Aishwarya Zingade, 2023, 11:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

A review on Adversarial Attacks on Deep Neural Networks in Image Classification

Aishwarya Zingade, Nishchay Nilekani, Mohammed Rafi
Department of Computer Science, University B.D.T. college of Engineering

Abstract- Deep neural networks can be abbreviated as DNN. They are the core component of the many machine learning algorithms. They have recently become so popular and successful due to the introduction of artificial intelligence, which is shortly called as the AI. Deep learning have got the success in machine learning tasks in different domains. In recent years we can see that the DNN models are more prone to vulnerabilities. So it is important to study and research them and take the comprehensive steps to find the solutions. In this survey paper we discuss about the adversarial attacks that effect the DNN and the counter measures against these.

Keywords- Model safety, Medical imaging, Adversarial attacks, Black-box attacks, White-box attacks

I. INTRODUCTION

Deep learning have provided the major breakthrough in problem solving in the field of machine learning and artificial intelligence. Due to this deep neural networks are efficiently used in the solving of scientific problems in an unpredictable way. They are efficiently being used in the reconstruction of brain circuits, prediction of structure activity of potential drug molecules, deep analysis on the mutations in DNA, analyzing the particle accelerator data.

DNN can be used in the process of natural image analysis tasks for example the image classification. DNN can be helpful in getting human-level-performance but not exactly like humans that can be nearer to the human performance level. DNN are used in the medical sector in the wider range. It is most powerful tool in the medical diagnosis, medical diagnosis in the sense medical image processing. The applications of DNN in medical fields are abundant to list some of their applications can be firstly diabetic retinopathy which means a complication of diabetes that affect the eyes, secondly cancer diagnosis, thirdly landmark localization.

DNN have superficial performance but even though they are more vulnerable to the to carefully crafted adversarial attacks such as the DNN can make wrong or incorrect decisions with higher level of confidence to some unknown input instances. Some input instances can fool the DNN which leads to the wrong assumptions of DNN system.

Deep neural networks perform a wide range of computer vision tasks with near accurate results. Szegedy et al [1] is the first person to discover the conspiring weakness of deep neural networks in the area of image classification. By his research he made confirm that even though they are precise in their working efficiency they are susceptible to adversarial attacks in the form of small perturbations to images that remain almost imperceptible human vision system. These kind of attacks can make the deep neural networks to predict the images falsely and cause the image classifier to completely change the prediction about the specified image. The worse condition is that the models will predict the image wrongly with more confidence. These wrong prediction of results by deep neural networks implications triggered many researches in the adversarial attacks and their protection for deep learning in general.

Deep neural networks are most widely used in the medical classification of image for medical diagnosis to assist the physicians and the doctors to accelerate the decision making skills in the clinical environment For example, DNN's are used to classify carcinoma based on the photographic images, referable diabetic retinopathy can be classified based on the optical coherence tomography which is shortly called as OCT in the medical terms which means the images of retina and pneumonia based on the chest x-ray images. Some analysis is made by the researches which can be named as the meta analysis confirms that the diagnosis or analysis of images made by deep neural networks is equivalent to the analysis made by the medical professionals.

Even though the deep neural networks shows high performance as that of the medical professionals their analytical ability is still debatable. High level of decisions are made by the DNNs and they will be based on the disease diagnosis. Complicated classifiers such as DNNs can impose a catastrophic harm to the well being of the community because they are hard to analyse. If we make it more precisely they are prone to vulnerabilities for examples which have adversarial affects such as input images that cause miss classifications by DNN and are generated by adding some perturbations to the real or original images that are been rightly diffrentiated by deep neural networks.

Investigations with more focus should be made on the vulnerability of deep neural networks. The studies which are made earlier only considered adversarial effects on deep neural networks which are input dependent that is an individual adversarial disturbance is used in a way that every single input image is classified wrongly. These kind of adversarial affects are very hard to find because they need high standards of computational levels.

In recent years many researchers made the research on the adversarial attacks on deep neural networks and they states that DNN models are vulnerable to the adversarial examples which can be defined as [2] "Adversarial examples are inputs to machine learning models that an invader on purpose has constructed to so that the model to make mistakes". In the domain of image classification more precisely the medical field these adversarial effects are intentionally created images which have no difference with the original images they look exactly same as the original images. These synthesized images which are the exact copy of original images will mislead the classifier and lead to the prediction of wrong outputs. In other application domains involving graphs, text or audio adversarial attacks also present and have the same patterns which attack the deep learning networks for instance perturbing only some of the edges can make the graph neural networks to predict wrong outputs and integrating typos or typing errors in paragraph or in a sentence or in a title can fool text classification or dialogue systems. Due to this errors in the right assumptions of the output has lead the appearance of adversarial attacks in existing applications and fields has made the researchers cautious against directly adopting DNNs in safety-critical-machine learning tasks.

