*THINK INDIA JOURNAL*

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

# A Comprehensive Review on The Different Types of The Attacks and Defense In Deep Learning

"**Mir Mohammad Yousuf[1], Sayantan Kar[2],Mamoon Rashid[3],Bisma Majid[4]**

[1, 3]Assistant Professor, "School of Computer Science and Engineering, Lovely Professional University, Punjab, India.

[2] Research Scholar, School of Computer Science and Engineering, Lovely Professional University, Punjab, India.

[4] Research Scholar, School of Electronics and Communication Engineering, N.I.T Srinagar, J&K, India."

***Abstract*-** In the growing phase of science and technology, it is very difficult for every developer to keep the information secret for everyone. Attacker break these information cycle through any loopholes and every single bit can be transferred from a medium which can be wired or wireless and in that medium loopholes can be created and attackers enter to these loopholes and access the secret information. These types of attacks occurred in the year May 2017 where the attacker can block the hard drive of the affected PC and demand high money which is about 500000 INR.It would provide a time span of 3 intervals and is increased by 50% after each interval is over.At the end of the third interval, the data in the PC is deleted automatically. The demandedmoney wasin the form of bitcoins and such types of organizations do not share any information about their customers according to their policy. There cannot be any trace of money in which account it will be transferred. These attacks are a kind of deep learning and this paper discusses the various types of such attacks and their defense in the technology of deep learning. The pape does a comprehensive study on each type of attacks and gives a summary of the defense as well.

***Keywords**:* Deep learning, attacks of deep learning, different defensive techniques in deep learning

## I. INTRODUCTION

Machine Learning provides some tasks through Deep Learning (DL) from the cyberspace to the communication world. These tasks like prediction and classification. Machine Learning algorithms like online classifiers which may receive inputs in the training and testing phase [1][2] of the machine. Then this becomes a threat of any kind of attack. Attacks in machine learning are very common to the Adversarial world. Attacks are classified into three categories viz,Exploratory Attack in which the attack works on the operation of algorithm in machine learning. As we say the learning boundaries of machine learning[3][7].The second category is the evasion attack which works to make the machine learning algorithm fool and try to prove the machine wrong. As we say about the spam emails, fooling the trained data of

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

the spam mail and try to except the spam mail as a legitimate mail[8][9]. The third is the causative attack whichis a type of attack in which it provides the wrong data to the ML algorithm. Based on these attacks a common attack will be followed in an attack and vindictive sample that is design to confuse a deployed machine learning model [3].

**Figure 1: An example of original input data (left) and its corresponding adversarial sample (right). The added noise is visually hard to see but can mislead the victim DL model [3].**

In this paper, an approach that should be taken is the black-box exploratory attack. In this process,the classifier is unknown"The samplesare sent to a classifier (online)T by an adversary and the returned_label of each sample s as T (s) is "observed and uses all these pairs (s, T (s)) to train a DL classifier ˆ T that is functionally equivalent to T by the adversary"""..

"Utilizing the above classifier ^T, the adversary has certain avoidable steps which should be followed for the evasion attack. In this attack, some samples through ˆ T are run by the adversary.The adversary uses those samples and feeds them to the classifier T in the enemy assault if their DL scores are bound to the chosen locale for class j and are near the choice district for class i.[14]. The causative attack has certain avoidable steps that should be followed.

a) In this attack, some samples through ˆ T are run by the adversary.

b) On the off chance that DL scores_are kept to the decision_region for class I and are far away from the choice area for class j.

c) The labels_of all those samples from_class I to class j are changed by the adversary and are sent to the classifier T.

These steps should be avoidable by the defense mechanism which is more reliable_inference of T_as ^T.[8] "

## II. PROBLEM OF CLASSIFICATION

"The main problem of classification is to classify one sample of data to the sample class C. This sample set is described in the form of sets Fs = {fm s} and belongs to a class C(s) ∈ C. In this context we should consider the following errors for the given set of test data with respect to the ground reality.

1) The Error_rate on the class i is given_by ei(T) = j=i j∈C eij (T) ni where ni is the number of samples with C(s) = i and eij(T ) is the number of samples with C(s) = i and T (s) = j.

