to represent the string to number which is having a unique value. This CLES is used to hold character and its converts into number.

Though different size of character CLES will be articulated with fixed size.

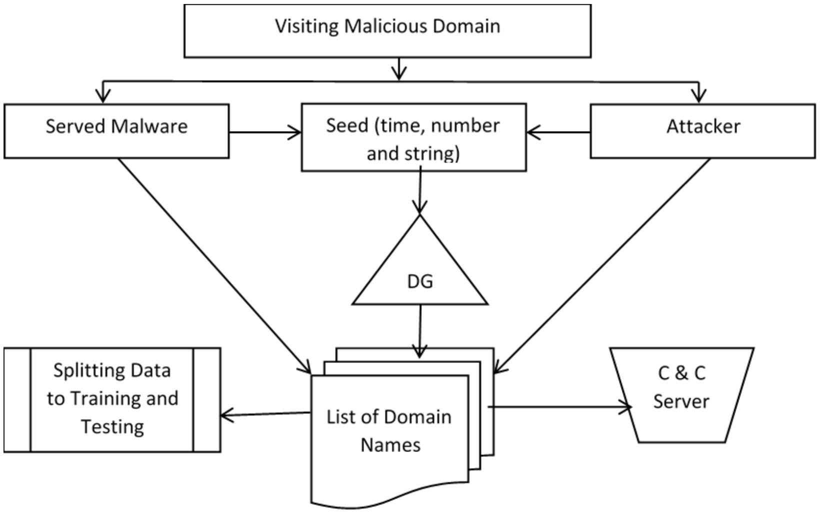
Visiting Malicious Domain
Served Malware
Seed (time, number and string)
Attacker
DG
Splitting Data to Training and Testing
List of Domain Names
C & C Server

Figure 5 Input Section of proposed Model

The figure 5 shows the input section for the proposed model. When users visits malicious domain, malware will be served to users. This scenario will know by attackers too. The malware requires seed to execute the DGA. Also attackers are practiced with the known seeds. An attacker selects and registers a domain that points to the IP address of the C & C server [23, 24, 25]. At the same time our proposed algorithm searches for domains registered by attackers among generated domains. Malware starts to detect the malicious codes effectively by using C & C servers. Our text character has embedded as input of classification shown in figure 6. The Preprocessing has done for feature extraction. The parameters are activating between convolutional and pooling layer of the neural network.

The kernel size also will be fixed between these layers with default values. Then the transformation verification is designed of hyper parameters to set the optimal. The selection of additional features is suitable for very effective classification. Also the changing of size of vector for kernel can be done by adjusting the layers for effectiveness. There is observation that smaller vector size comes from more number of convolution and pooling layer and vice versa.

Here more number of features suitable for classifying the classes and to set the number for dense node. We realized that the number of dense layers and nodes increases then processing speed gets down.