II. METHODS

1. Datasets for Medical Image

We can consider the images in medical field such as lesion of skin and the images for carcinoma classification, OCT images for diabetic retinopathy differentiation and X-ray Images of chest for pneumonia classification.

In previous studies [3] skin lesion images were secured from the International Skin imaging Collaboration 2018 datasets and the images were diffrentiated into seven different classes: melanoma (MEL), basal cell carcinoma (BCC), melanocytic nevus (NV) actinic keratosis/Bowens diseasem, benign keratosis, dermatofibroma, and vascular lessions.

The OCT images and x-ray images of chest were diffrentiated into four classes: choroid neo vascularisation with neo vascular layer and connected sub-retinal fluid (CNV), diabetic masuclar

edema with retinal thickening-associated in raretinal fluid (DME), multiple drusen appearing in early duration of life related macular degeneration (DRUSEN) and normal retinal fluid/edema (NM).
The x-ray images of chest were differentiated into binary classes: no pneumonia (NORMAL) or bacterial or viral pneumonia (PNEUMONIA)

2. Background of Medical Image Analysis

We know the deep learning domain is successful in the current scenario. The DNN models are extensively used in medical field in medical imaging analysis.

The methods like diagnosis of any medical condition through the medical image classification adopt roughly the same pipeline for example [4] ophthalmology(Kaggle,2015),dermatology(ISIC,209. These pipelines have got the great success and they are synonymous to the computers vision object recognition even though it is criticized for its lack of transparency and accuracy. The individuals still face the problem to analyse the predictions made by the DNN which is the important aspect for clinical applications which needs higher amount of faith which can be forgotten due to the adversarial examples.

III. PRELIMINARIES

If we focus on the classification of medical image tasks using DNN ,for a K-class that is $k>=2$ classification problem and if the given datasets is $\{(x_i ,y_i)\}$ where $i=1$ to N and x_i belongs to R^d and y_i belongs to the set $\{1,\ldots\ldots k\}$ then the DNN with parameter theta predicts the class of input x_i:

$$h(x_i) = \arg\max_{k=1,\ldots,K} p_k(x_i , \theta), \quad (1)$$

$$p_k(x_i , \theta) = \exp(z_k(x_i , \theta))/ \exp(z_{k0} (x_i , \theta)), \quad (2)$$

Where (zk'(x,theta)) is the lo-gits output of the network. Pk(xi,theta) is the probability of xi.

The model performance parameters theta are updated using back-propagation to minimize the classification loss such as commonly used cross entropy

$$\text{loss}\ `\ (h, x) = 1 /N \ P\ N\ i\ -y_i \log p_{y_i} (x_i , \theta).$$

1. Datasets, DNN models and Classification tasks,

Here we are considering three successful applications of DNN for image classification. 1) diabetic retinopathy a eye disease using retinal fundos copy classification 2)detecting diseases of thorax from chest X-rays and doing their classification 3)a type of skin cancer called melanoma from the photographs of desmo scope and their classification.

2. Datasets

Here we use five datsets which are avilable publically for all the three classifications For these experiments we have two sets of data for each dataset they can be put as follows 1) for pre-training the DNN model we need a train dataset 2) for evaluating the DNN models and for testing the advarsarial attacks we need test sub set.

Table I Classes and Images in Each Sub set

Datasets	Classes	Train	Test AdvTrain AdvTest
Fundoscopy Chest X-ray Dermoscopy	2 2 2	75397 53219 18438	8515 2129 6706 1677 426 107
Chest X-ray-3 Chest X-ray-4	3 4	54769 57059	9980 10396

After conducting detection experiments the test data is partitioned into two parts further 1) Adv Train for learning adversarial detectors 2) Adv Test for evaluating the adversarial detectors

For classification task of diabetic retinopathy the Kaggle date set fundos copy is used which consists of nearly 80000 images of retina with higher amount of resolution. The images were classified under various conditions and each image is named as"NoDR"to"moderate/

