International Journal of Scientific Engineering and Science
Volume 5, Issue 7, pp. 23-27, 2021.
ISSN (Online): 2456-7361

A Robust-Based Framework towards Resisting Adversarial Attack on Deep Learning Models

P. S. Ezekiel[1], O. E. Taylor[2], F.B. Deedam-Okuchaba[3]
[1, 2, 3]Department of Computer Science, Rivers State University, Port Harcourt, Nigeria
*Corresponding Author: tayonate @ yahoo.com, Tel.: +2348034448978

Abstract— *Adversarial attack is a type of attack executed by an attacker in other to confuse a deep learning model to falsely classify a wrong input data as the correct data. This attack is being executed in two ways. The first one is the Poisoning attack which is being generated during training of a deep learning model. And the second one is the Evasion attack. In the evasion assault, the assaults' is being done on the test dataset. An evasion assault happens when the computer network is taken care of an "adversarial model", a painstakingly perturbed info that looks and feels precisely equivalent to its untampered duplicate to a human however that totally loses the classifier that. This system presents a robust based model towards the resistance of adversary assaults on deep learning models. The system presents two models using convolutional neural network algorithm. This model was trained on a Modified National Institute of Standards and Technology dataset (MNIST). An adversary (evasion) attacks was generated to in other to fools this models to misclassify result, therefore, seeing the wrong input data to be the right one. This adversarial examples was generated using a state-of-the-art library in python. The generated adversarial examples was being generated on the test data, in which the first model fails in resisting the attack, while the second model, which is our robust model resisted the adversarial attack on a good number of accuracy when tested for the first 100 images.*

Keywords— *Adversarial attack, Deep Learning, Evasion attack, Poisoning Attack, MNIST Dataset*.

I. Introduction

Adversarial attacks are explicitly intended to feat weaknesses in a given Deep Learning Algorithm. These weaknesses are recreated via preparing the learning algorithm under different attacks and approaches. The attack sequence are thought to be defined by a smart adversary. The ideal attack strategy is defined to tackle one or more optimization issues over one or more attack sequence. A learning method built over adversarial settings gets robust to such weaknesses in training and testing dataset. The different adversarial learning methods is different in presumptions, in regards to the adversary's information, security infringement, attack systems and attack impact [1]. The surveillance application of a deep neural Network (DNNs) like Intrusion Detection System (IDS), malware identification, spam-sifting have become fundamentals in protection of information, characterization, and prediction. These distinctive kind of errands are depending on the insight to construct a model that ordinarily group and separate among "benign" and "malware" samples, similar to attack and benevolent bundles. With the quick increment of utilizing DNNs and the weakness of DNNs to adversarial attack, the complexity of attack procedures apparatuses is additionally expanded. Accordingly, different explores [2] [3] track down that various attacks add extreme difficulties to weaknesses of DNN engineering plan. The way that the preparation of DNNs depends on information, the arrangement undertaking can be controlled by perturbation inputs called adversarial samples. Adversarial samples are regularly obscure, and they are intended to dependably delude a machine learning model toward mistaken characterization and dodge discovery. Toward covering the issue of getting classifiers against antagonistic assaults, we study the ill-disposed assaults against DNNs and their vigor.

The new advances in deep learning [4] have prompted forward leaps in a scope of long-standing machine learning assignments e.g., natural language processing [5], image classification [6], and in any event, playing Go, empowering numerous sequence recently thought to be stringently exploratory. Notwithstanding, it is notable that deep neural organization (DNN) models are innately helpless against adversarial sources. Normally, adversarial data sources are made via cautiously perturbing legitimate examples under the direction of the gradient data of target deep neural networks (i.e., "white-box" attacks) [7]. In the interim, many cloud-based specialist co-ops, including Amazon, Google, Microsoft, BigML, and others all have emerged to give Machine Learning-as-a-administration (MLaaS) stages. On such stages, progressively numerous business and exclusive DNN models are being conveyed with openly open interfaces ("prescient APIs"), which permit clients to inquiry the backend models with contributions of interests and charge clients on a compensation for each question premise. In this paper we created a robust model towards resisting adversary assaults on deep learning models.

II. Related Works

Investigating Resistance of Deep Learning-based IDS against Adversaries using min-max Optimization [8] applied the min-max (or saddle point) systemization in the Intrusion Detection System space and researched the viability of the internal augmentation issue on the power of adversarial trained model which is the external minimization issue. They created adversarial examples utilizing four existing techniques dF GSMs, rF GSMs, BGAs and BCAs. They investigated if BGAs and BCAs, which are intended for the discrete component area can create adversarial examples in the ceaseless features space. Their experimental result shows that doing dimensionality decrease utilizing Principal Component Analysis on the dataset helped in diminishing avoidance rates.