ISSN: 2582-337X
Submitted: 04.12.2020
Accepted: 30.12.2020
Published: 06.01.2021

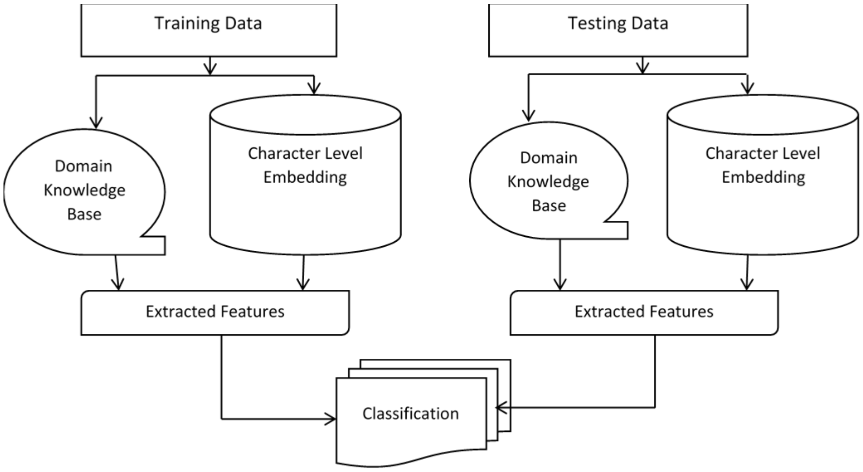

Figure 6 Training and testing section of proposed Model

After setting the optimum number of nodes and dense layers, the classification and overall processing speed increases. For the effectiveness, the features can be extracted the domain with knowledge based extraction. Domain Name System (DNS) is used to map the domain name instead of IP addresses. Generally, the subdomain, domain name, SLD, TLD comprised version called Domain address. The user can register any type of string character in flexible area. The SLD and TLD indicate the information of country belongs to [5, 21, 22]. The DGA creates domain names randomly with letters with numbers. Most of the malicious websites generates by DGA when it was with a fixed length of word and number of dot was two or less. Some more additional features are extracted from this structured format of domain address. This structured format of domain address is an investigated to feature extraction. It provides better accuracy and effective of the proposed algorithm.
We achieved the effective classification for DGA for cyber security.

5. EXPERIMENTAL ANALYSES

The classification metrics are measured here to find the overall performance of the proposed model. There are many metrics present to evaluate the performances of the model. The metrics for our proposed methods includes accuracy, precision, recall, F1 score and Detection time. These evaluation metrics are derived from description of predicted class as shown in table
2. The f1 score is measuring for the unbalanced classes' rate. The detection time is very less compared with other deep learning model.

Table 2 The description of predicted Class

		Malicious	**Benign**
Actual Class	Malicious	True Positive (TP)	False Negative
	(FN) Benign	False Positive (FP)	True Negative (TN)

The following formulas are used to measure the metric measurement for our proposed model.

$$Precision = \frac{TP}{TP + FP}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{TP + TN}$$

$$\boldsymbol{F1\ Score = \frac{2.TP}{2.TP + FP + FN}}$$

By varying the classification threshold, we can calculate false positive (FP), True Positive (TP) respectively. The accuracy will be calculated for multiclass problems easily. The remaining metric parameters are not calculated for multi class problems. So the remaining metrics are determined for each class. Mainly accuracy will be calculated for multi class problem. But our proposed model can be evaluated with only accuracy. Basically our proposed algorithm is designed for better effectiveness and uses a different datasets, subset of a given dataset. So this is not appropriate comparison between the models developed with the accuracy only. If training and testing data sets are exactly same, comparison could be materialized. The efficiency percentages are tabulated and values are plotted in the graph is shown in figure 7. The efficiency is calculated by the algorithm conducts the iteration for number of attacks and success rate of it is called efficiency.

Table 3 The Comparison of proposed methods with measuring metrics

S.No	Measuring Metrics	CNN	RNN	Proposed Method
1	Efficiency	91.68	90.5	94.9
2	Accuracy	95.6	96.7	98.3
3	Precision	0.7	0.67	0.89
4	Recall	0.65	0.6	0.864
5	F1-Score	90.8	92.1	95.56
6	Detection time (sec)	1.88	2.19	0.876