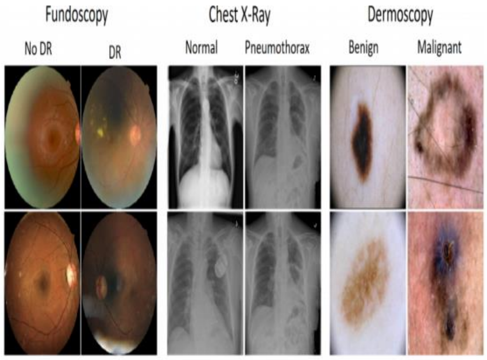


Figure: 1 Image examples for the classification of fundoscopy, chest x-ray, Dermoscopy[4]

mid/severe/prolifeartive.

For the classification task of thorax disease Chest x-ray database is used which consists of a huge number of images nearly 112,120 frontal-view X-ray Images. These images were also given different names on the basis of different conditions and each of the image had multiple labels. And the images were differentiated simply as "No finding" or pneumopthorax.

For the classification task of melanoma the images related to the melanoma wedre collected and classified on different conditions and they are named as "benigns" and "malignant" by considering the database of International skin Imaging Collaboration .

3. DNN Models

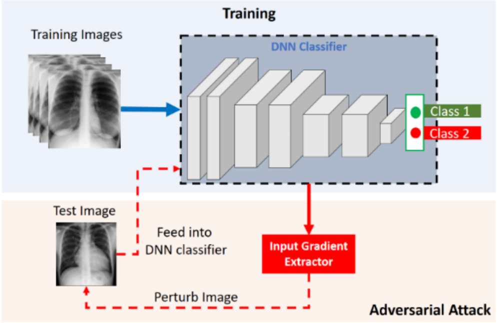


Figure: 2 The training DNNs pipeline , generation of adversial attacks

For all the datasets used in the experiment a model named Image Net pretrained ResNet-50 is used as the base network and its top layer is made of a new and dense layer which is made up of approximately 128 neurons, and following this layer another layer is present whose rate is 0.2 which is a dropout layer and there is a layer with K numbers of neurons.

4. Attack Results

The experiments conducting on medical images in Image Net compared to natural images is difficult. The difficulty of the underlying attack is checked and measured by least maximum perturbations and it is most important for the attacks to succeed.

IV. ADVERSARIAL ATTACKS

1. Fast Gradient Sign Method (FGSM)

By the research made by [5] Szegedy et al he confirmed that the adversarial training can be used to improve the robustness of the deep neural networks .After his research another researcher named Good Fellow et al designed or constructed a method to compute a perturbation effectively of a specified image by solving the problem given below.

$$p=e \text{ sign } (\nabla J (\theta, Ic, l`))$$

Here the ∇J is used to compute the gradient of cost function with the current value of model parameters. Here e is the scalar value and it limits the perturbation. The above equation can be depicted as the 'Fast gradient Sign Method' which is shortly named as FGSM.

The FGSM generates adversarial instances which can cause a huge harm to the deep neural networks because they harm corrupt the linearity of deep neural networks and their models in the three dimensional space which are earlierly considered as non-linear in nature. The sign function is applied to enlarge the loss of the magnitude.

2. Box Constrained L-BFGS

The researcher Szededy et al made the research and proved the existence of little perturbations in the images and these perturbed images can easily fool the models of deep learning and lead to mis classifications. Szedety et al described and solved the problem given below.

$$\min \rho \ ||\rho||2 \text{ s.t. } C(Ic + \rho) = l \ ; \ Ic + \rho \in [0, 1]m$$

In the above given equation Ic belongs to Rm and it indicates a vector clean image.The subscript c clearly indicates that the image is clean image.Here 'l' denotes the name of the image and C is the deep neural network classifier. For the given above equation the researchers tried to solve its non-trivial solution where 'l' is not similar to the original label Ic.

3. The One Pixel Attack

The one pixel attack can be defined as a case which is extreme in the sense of adversarial attacks and it happens only when a single pixel in an image is modified and lead to fool the classifier. A researcher named (1) Su et al during his research tested on the images by modifying a single pixel which lead to the fooling of three different network models. And he also noticed that theses three modules even after fooled by theses modified images gives predictions with a higher confidence level which is about 97.47%.Su et al used a concept named Differential Evolution for the computation of the adversarial examples that effects the deep neural network and led them to the wrong prediction.