2)TheOverall error rate is given by e(T) = i ∈C j=i j∈C eij (T) i∈C ni [1]3][4][5]

Text Classification: Earlier people use software to classify the text named as "Reuters-21578" from the two documents and for accessing these documents two types of

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

classes are predefined classes as “earn” and “acq”. Learning a particular machine there isa need of 4234 documents for machine learning purposes and need of 1718 documents for testing purposes. In this classification, we need Naive Bayes (NB) and implemented with a toolbox kit named Use the "Insert Citation" button to add citations to this document."
“Natural language Toolkit (NLTK) [2].Image Classification: Image can be of any types but for the convince of the machine we use flower images (like rose, jasmine, daisy). Learning for the machine we need 638 images and for testing, we can use 636 images.[6] ”

## III. EXPLORATORY ATTACK

Train a machine is a wonderful thing but the machine cannot be fulfilled without a classifier. To do more effort on the train a machine, rather than put the effort to train a classifier which is a replicate of a machine. The classifier acts as the heart of a machine. Attacks on classifiersare considered as an attack on the machine. There are some steps of attacks which can ruin the goal of a particular machine[5][8]

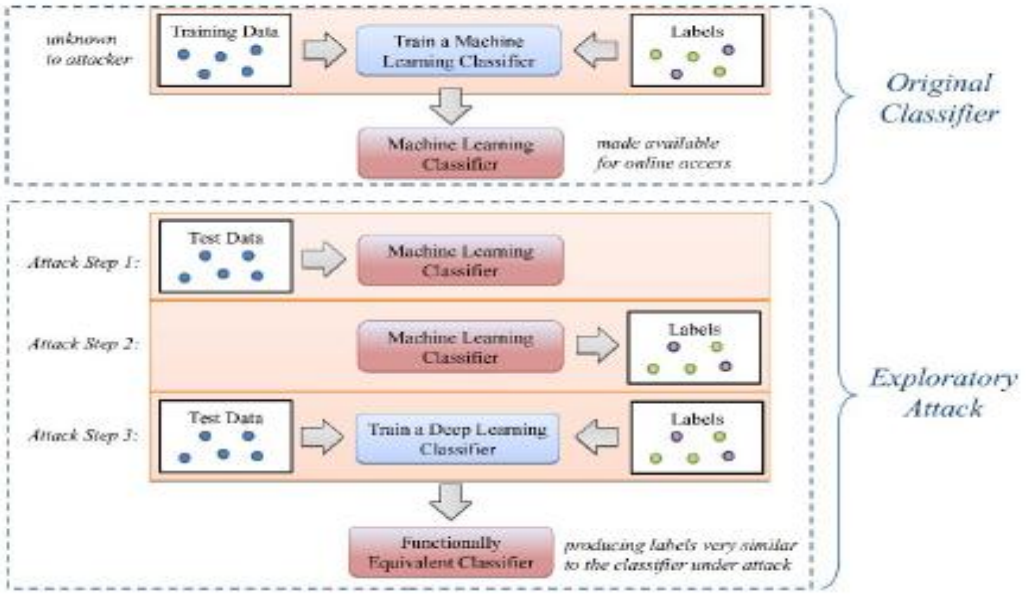


**Figure 2: Steps of Exploratory Attack [4]**

From the above steps, it was cleared that theinformation regardingthe originalclassifier T is unavailable tothe adversary. It can easily say that the information of other sample data is not available. The adversary can query some of the sample data and collects its label which is given by T. It was easily explained with an example, suppose the adversary performs N queries for sample a1,a2,.....an and obtain the results as $T(a_1)$, $T(a_2)$, ..... $T(a_n)$. Then the adversary builds trainingand test data by spilling the N sample into the two datasets.

**a) <u>Exploratory Attack on Test Classification</u>**: Deep Learning and the Feedforward network used as ^T thathas been developed by Naive Bayes classifier and the_difference between the T and the ^T is 2.10% which is obtained by a researcher. The optimized threshold results of

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

ˆθ of deep learning classifieri^T to reduce the results of the originalclassifier of Naive Bayes of T will be 0.55.[3][5][7]

**b) Exploratory Attack on Image Classification**: Deep learning and the feed-forward network is used as ^T to reduce the classifier T which can be built by another DL classifier. The difference between T and ^T is 9.12% obtained by a researcher. The optimized threshold value of ˆθ in DL classifier ˆ T to reduce the original DL classifier results of T is 0.6 [4][5].

## IV. EVASION ATTACK

This attack is tested and launchedupon a trained classifier by providing it with an input which is test data and the result is obtained in an incorrect output which is labels. This attack is tested upon on the three steps. These are listed below:
1) Applyithe classifieriˆ T on the set of thetest samples to_obtain a DL score ˆr(s) on_each sample s.
2) Find the "attack region" of scores, Aˆij, to cause misclassification_from class i to class j.
3) Feed the samples s with the scores r(s) ∈ Aˆij to the classifier T under the attack.