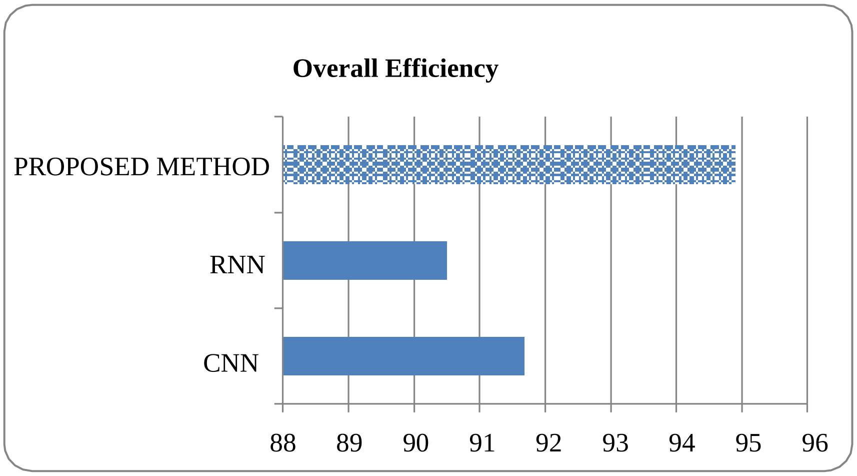

Figure 7 Proposed Method Overall Efficiency Comparison

6. CONCLUSION

This paper surveys the significant effective algorithm is designed to classify the domain generation algorithm very successfully. Our algorithm is more efficient because of that attackers know only how DGA algorithm operates. Mostly malicious codes are received by invaders command after spoil the server. So the attackers use C&C servers to give commands. The malware needs IP address for communication. The attackers will continue with the blocking technique and randomly generates domain names with the help of DGA. Our algorithm is working with text classification with additional features provides better efficiency to detect the malicious domain in the network. The damage in the server or system directly depends on the increment of intelligent malware action. Our idea is that special features and additionally knowledge base feature extraction provides good results at the end of this research. It is proved and gives new effective method to classify the DGA. Our proposed model achieved 94.9% accuracy for classification. The further improvement of the proposed techniques should have more training and test data set. Introduce the feature refinement technology in our proposed algorithm gives better speed performance and detection time in a real time environment.

ACKNOWLEDGEMENT

We are most grateful to Eng. Mussie, (Retired) Ministry of Telecommunication, Department of Electrical & Electronics Engineering, Mainefhi College of Engineering, for his assistance and encourage this research.

REFERENCES

[1] M. A. Al-Garadi, A. Mohamed, A. K. Al-Ali, X. Du, I. Ali and M. Guizani, "A Survey of Machine and Deep Learning Methods for Internet of Things (IoT) Security," in IEEE Communications Surveys & Tutorials, vol. 22, no. 3, pp. 1646-1685, thirdquarter 2020, doi: 10.1109/COMST.2020.2988293.

ISSN: 2582-337X
Submitted: 04.12.2020
Accepted: 30.12.2020
Published: 06.01.2021

[2] T. T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," in IEEE Communications Surveys & Tutorials, vol. 10, no. 4, pp. 56-76, Fourth Quarter 2008, doi: 10.1109/SURV.2008.080406.

[3] Wu, Shelly & Banzhaf, Wolfgang. (2010). The use of computational intelligence in intrusion detection systems: A review. Applied Soft Computing. 10. 1-35. 10.1016/j.asoc.2009.06.019.

[4] Berman, Daniel & Buczak, Anna & Chavis, Jeffrey & Corbett, Cherita. (2019). A Survey of Deep Learning Methods for Cyber Security. Information. 10. 122. 10.3390/info10040122.

[5] Babu R, Mohammed & R, Vinayakumar & Kp, Soman. (2018). A short review on Applications of Deep learning for Cyber security.

[6] Chen, Mu-Yen & Chiang, Hsiu-Sen & Lughofer, Edwin & Egrioglu, Erol. (2020). Deep learning: emerging trends, applications and research challenges. Soft Computing. 24. 1-4.10.1007/s00500-020-04939-z.

[7] A. Humayed, J. Lin, F. Li and B. Luo, "Cyber-Physical Systems Security—A Survey," in IEEE Internet of Things Journal, vol. 4, no. 6, pp. 1802-1831, Dec. 2017,doi:10.1109/JIOT.2017.2703172.

[8] Hwang, Chanwoong & Kim, Hyosik & Lee, Hooki & Lee, Taejin. (2020). Effective DGA- Domain Detection and Classification with TextCNN and Additional Features. Electronics. 9.1070. 10.3390/electronics9071070.

[9] Deng, Li & Yu, Dong. (2013). Deep Learning: Methods and Applications. Foundations and Trends in Signal Processing. 7. 10.1561/2000000039.