For a clean image Ic they created approximately 400 vectors and all these vector contained x and y-coordinates and also the RGB (Red Green Blue) values of each vector. They randomized and modified the constituent element in the vectors so to get the corresponding children. These children are formed in such a way that the child will be competing in the sense of fitness with its own.

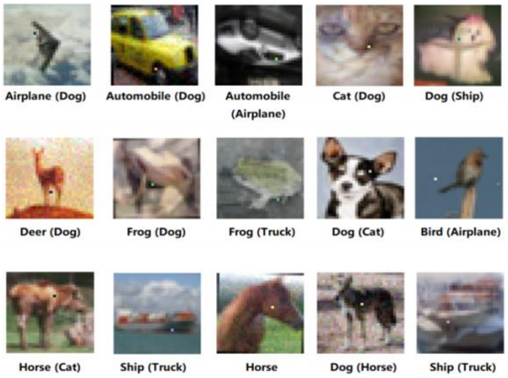


Figure: 3 illustration for adversarial attacks regarding to one pixel. The correct label of iamge is specified with the image and wrongly predicted name is mentioned in the square brackets [6].

4. Model Safety

It refers to the strength of the DNN against the adversarial examples. [6]

The concepts related to Model safety are:

Adversarial Examples

Here some specially crafted input data are designed to mislead the model into making incorrect prediction.

Strength

The robust model is less harmed to adversarial attacks. It's the ability of the model to maintain it's performance on both clean and clean and corrupted inputs.

Adversarial Training

It is the technique where the model is trained on both clean and adversarial data to improve its robustness against adversarial attack.

Transferability

The adversarial example of one models are effective for others. By Understanding transferability its will be easier to assessing the generalization of attacks.

White-Box Attacks

The attacker has the complete knowledge of the model such as parameters and architecture.

Black-Box Attack

The attacker has minimum or no knowledge about the interior of the system and tries to generate adversarial examples.

Defense Mechanism

These are strategies used to improve the strength of the system against adversarial attack. It may include training, input processing and use of defenses.

Certified Defenses

These provides the guarantees about models robustness.

Non Targeted Universal Adversarial Attack

After evaluating the [7] sensitivity of DNN models to untargeted UAPs. First, we will use the V3 model.

Giant. 1 describes an example of untargeted p=2 UAPs in opposition to Inception V3 models. The deception rate for the learning and test images increased rapidly with size and reached a high Rf, despite the low size. UAPs with magnitude=4% and Rf > 80% are obtained for skin lesions (Fig. 1a) and chest X-ray image.

Most images of skin lesions should have been mostly distinguished as AKIEC or DF (Fig. 1d); however, most OCT images were differentiated as CNV (Fig. 1e). For the chest X-ray dataset, the model mis predicted real names (Fig. 1f). High Rf at low magnitude and most dominant names were observed in the UAP case with p=∞ in contrast to the Inception V3 models for the medical image datasets.

Table 1 shows graph on medical image dataset which descibes Rf UAP versus DNN models [8]

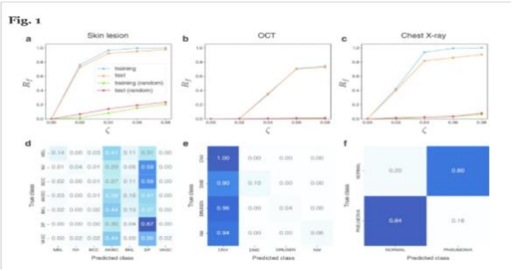
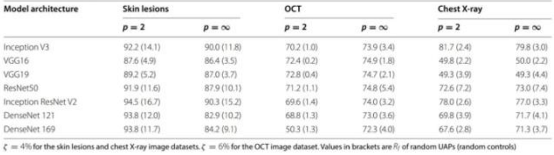


Table. 2 DNN models for test images of skin lesions, OCT, and chest X-ray image datasets and fooling rates for non-targeted UAP,s