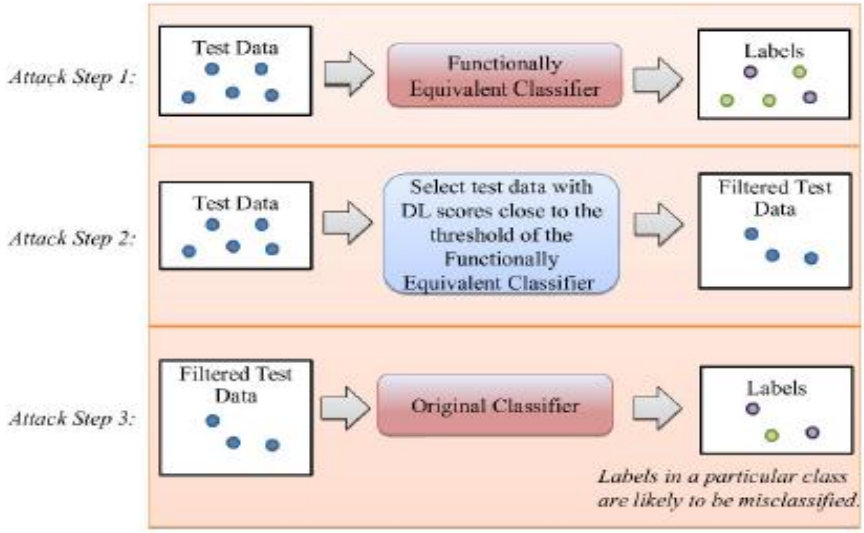


**Figure 3: Steps of Evasion Attack [5].**

This attack isican beclassified under certain conditions as per the situation but there are two common conditions which is found under every situation.
Condition 1: A sample s_with r(s) ∈ Aˆij is to be classified in class j by ^T which means ˆ Aij.
Condition 2: There is a_chance that s is belongs to class i which means that this condition interprets as a DL scores r(s) being "close" to ˆRi.
**a)Evasion Attack on Text Classification**: The adversary's main goal is to select sample form the class 1 without the knowledge about the ground reality and it is not classified as class 2 by the original classifier._For obtained the good results, we assume that = ˆθ /2 and their target attack_region is Aˆ12 is ˆθ /2 ˆθ.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

**Table 1: Error rate of the text classification**

| Attack_region of Scores | Error_rate |
|---|---|
| [0, 0.275] | 5.15% |
| [0.275, 0.55] | 57.14% |
| [0.55, 0.775] | 6.67% |
| [0.775, 1] | 0.82% |

**b) EvasioniAttack on Image Classification**: For the numerical results on the image classification, it's assumed to beˆθ/ 2 and their target attack region is Aˆ12 is ˆθ/ 2, ˆθ]. When its compared with the error probabilities with the image values [0, 0.3], [0.6, 0.8], and [0.8, 1] respectively. Obtained results are verified through the error.

**Table 2: Error rate of the image classification**

| Attackiregion of Scores | Error_rate |
|---|---|
| [0, 0.3] | 13.57% |
| [0.3, 0.6] | 29.17% |
| [0.6, 0.8] | 20.0% |
| [0.8, 1] | 4.32% |

## V. ADVERSARIAL EXAMPLES

To analyze the systematic approach for generating the adversarial example and to analyze this example we need to classify the three model approaches of adversarial example

A) Threat Model.

B) PerturbationiModel.

C) Benchmark_Model.

### A) Threat_Model

Threat Model is based on a certain condition, scenario, assumptions, quality requirements etc. We can classify the threat model under four aspects of attacks which are adversarial_falsification., adversary's_knowledge, adversarial_specificity. and attack frequency.

The adversarial classification is further classified into two types. These can be false-positive attack and false-negative attack Afalse positive attack isa negative effect over the positive attack is created by this sort of the assault and is otherwise called sort I error. In a malware discovery task, programming being delegated a malware is a bogus positive.It can be an ill-disposed picture unrecognizable to humans in a picture characterization task, while DNNs foresee it to a class with a high certainty score [2][3][5]. A False Negative AttackIt creates a positive effect on the negative attack and is otherwise called type II error. In a malware recognition task, a bogus negative can be the condition that a malware (typically considered as positive) can't be identified by the prepared model[1]. The_false negative assault is

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

additionally called ML avoidance. This error appears in most ill-disposed pictures, where a_human_can perceive the picture, however, the neural systems can't recognize it[2][5][13].