AdvMind: Inferring Adversary Intent of Black-Box Attacks [9] presented AdvMind, which is another class of assessment models that construe the adversary expectation of black-box adversarial attacks in a smart and brief way through a broad exact assessment on a benchmark datasets and state-of-the-art black-box-attack. Their findings shows that on average AdvMind identifies the adversary motive with more than 75% accuracy subsequent to observing three query batches and then builds the expense of versatile attacks by more than 60%. The exact assessment with respect to benchmark datasets, well known Deep Neural Networks, and state-of-the-art attacks approves the adequacy of AdvMind.

Adversarial Attacks with Multiple Receivers Against Deep Learning-Based Modulation Classifiers [10] carried a survey on a remote communication framework, where a transmitter conveys messages to a recipient with various regulation types, while the recipient groups the modulation types of the received message utilizing its deep learning classifier. Simultaneously, an adversary sends adversarial perturbation utilizing its different receivers to trick the classifier into giving false result of the received message. For adversarial machine-learning viewpoint, they use different receivers at the adversary to uplift the adversarial (evasion) attack execution. Two primary concerns are thought of while attaining the various receivers at the adversary, in particular the force distribution among receivers and the usage of channel variety. In the first place, they show that autonomous adversaries, each with a solitary receivers can't uplift the attack execution contrasted with a solitary adversary with various receivers utilizing a similar complete force. They also acquaint an attack that will send the perturbation adversarial attack through the channel with the biggest channel acquired at the image level.

Adversary for Social Good: Protecting Familial Privacy through Joint Adversarial Attacks [11] proposed a novel adversarial attack algorithms for the good of social network. In the first place, they start from traditional visual family problem, and show that familial data can without much of a stretch be presented to attackers by associating sneak shots to social networks. Second, to secure family protection on informal organizations, they propose a novel algorithm for adversarial attack that produces both features of adversary and chart under a given spending plan. In particular, the two features on the hub and edges, between hubs was perturbed steadily with the end goal that the test pictures and its family data can't be distinguished accurately through customary GNN. Their exploratory result on a well-known visual social dataset shows that their defense methodology can altogether moderate the effects of family data spillage.

Deep Learning Based Intrusion Detection with Adversaries [12] evaluated state-of-the-art attack methods in the deep learning put together intrusion identification space with respect to the NSL-KDD informational index. The weaknesses of neural organizations utilized by the intrusion recognition frameworks are tentatively approved. The jobs of individual features in producing adversarial models were investigated in their study. For adversarial attack, they executed a DeeFool attack on which they acquired an accuracy of about 17% on Dos and 87.25% on R2L.

Adversarial Deep Learning Models with Multiple Adversaries [13] developed an algorithm for adversarial learning that will be administered for Convolutional Neural Networks (CNN) specifically with a goal of creating little changes to the information conveyance characterized over positive and negative class names so that the subsequent data distributed is misclassified by the CNN. The learning algorithm creates adversarial controls by systemizing a multiplayer stochastic game focusing on the grouping execution of the CNN. The multiplayer stochastic game is communicated as far as two-player consecutive games. Each game comprises of associations between two players a smart adversary and the learner CNN to such an extent that a player's result work increases with collaborations. Following the intermingling of a successive non-helpful Stackelberg game, every two-player game is tackled for the Nash balance. The outcomes propose that game hypothesis and developmental algorithms are viable in getting deep learning models against execution weaknesses mimicked as attacks situations from numerous adversaries.

Adversarial Examples Against the Deep Learning Based Network Intrusion Detection Systems [14] carried a survey on the reasonableness of adversarial model in the space of organization intrusion detection systems (NIDS). In particular, they executed a survey on how adversarial models influence the presentation of deep neural network (DNN) trained to identify unusual practices in the black-box model. They exhibit that adversaries can produce powerful adversarial models against DNN classifier prepared for NIDS in any event, when the inward data of the objective model is confined from the adversary.

Addressing Adversarial Attacks against Security Systems Based on Machine-Learning [15] focused on adversarial assaults that expect to influence the discovery and expectation abilities of machine-learning models. They considered reasonable kinds of poisoning and evasion attacks focusing on security arrangements dedicated to malware, spam and intrusion recognition. They investigate the potential harms that an assailant can cause to a digital indicator and present some current and unique guarded strategies with regards to intrusion identification frameworks.