Model architecture	Skin lesions		OCT		Chest X-ray	
	$p = 2$	$p = \infty$	$p = 2$	$p = \infty$	$p = 2$	$p = \infty$
Inception V3	92.2 (14.1)	90.0 (11.8)	70.2 (1.0)	73.9 (3.4)	81.7 (2.4)	79.8 (3.0)
VGG16	87.6 (4.9)	86.4 (3.5)	72.4 (0.2)	74.9 (1.8)	49.8 (2.2)	50.0 (2.2)
VGG19	89.2 (5.2)	87.0 (3.7)	72.8 (0.4)	74.7 (2.1)	49.3 (3.9)	49.3 (4.4)
ResNet50	91.9 (11.6)	87.9 (10.1)	71.2 (1.1)	74.8 (5.4)	72.6 (7.2)	73.0 (7.4)
Inception ResNet V2	94.5 (16.7)	90.3 (15.2)	69.6 (1.4)	74.0 (3.2)	78.0 (2.6)	77.0 (3.3)
DenseNet 121	93.8 (12.0)	82.9 (10.2)	68.8 (1.3)	73.0 (3.6)	69.8 (3.9)	71.7 (4.1)
DenseNet 169	93.8 (11.7)	84.2 (9.1)	50.3 (1.3)	72.3 (4.0)	67.6 (2.8)	71.3 (3.7)

$\zeta = 4\%$ for the skin lesions and chest X-ray image datasets. $\zeta = 6\%$ for the OCT image dataset. Values in brackets are R_f of random UAPs (random controls)

Targeted Universal Adversarial Attacks

A study (9) shows targeted UAP. DNNs are vulnerable not only to non-targeted UAPs, but also to targeted ones. Table 2 shows the targeted success rates of Rs UAP with p=2 against DNN models for test images in the medical data set. Targeted attacks on MEL and NV were considered for the skin lesion image dataset. Targeted attacks on CNV and NM were considered for the OCT image dataset. Targeted attacks on PNEUMONIA and NORMAL were considered for the chest X-ray dataset.

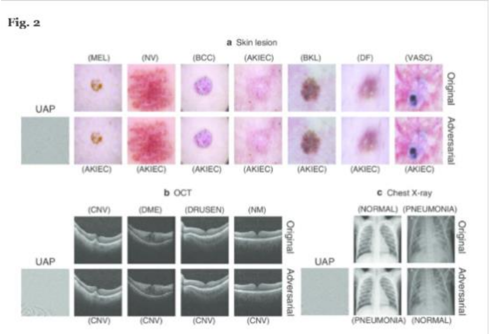
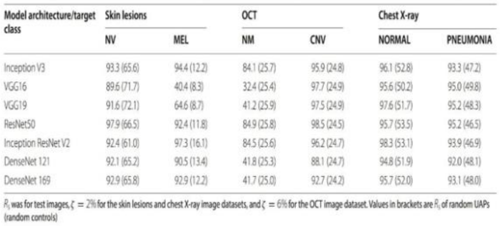


Table 3 targeted attack success rate R_s (%) of targeted UAPs with p=2 against various DNN models to each target class



Model architecture/target class	Skin lesions		OCT		Chest X-ray	
	NV	MEL	NM	CNV	NORMAL	PNEUMONIA
Inception V3	93.3 (65.6)	94.4 (12.2)	84.1 (25.7)	95.9 (24.8)	96.1 (52.8)	93.3 (47.2)
VGG16	89.6 (71.7)	40.4 (8.3)	32.4 (25.4)	97.7 (24.9)	95.6 (50.2)	95.0 (49.8)
VGG19	91.6 (72.1)	64.6 (8.7)	41.2 (25.9)	97.5 (24.9)	97.6 (51.7)	95.2 (48.3)
ResNet50	97.9 (66.5)	92.4 (11.8)	84.9 (25.8)	98.5 (24.5)	95.7 (53.5)	95.2 (46.5)
Inception ResNet V2	92.4 (61.0)	97.3 (16.1)	84.5 (25.6)	96.2 (24.7)	98.3 (53.1)	93.9 (46.9)
DenseNet 121	92.1 (65.2)	90.5 (13.4)	41.8 (25.3)	88.1 (24.7)	94.8 (51.9)	92.0 (48.1)
DenseNet 169	92.9 (65.8)	92.9 (12.2)	41.7 (25.0)	92.7 (24.2)	95.7 (52.0)	93.1 (48.0)

R_s was for test images, $\zeta = 2\%$ for the skin lesions and chest X-ray image datasets, and $\zeta = 6\%$ for the OCT image dataset. Values in brackets are R_s of random UAPs (random controls)

Black Box Attack and White Box Attack

In Black Box attacks, adversaries can generate adversarial examples without accessing the parameters of the target model, which makes them more threatening (10).

In 2017, a group of researchers proposed a new technique based on local search to construct a numerical approximation to the gradient of the network, which is then used to construct a small set of pixels in the image. They show how it can be used for image classification (10).

