Metaverse (2023) Volume 4 Issue 1, *17 pages*.
doi: 10.54517/m.v4i1.2156

Asia Pacific Academy of Science Pte. Ltd.

REVIEW ARTICLE

A survey on adversarial attack and defense of deep learning models for medical image recognition

Jipeng Hu, Jinyu Wen, Meie Fang*

Metaverse Institute, School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 511442, China. E-mail: fme@gzhu.edu.cn

ABSTRACT

The advancement of hardware and computing power has enabled deep learning to be used in a variety of fields, particularly in AI medical applications in intelligent medicine and medical metaverse. Deep learning models are aiding in many clinical medical image analysis tasks, including fusion, registration, detection, classification and segmentation. In recent years, many deep learning-based approaches have been developed for medical image recognition, including classification and segmentation. However, these models are susceptible to adversarial samples, posing a threat to their real world application and making them unsuitable for clinical use. This paper provides an overview of adversarial attack strategies that have been proposed against medical image models and the defense methods used to protect them. We assessed the advantages and disadvantages of these strategies and compared their efficiency. We then examined the existing state and restrictions of research methods involving the adversarial attack and defense of deep learning models for medical image recognition. Additionally, several suggestions were given on how to enhance the robustness of medical image deep learning models in intelligent medicine and medical metaverse.

Keywords: adversarial samples; attack; defense; deep learning; medical image

The medical metaverse and intelligent medicine will solve the problems of improving the uneven distribution of medical resources and the inefficient treatment. However, as the most important medical image recognition model in the medical metaverse and intelligent medicine, its safety is crucial. If its safety cannot be guaranteed, it will greatly hinder the development of the medical metaverse and the promotion of intelligent medicine. Ma et al.[1] contributed to the knowledge of contrastive examples in medical imaging, raising questions about the applicability of deep learning-based systems for classifying medical images. Li et al.[2] proposed an unsupervised learning technique for identifying malicious attacks on medical images. This approach does not require labeled data for detection. Paul et al.[3] introduced a method for defending adversarial attacks on lung nodule malignancy prediction. Park et al.[2] present a distinctive and successful safeguard approach for segmentation models against adversarial attacks in medical imaging. Park et al.[4] had developed a novel and highly effective security architecture for medical imaging segmentation models in order to reduce vulnerability to adversarial attacks. Li et al.[5] presented a robust Artificial Intelligence (AI)

ARTICLE INFO
Received: 17 March 2023 | Accepted: 29 March 2023 | Available online: 17 April 2023

CITATION
Hu J, Wen J, Fang M. A survey on adversarial attack and defense of deep learning models for medical image recognition. Metaverse 2023; 4(1): 17 pages. doi: 10.54517/m.v4i1.2156

framework for medical imaging. This framework was based on SSAT and UAD, and incorporates a novel metric for measuring adversarial risk. The proposed setup serves to reduce the repetition rate and improve protection from attacks. Minagi et al.[6] assessed if DNNs used for medical image classifications (i.e. skin cancer, referable diabetic retinopathy, and pneumonia) are vulnerable to universal adversarial perturbations (UAPs) when employing transfer learning. Apostolidis et al.[7] proposed digital watermarking as a possible solution for black-box adversarial attacks. They presented a convincing viewpoint on this. They introduced a novel adversarial attack, called watermarking attacks, and investigate the potential of utilizing digital watermarking as a protection against them. They also examined how this approach could be used to reduce the frequency. Apostolidis et al.[7] pointed out a major issue, as the heavy application of watermarks for safety reasons could potentially be a hazard for vision systems. Finlayson et al.[1] , Winter[8] , Desjardins et al.[9] , and Yao et al.[10] explored the potential of a new adversarial example defense system called Medical Aegis. Research indicates the necessity of continuing to look into safety concerns surrounding AI models in healthcare and creating defensive strategies against adversarial attacks[11] . Selvakkumar et al.[12] investigated the effects of adversarial attacks on smart healthcare systems and found that they can have a significant impact. To defend against model inversion attacks, Khowaja et al.[13] suggested using the proximal gradient split learning (PSGL) method. The risk of adversarial attacks prevents DNNs from being used for significant tasks such as diagnosis, however, this is likely to be mitigated due to the difficulty of obtaining medical images, which are often necessary for such attacks, as privacy and security needs to be maintained. Minagi et al.[6] found that medical deep neural networks (DNNs) using transfer learning are still susceptible to adversarial attacks even when natural images are not accessible. Rodriguez et al.[14] explore the impact of model complexity on adversarial scenarios. Jin et al.[10] proposed utilizing a data poisoning strategy from backdoor attack classification to increase the efficacy of FedGAN. Xu et al.[15] developed a Durable and No-Retraining Diagnostic Framework for Medical pretrained models that is resistant to adversarial sample.