III. Methodology

Data: The system architecture is made up of the Modified National Institute of Standards and Technology dataset (MNIST) dataset which we used as our training data. This dataset was imported from keras library in python. The dataset contains images of digits starting from zero (o) to nine (9). This dataset will be trained using a convolutional neural network algorithm with a total layer of three.

Adversarial Attack: Adversarial attack is a type of attack executed by an attacker in other to confuse a deep learning model to falsely classify a wrong input data as the correct data. This attack is being executed in two ways. The first one is the Poisoning attack which is being generated during training of a deep learning model. And the second one is the

Evasion attack. In the evasion assault, the assaults' is being done on the test dataset. An evasion assault happens when the computer network is taken care of an "adversarial model", a painstakingly perturbed info that looks and feels precisely equivalent to its untampered duplicate to human, however that totally loses the classifier. Evasion assault can also be seen as the bypassing of data security for the purpose of attack or exploit.

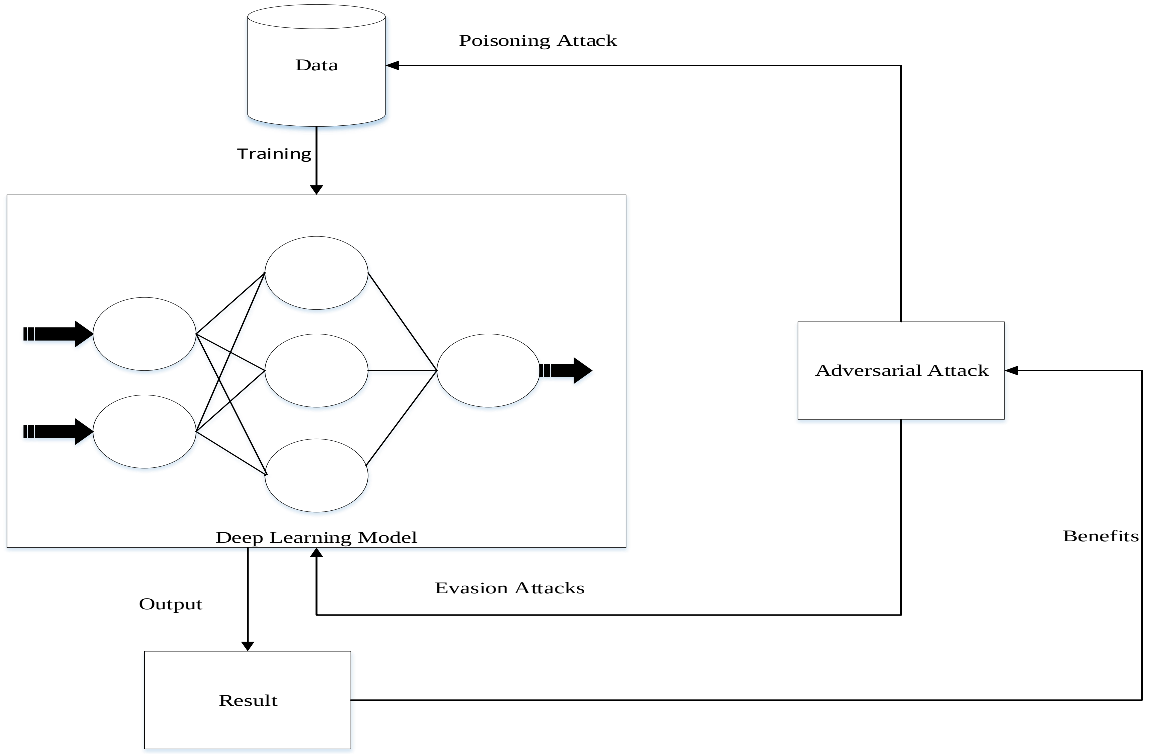


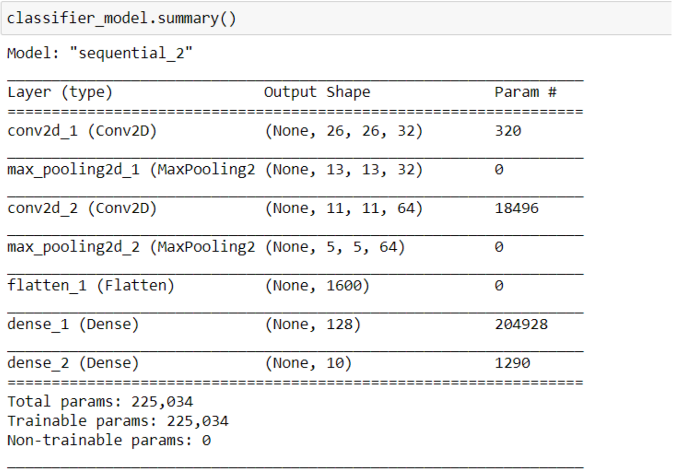
Figure 1: Architecture of the Proposed System