White box attacks are a type of security threat where adversaries have access to the parameter of the target model. There makes them more powerful than black box attack. In white-box attacks, adversarial can generate adversarial examples by knowing the target model's architecture and parameter (11).

The image data used in this study are digital chest X-Ray images of norm of different age groups, 2D slices of Computed Tomography (CT) images representing the norm and lung tuberculosis as well as colour histology images sampled from normal and cancerous tissue of thyroid glands and the ovary. An additional benchmarking image dataset consisted of 6 classes of histological images stained with different histo chemical markers. A detailed description of image data and the number of images in each class is given in table 1

The black-box settings in this study require complete knowledge of the training image dataset of the network to be attacked but we don't know the architecture which we are going to attack. The entire attack channel is based on the Projected Gradient Descent white-box algorithm.

- Train a "base" CNN on which to generate hostile (i.e., attacking) images using the training dataset of the target network that is attacked.
- Perform PGD attack on the trained network for each image in particular testing dataset to obtain a set of adversarial examples.
- Pass both the testing dataset and its adversarial examples to the target network to assess the rate of successful attacks

To make result clear we will use a single testing data-set for each of the training dataset. We have performed a large number of experiments on based on each image data-set for 5 network architecture: InceptionV3, ResNet50, DenseNet121, Mobil E-net, xception.

To imitate black-box following Black-Box pipeline were used as:

- Select one network as the target.
- Perform the defined black-box algorithm with each network left as an attacking ones separately.
- Carry out steps 1-2 with subsequent selection of every network as a target one.

Table 4 Descriptions about datasets and classification task configurations

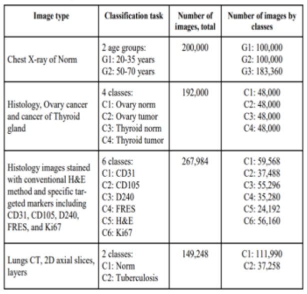


Image type	Classification task	Number of images, total	Number of images by classes
Chest X-ray of Norm	2 age groups: G1: 20-35 years G2: 50-70 years	200,000	G1: 100,000 G2: 100,000 G3: 183,360
Histology, Ovary cancer and cancer of Thyroid gland	4 classes: C1: Ovary norm C2: Ovary tumor C3: Thyroid norm C4: Thyroid tumor	192,000	C1: 48,000 C2: 48,000 C3: 48,000 C4: 48,000
Histology images stained with conventional H&E method and specific targeted markers including CD31, CD105, D240, FRES, and Ki67	6 classes: C1: CD31 C2: CD105 C3: D240 C4: FRES C5: H&E C6: Ki67	267,984	C1: 59,568 C2: 37,488 C3: 55,296 C4: 35,280 C5: 24,192 C6: 56,160
Lungs CT, 2D axial slices, layers	2 classes: C1: Norm C2: Tuberculosis	149,248	C1: 111,990 C2: 37,258

V. CONCLUSION

It pose a significant challenge to the robustness of these system. It is effective for researchers to develop defense mechanism.

REFERENCES

1. Adversarial Deep Learning: A Survey on Adversarial Attacks and Defense Mechanisms on Image Classification Samer Y. Khamaieshi1,, derek Bagagem2 , Abdullah Al-Alaj3, Mathew Mancino4, Hakam W. Almorari5
2. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review by Han Xu1, Yao Ma2, Hao-Chen Liu3, Debayan Deb4,Hui Liu5 ,Ji-Liang Tang6, Anil K. Jain 7
3. Universal adversarial attacks on deep neural networks for medical image classifcation Hokuto Hirano, Akinori Minagi and Kazuhiro Takemoto
4. Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems Xingjun Ma1,Yuhao Niu2, Lin Gu 3 ,Yisen Wang4 Yitian Zhao5 , James Bailey 6, Feng Lu7
5. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey by Naveed Akhtar and Ajmal Mian
6. International journal of Automation and Computing, By Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, DOCS Michigan State University,USA
7. Hokuto Hirano,Akinori Minagi and Kazuhiro Takemoto. BMC Me Imagiing
8. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2012;2017(42):60–88.
9. Hirano H, Takemoto K. Simple iterative method for generating targeted universal adversarial perturbations. Algorithms. 2020;13:268.
10. Black-Box Adversarial Attacks on Deep Neural Networks: A Survey
11. DI-AA: an Interpretable White-Box attack for Fooling deep Neural network by Yixiang Wang, Jiqiang Liu,