**2) Adversary Knowledge**: Adversary classification can be further classified into 2 types. Viz, White box testing and Black box testing[2][3][8].In white-box testing, we can realize that the developer/researcher can think about everything, i.e trained neural organize models, including preparing information, model designs, hyperparameters, quantities of layers, enactment capacities, and model loads. The adversarial model is determined by the model gradient.[9][12].The Black Box testing is testing where we can know that the developer/researcher doesn't know about everything of the machine. Thisassumption is commonly for attacking_online services like online ML services, Microsoft Azure, online cloud services etc. [9][11][13]

**3) Adversarial Specificity:** It is further classified into 2 types viz, Targeted Attacks and Non-Targeted attacks. Targeted Attacks are specially targeted by the multi-class classification problem. An adversary fool system can predict also the targeted image[1][6][8]. Non-targeted Attacks do not specify the neural network_output. The adversarial example can be a random choice for the original one.[2]

**4) Attack Frequency:** It is further classified into two types viz, one time Attack and Multi-time attack. One time attackcan be done one time on the adversary example to get an optimized output.[6][9]. The multi-time attack is done multiple times on the adversarial example to update the optimized output. These attacks can be called an Iterative attack. It is considered as a good attack as compared to the one-time attack process.[2][4][7]

**B)Perturbation:** It is considered as a fundamental premise of the adversarial example and designed to close the original sample[4]. Some researcher placed their viewpoints that it affectsthe deep_learning model when compared with the human. They said that none of the deep learning models is compared with the human. If a human is considered a machine then there is no deep learning model that can be compared_with humans. It is classified into 3 types. These are listed below viz,Perturbation_Scope,Perturbation_Limitation. and Perturbation_Measurements. The Perturbation Scopecan be classified into 2 types viz,Individual attack in which the user can get different output every time for each clean input. It can be applied to the single dataset. [8][9]. TheUniversal attack in which the user cannot get different outputs for any clean input. It is applied to the whole dataset. [10][12]. The second type of perturbation is perturbation scope which is further classified into 2 types viz,Optimized Perturbation set,which can be the goal of the optimized problem. In this phase, it can minimize the perturbation problem, so that human cannot recognize it.[5][13]. The second type is the Constraint Perturbation set. In this phase, it can act as a constraint of the optimized perturbation problem where the input can be small.[5][13]

The third type of perturbation is Perturbation Measurements. It can be measured under two conditions. These conditions are listed below:

a)The Euclidean distance between the example of the advertised image and the original image can move the maximum change for all pixels in the advertised example. It is denoted in the mathematical form.[9]

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

$$\|x\|_p = \left(\sum_{i=1}^{n} \|x_i\|^p\right)^{\frac{1}{p}}.$$

b) Psychometric_perceptual adversarial similarity score (PASS) is a new metric and is consistent_with human perception.

**C) Benchmark**: The benchmark is classified into 2 types of model viz, dataset model and the victim model.

a) Dataset Model: "MNIST", "CIFAR-10", and "Image Net" are the three most widely used image classification data sets to evaluate adversarial attacks.

b)Victim Model:Various well-known DL models, such as "LeNet", "VGG", "Alex Net", "Google Net"," CaffeNet", and "ResNet" are usually attack by the adversaries [3][4][5].

## VI. ATTACKS IN MACHINE

Attacks can be done to the machine where it can be targeted to the cognitive radio. Radio generally hasa simple transmission medium to communicate and it is easy for the attackers to attack such medium where it is easy to attack and the effected rate should be high. General operation model for the cognitive radio will be

a) Transmission Operation: It is the basic operation for the cognitive radio and it can add background noise where the channel needs T and A do not transmit. In this state, two modes can be possible either idle or busy.[2][7][8]
b) Receiver Mode: In this mode, the transmission is successful where the threshold is smaller than the channel. In this mode, feedback is sent back. [2][3][9]
c) Attacker Mode: In this mode, it checks for the successful transmission from the channel. It depends on the feedback [1].