Algorithm for Adversarial Examples

```
Input: Attack and Benign training set from D
Output: Adversarial trained model, $x^*$ adversaries
Load attack and benign data (train, test & validation)
Extract features x = {x1,x2,...,xn}
Construct CNN model C
Define loss function T using"RMSprop"
Define inner-maximization M
Batch ← 100
repeat
    Read Batch of samples
    if Evasion method != Natural then
        Batch∗ ← M(Attack Batch, T, Evasion)
        start Adversarial Learning (Batch$^*$)
        do Test(Batch*)
    else Evasion method == Natural
        start train (Batch)
        do Test (Batch)
    end if
        until epoch=50 and C network converge
        procedure INNER-MAXIMIZATION(x,y,T,s,method)
        Computer natural loss for original samples
        Initialize starting point
        Compute natural loss for original samples
        Compute gradient for the loss T
        Compute the new adversarial sample
end procedure
```

IV. Result and Discussion

Adversarial attack has become a major challenge in deep neural network has it fools a deep neural network algorithm in classifying a wrong input data as the right data. The dataset used in this study is that of the Modified National Institute of Standards and Technology dataset (MNIST). The dataset comprises of images that contains digits or numbers ranging from 0-9. The dataset is being divided into a training data and a test data. This system proposed two deep learning model on which an adversarial attack will be executed on the both models using Fast Gradient Sign Method. The Modified National Institute of Standards and Technology dataset (MNIST) will be used as input data for the both model. The both models will be trained using a convolutional neural network algorithm, but with a different dense layer. Figure 2 and 3 shows the summary of the first and second model. The first model was trained normally while the second model was trainedin other to resist adversarial attack. The adversarial examples that was executed is the evasion attack. This attack is being executed on the test data with the aim of fooling the deep learning based models for seeing the wrong input data as the right input data. The evasion attack was executed on the first model for the first 100 images and the model wasn’t able to resist the attack. The first model classified the input data wrongly for 79 times and it classified the input data to be true (correctly) for 21 times. This shows that the first model was vulnerable to adversarial attack. For the second model which is we made robust in other to resist adversarial attack misclassified the input data to be true for just 1 (one) time and classified the input data to be true for 99 times. Figure 4 shows a graphical representation of the adversarial attack on the first and second model on the first 100 images.

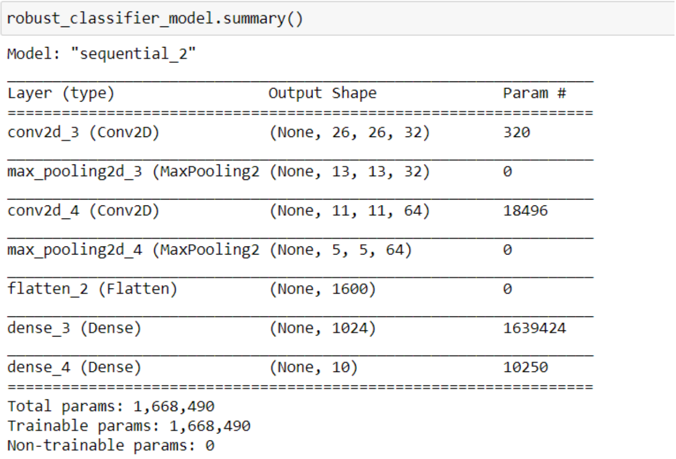


```
classifier_model.summary()

Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_1 (Conv2D)            (None, 26, 26, 32)        320
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 13, 13, 32)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 11, 11, 64)        18496
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 5, 5, 64)          0
_________________________________________________________________
flatten_1 (Flatten)          (None, 1600)              0
_________________________________________________________________
dense_1 (Dense)              (None, 128)               204928
_________________________________________________________________
dense_2 (Dense)              (None, 10)                1290
=================================================================
Total params: 225,034
Trainable params: 225,034
Non-trainable params: 0
_________________________________________________________________
```