[10] Amara, Dinesh & Thodupunoori, Harish & R, Vinayakumar & Kp, Soman & Poornachandran, Prabaharan & Alazab, Mamoun & Venkatraman, Sitalakshmi. (2019). "Enhanced Domain Generating Algorithm Detection Based on Deep Neural Networks" 10.1007/978-3-030-13057-2_7.

[11] Tran, D.; Mac, H.; Tong, V. A LSTM based framework for handling multiclass imbalance in DGA botnet detection. Neurocomputing 2018, 275, 2401–2413.

[12] Liou, C.Y.; Cheng, W.C.; Liou, J.W.; Liou, D.R. Autoencoder for words. Neurocomputing 2014, 139, 84–96.

[13] Chin, T.; Xiong, K.Q.; Hu, C.B.; Li, Y. A machine learning framework for studying domain generation algorithm (DGA)-based malware. In Proceedings of the International Conference on Security and Privacy in Communication Systems, Singapore, 8–10 August 2018.

[14] Woodbridge, H.S.J.; Anderson, A.A.; Grant, D. Predicting domain generation algorithms with long short-term memory networks. arXiv 2016, arXiv:1611.00791.

[15] Qiao, Y.; Zhang, B.; Zhang, W.; Sangaiah, A.K.; Wu, H. DGA Domain Name Classification Method Based on Long Short-Term Memory with Attention Mechanism. Appl. Sci. 2019, 9,4205.

[16] Yu, B.; Daniel, L.G.; Pan, J.; Martine, D.C.; Anderson, C.A.; Nascimento, Y. Inline DGA detection with deep networks. In Proceedings of the 2017 IEEE International Conference on Data Mining Workshops (ICDMW), New Orleans, LA, USA, 18–21 November 2017; pp. 683–692.

[17] Pechenizkiy, Mykola & Puuronen, Seppo & Tsymbal, Alexey. (2003). Feature Extraction for Classification in Knowledge Discovery Systems. 526-532. 10.1007/978-3-540-45224-9_72.

[18] Alkahtani, Hasan et al. "Adaptive Anomaly Detection Framework Model Objects in Cyberspace." Applied Bionics and Biomechanics 2020 (2020): https;//doi.org/10.1155/2020/6660489

[19] Dasgupta D, Akhtar Z, Sen S. Machine learning in cybersecurity: a comprehensive survey. The Journal of Defense Modeling and Simulation. September 2020. doi:10.1177/1548512920951275

[20] Bi, Mengxiao et al. "Very deep convolutional neural networks for LVCSR." INTERSPEECH (2015).

[21] P. Raghavan and N. E. Gayar, "Fraud Detection using Machine Learning and Deep Learning," 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), Dubai, United Arab Emirates, 2019, pp. 334-339, doi: 10.1109/ICCIKE47802.2019.9004231.

[22] Sainath, Tara & Kingsbury, Brian & Mohamed, Abdel-rahman & Dahl, George & Saon, George & Soltau, Hagen & Beran, Tomas & Aravkin, Aleksandr & Ramabhadran, Bhuvana. (2013). "Improvements to Deep Convolutional Neural Networks for LVCSR" 10.1109/ASRU.2013.6707749.

[23] Lai, Yingxu & Zhang, Jingwen & Liu, Zenghui. (2019). "Industrial Anomaly Detection and Attack Classification Method Based on Convolutional Neural Network". Security and Communication Networks 2019. 1-11. 10.1155/2019/8124254.

[24] Wu, Yirui et al. "Network Attacks Detection Methods Based on Deep Learning Techniques: A Survey." Security and Communication Networks 2020 (2020): 1-17.

[25] Bakhshi, Taimur & Ghita, B.V.. (2016). "On Internet Traffic Classification: A Two-Phased Machine Learning Approach" Journal of Computer Networks and Communications. Volume 2016 (2016). 21 pages. 10.1155/2016/2048302.