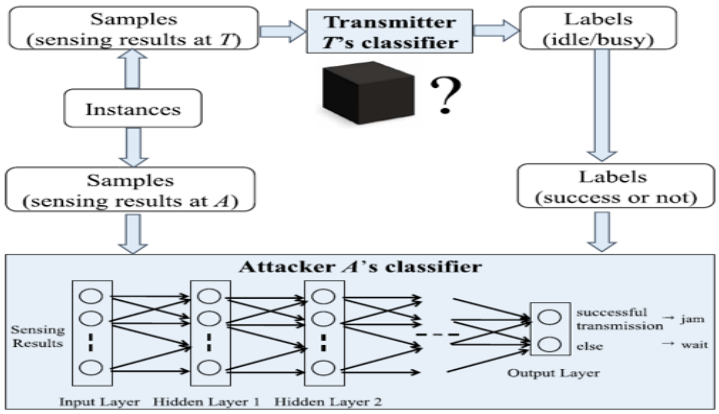


**Figure 4: Systems model for attackers learning**

Cases of the cognitive radio depend on these 4 cases:

1) The channel being idle and T is transmitting [1].
2) The channel being busy and T is not transmitting [1].
3) The channel being idle and T is not transmitting [1]

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

4) The channel being busy and T is transmitting [1].

**Defense against evasion attack along with exploratory attack**

It is difficult to defend this type of attack along with an exploratory attack and it is done for the 2 types of classification. This type of attack is shown in the below-mentioned figure where the graph shows the correlation between the inferred classifiers Error and the perturbation. The figure shoes the defense against the exploratory attack on the text classification.

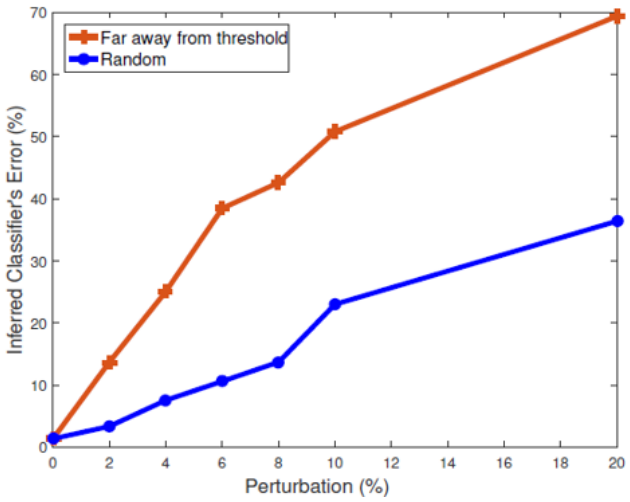


**Figure 5:Defense against the exploratory attack on the text classification**

This is a type of exploratory defense attack with text classification where we can see that the deflection comes in the statistical data.

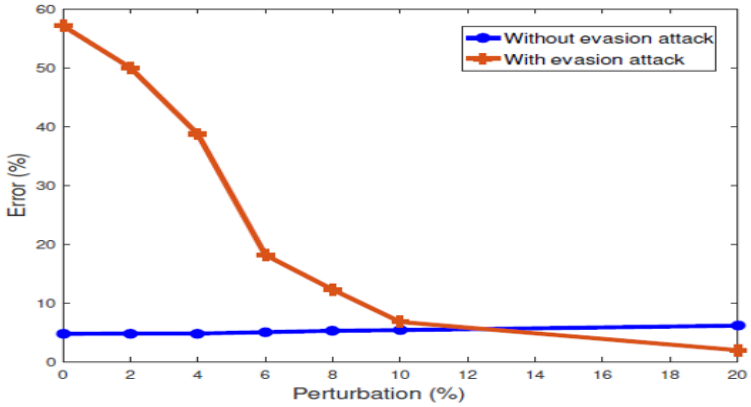


**Figure 6: Defense against the evasion attack on the text classification**

This is a type of evasion attack with text classification where the data is high to low it can determine the statistical data is not reliable along with this type of attack [1].

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

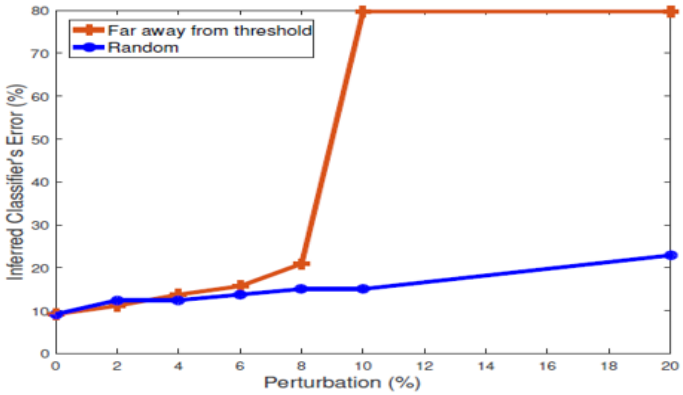


**Figure 7:Defense against the exploratory attack on the image classification**

In this type of attack where the data is gradually increasing at a certainpoint of the graph and it is the opposite of the text classification.