Figure 2: Model 1: Summary

This shows the total parameters, layers used in training our first model to resist adversary attack



```
robust_classifier_model.summary()

Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d_3 (Conv2D)            (None, 26, 26, 32)        320
_________________________________________________________________
max_pooling2d_3 (MaxPooling2 (None, 13, 13, 32)        0
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 11, 11, 64)        18496
_________________________________________________________________
max_pooling2d_4 (MaxPooling2 (None, 5, 5, 64)          0
_________________________________________________________________
flatten_2 (Flatten)          (None, 1600)              0
_________________________________________________________________
dense_3 (Dense)              (None, 1024)              1639424
_________________________________________________________________
dense_4 (Dense)              (None, 10)                10250
=================================================================
Total params: 1,668,490
Trainable params: 1,668,490
Non-trainable params: 0
```

Figure 3: Model 2 Summary

This shows the total parameters, layers used in training our second model to resist adversary attack

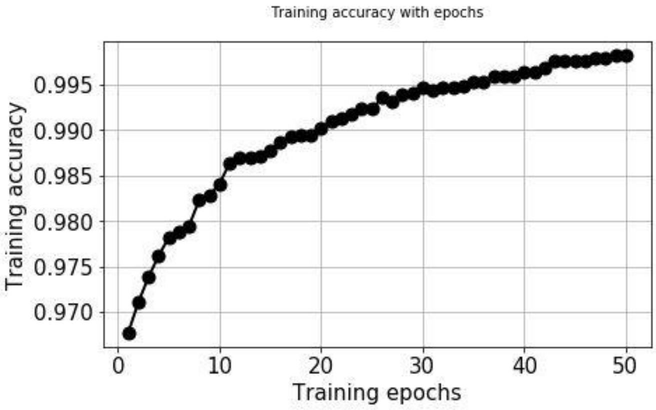


Figure 4: Model 1 Training Accuracy on 50 epoch

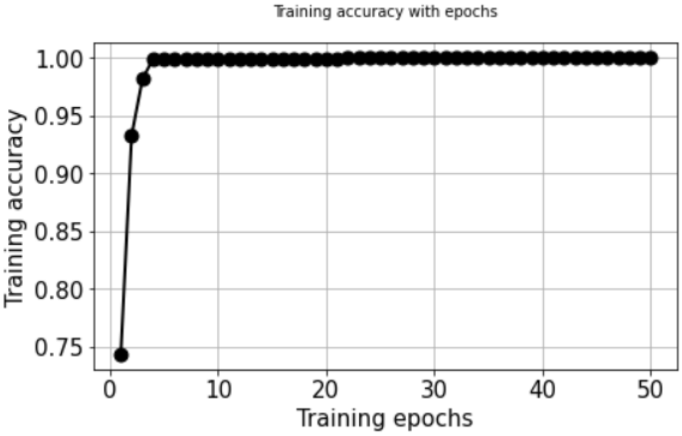
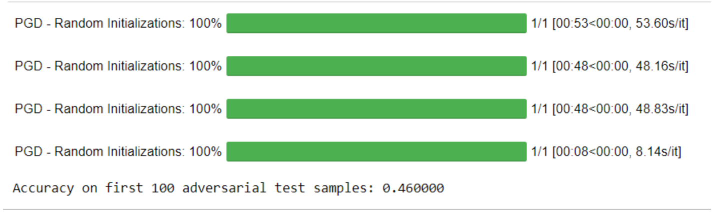


Figure 5: Model 2 training Accuracy



```
PGD - Random Initializations: 100% 1/1 [00:53<00:00, 53.60s/it]
PGD - Random Initializations: 100% 1/1 [00:48<00:00, 48.16s/it]
PGD - Random Initializations: 100% 1/1 [00:48<00:00, 48.83s/it]
PGD - Random Initializations: 100% 1/1 [00:08<00:00, 8.14s/it]
Accuracy on first 100 adversarial test samples: 0.460000
```