## VII. DEFENSE ALONG WITH A NEURAL NETWORK

A neural network is based on a feed-forward network where there is a forward of data through one or more network layers. In the neural network, we need 3 layers for feed forwarding

- Input Layer: This layer deals with the input values in the network. [2][3]
- Hidden Layer: This layer deals the output value from the input layer and it should be taken as the input for the hidden layer[2][3]
- Output Layer: This layer consists to take the output of the hidden layer as the input of the output layer [2][3].

To make a defensive approach it can we make through the distillation defense approach, weight defense, and the ensemble defense. In a defensive approach, we can analyze the performance of the distillation defensive**.** This defensive technique is based on the ROC curves of the DNN system where T=10 for the range of hidden layers where it can receive the input from the input layer. The F=0.001 is the default value of this layer.[2][5]. In Weight,the Defense approach can be proposed from the inspiration of the previous defensive approach. This approach can be based on the weights of the classificationof the image of the attacks [5][7].In the Ensemble Defense approach where this approach is especially based on the dataset. Where the dataset is a taken special of any classifier [4][8].

## IX. CONCLUSION

This attack and defensive approach in deep learning can determine the various defensive approach for each of the attacks and it is not a limited defensive approach. The defensive approach of these two techniques is very popular among all other defensive techniques.From that approach, we can say that this attack in deep learning is easier than the defense in deep learning. It is agreed by many researchers that results should vary from researcher to researcher. It is not a static outcome of the same example. It is difficult to say that which is results is better than all the defensive approach.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

## X. REFERENCES

[1] Rouhani, B. D., Samragh, M., Javaheripi, M., Javidi, T., & Koushanfar, F. (2018, November). Assured deep learning: practical defense against adversarial attacks. In *2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)* (pp. 1-4). IEEE.

[2] Shi, Y., & Sagduyu, Y. E. (2017, October). Evasion and causative attacks with adversarial deep learning. In *MILCOM 2017-2017 IEEE Military Communications Conference (MILCOM)* (pp. 243-248). IEEE.

[3] Shi, Y., Sagduyu, Y. E., Erpek, T., Davaslioglu, K., Lu, Z., & Li, J. H. (2018, May). Adversarial deep learning for cognitive radio security: Jamming attack and defense strategies. In *2018 IEEE International Conference on Communications Workshops (ICC Workshops)* (pp. 1-6). IEEE.

[4] Stokes, J. W., Wang, D., Marinescu, M., Marino, M., & Bussone, B. (2017). Attack and defense of dynamic analysis-based, adversarial neural malware classification models. *arXiv preprint arXiv:1712.05919*.

[5] Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*.

[6] Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I., & Tygar, J. D. (2011, October). Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence* (pp. 43-58). ACM.

[7] Shi, Y., Sagduyu, Y., & Grushin, A. (2017, April). How to steal a machine learning classifier with deep learning. In *2017 IEEE International Symposium on Technologies for Homeland Security (HST)* (pp. 1-5). IEEE.

[8] Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., ... & Roli, F. (2013, September). Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 387-402). Springer, Berlin, Heidelberg.

[9] Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*.

[10] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016, March). The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)* (pp. 372-387). IEEE.

[11] Pi, L., Lu, Z., Sagduyu, Y., & Chen, S. (2016, December). Defending active learning against adversarial inputs in automated document classification. In *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)* (pp. 257-261). IEEE.

[12] Lewis, D. D. (1997). Reuters-21578, distribution 1.0 http://www. daviddlewis. com/resources/testcollections/reuters21578.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

[13] Shi, Y., & Sagduyu, Y. E. (2017, October). Evasion and causative attacks with adversarial deep learning. In *MILCOM 2017-2017 IEEE Military Communications Conference (MILCOM)* (pp. 243-248). IEEE.

[14] Pathak, A. R., Pandey, M., & Rautaray, S. (2018). Application of deep learning for object detection. *Procedia computer science*, *132*, 1706-1717.

[15] Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction apis. In *25th {USENIX} Security Symposium ({USENIX} Security 16)* (pp. 601-618).