Figure 6: Testing accuracy of the adversarial attack of the first 100 smaples

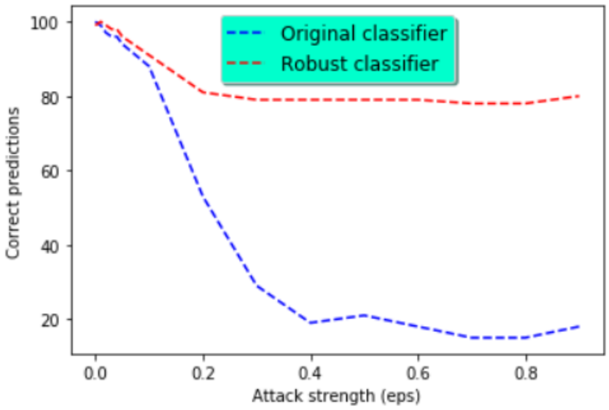


Figure 7: Model 1 vs Model 2 Comparison

V. Conclusion and Future Work

Adversarial attacks are explicitly intended to feat weaknesses in a given Deep Learning Algorithm. These weaknesses are recreated via preparing the learning algorithm under different attacks and approaches. This system presents a robust based model towards the resistance of adversary attacks on deep learning models. The system presents two models using convolutional neural network algorithm. This model was trained on a Modified National Institute of Standards and Technology dataset (MNIST). An adversary (evasion) attacks was generated to in other to fools this models to misclassify result, therefore, seeing the wrong input data to be the right one. This adversarial examples was generated using a state-of-the-art library in python. The generated adversarial examples was being executed on the test data, in which the first model fails in resisting the attack, while the second model, which is our robust model resisted the adversarial attack on a good number of accuracy when tested for the first 100 images. This work can further be extended by carrying out poisoning attack, which is an adversarial attack that is being executed on the training data, and also a detailed review of black-box and white-box attack should also be look into.

References

[1] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. Tygar, “Adversarial machine learning,” in Proceedings of the 4th ACM workshop on Security and artificial intelligence. ACM, 2011, pp. 43–58.

[2] Z. Wang, “Deep learning based intrusion detection with adversaries,” IEEE Access, vol. 6, pp. 38367–38384, 2018.

[3] I. Homoliak, M. Teknos, M. Ochoa, D. Breitenbacher, S. Hosseini, and P. Hanacek, “Improving network intrusion detection classifiers by non-payload-based exploit-independent obfuscations: An adversarial approach,” ICST Trans. Security Safety, 2018.

[4] B. Biggio, G. Fumera, and F. Roli, “Security evaluation of pattern classifiers under attack,” IEEE transactions on knowledge and data engineering, vol. 26, no. 4, pp. 984– 996, 2014.

[5] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, “ImageNet: A Large-scale Hierarchical Image Database”, In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009.

[6] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, “SQuAD: 100,000+ Questions for Machine Comprehension of Text”, In Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP) 2016.

[7] N. Carlini, D. A. Wagner, “Towards Evaluating the Robustness of Neural Networks”, In Proceedings of IEEE Symposium on Security and Privacy (S&P), pp.39-57, 2017.

[8] R. A. Khamis, M. O. Shafiq , A. Matrawy, “Investigating Resistance of Deep Learning-based IDS against Adversaries using min-max Optimization”, IEEE International Conference on Communications (ICC), pp.1-7, 2020.

[9] P. Pang, X. Zhang, J. Shouling, “AdvMind: Inferring Adversary Intent of Black-Box Attacks”, Proceeding of the 26th International Conference on Knowledge Discovery and Data Mining, pp.1899-1970, 2020.

[10] B. Kim, Y. E. Sagduyu, T. Erpek, K. Davaslioglu, S. Ulukus, “Adversarial Attacks with Multiple Receivers Against Deep Learning-Based Modulation Classifiers”, arXiv preprint arXiv:2007.16204, 2020.

[11] C. Kumar, R. Ryan, M. Shao, “Adversary for Social Good: Protecting Familial Privacy through Joint Adversarial Attacks”, Proceeding of the AAAI Conference on Artificial Intelligence 34(07), pp.11304-11311, 2020.

[12] Z. Wang, “Deep Learning Based Intrusion Detection with Adversaries”. Special section on challenges and opportunities of Big Data against Cyber Crime, vol.6, pp.38367-38384, 2018.

[13] A. S. Chivukula, W. Liu, “Adversarial Deep Learning Models with Multiple Adversaries”, Ieee Transactions On Knowledge And Data Engineering, 10(30), pp. 1-14, 2018.

[14] K. Yang, J. Liu, C. Zhang, Y. Fang, “Adversarial Examples against the Deep Learning Based Network Intrusion Detection Systems”, Milcom 2018 Track 5 - Big Data and Machine Learning, 559-564, 2018.

[15] G. Apruzzese, M. Colajanni, L. Ferretti and M. Marchetti, "Addressing Adversarial Attacks Against Security Systems Based on Machine Learning," *11th International Conference on Cyber Conflict (CyCon)*, pp. 1-18, 2019.