1. Related works

Kos et al.[16] investigated an innovative approach to the issue of adversarial attacks on deep reinforcement learning policies. Although many existing adversarial attacks can only mislead a black-box model, they have a relatively low success rate. Dong et al.[17] offered a wide set of momentum-based iterative procedures to heighten adversarial attacks in order to reduce the rate of repetition. Recently, LaVAN and Adversarial Patch have been implemented, introducing a new challenge to deep learning security. These techniques introduced adversarial noise, focused at a specific region of an image, while leaving out important features. This leaded to high-frequency changes in that area, while the rest of the image remains untouched, Naseer et al.[18] created an approach for accurately estimating the adversarial noise's position in the gradient domain and transforming the high activation regions caused in the image domain, with minimal impact on the essential object for classification. Dai et al.[19] suggested a gradient-based search method to generate the adversarial noise patches. IDSGAN, a generative adversarial network framework, was proposed to create adversarial malicious traffic records, aiming to fool deep learning models and evade intrusion detection systems. This process was done by altering the combination of data. Lin et al.[20] Qiu et al.[21] strived to provide an overview of the most recent developments in deep learning concerning adversarial attack and defense strategies. Dong et al.[22] proposed a translation-invariant attack technique to produce transferable malicious samples capable of evading defense models. Laidlaw et al.[23] presented TextAttack, a Python framework for creating adversarial examples, augmenting data, and training models resilient to adversarial attacks. This framework was part of a novel class of threat models, known as functional adversarial attacks, which aim to reduce the rate of repetition and strengthen the fooling of machine learning models.

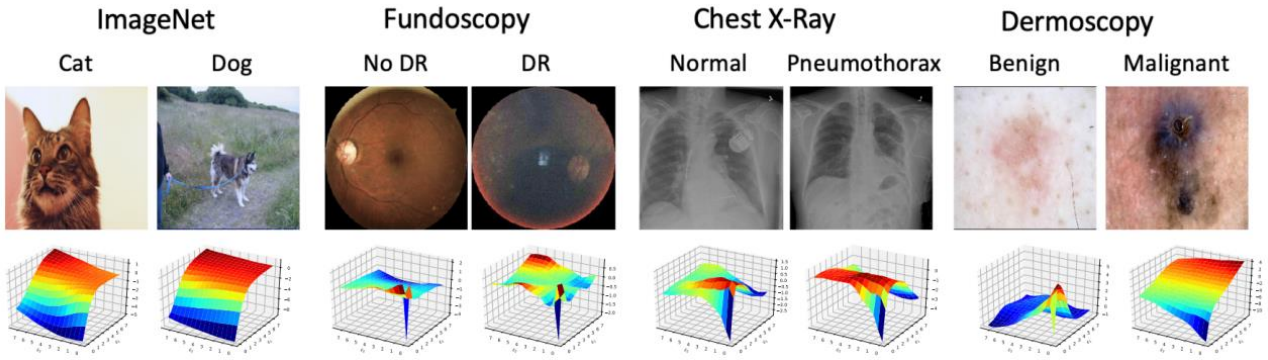

Figure 1. The bottom row of the scene shows the variation in losses depending on the corresponding input examples in the top row. The x and y-axes of the loss landscape graphs are denoted by ϵ_1 and ϵ_2, which indicate the magnitude of perturbations that have been injected into two adversarial directions and $^{\perp}$ respectively: = + ϵ_1 + $\epsilon_2{}^{\perp}$, where is the adversarial direction (sign of the input gradients) and $^{\perp}$ is the adversarial direction found from the surrogate models. The z-axis of the loss landscape, or the classification loss, is more pronounced when highly parameterized deep networks are used for medical images rather than natural images[1].

2. Attacks to medical image learning model

In recent times, deep learning has been widely used in healthcare. Studies have revealed its susceptiblity to adversarial sample attacks, similar to those targeting deep learning models for natural images.

2.1. Classification

Xing et al.[1] discovered that DNNs are more prone to adversarial samples than natural image models. DNNs have demonstrated their effectiveness in achieving near-human performance on various image analysis tasks, including image classification, object identification, image retrieval, and 3D analysis. Constructing adversarial samples on medical images is likely to be harder than on non-medical images because of several reasons. Medical images often feature intricate biological textures, creating a high number of regions that are sensitive to minor adversarial noises. Additionally, the utilization of outstanding DNNs developed for natural image processing on medical imaging tasks can result in an unpredictable interpretation. This complexity produces a steep loss landscape, making medical images highly vulnerable to such attacks. Attacks such as BIM (Basic Iterative Method), PGD (Project Gradient Descent) and CW[41] are particularly successful with minimal perturbations of $\epsilon < 1.0/255$. This makes attacking medical imaging simpler than attempting to compromise images from datasets like CIFAR-10 and ImageNet, which necessitate a much greater degree of distortion. In order to be successful with targeted attacks, it is usually necessary to have a perturbation of $\epsilon > 8.0/255$. The author increases ϵ_1 and ϵ_2 from 0 to 8.0/255, visualizing the classification loss curve for each combination in **Figure 1**.

Attention maps were also created using Grad-CAM to illustrate the contribution of the key areas in the image to the network output (**Figure 2**). The AUC of excellent detectors on medical images was usually lower than 80% in the face of certain attacks like FGSM and BIM, suggesting that adversarial examples can be easier to spot in this field. The author studied the efficacy of deep features and their quantized versions in detecting adversarial examples. Mini-batches of 100 were taken to extract the detection features. To help Visualize the contrast between adversarial and regular features, the author provided a graph in **Figure 3** of the two-dimensional representations of the deep features.

Figure 1 reveals that adversary-crafted features were nearly linearly distinguishable when the data had been through a non-linear transformation, which contrasted to natural images that can be difficult to divide even in the presence of a non-linear

transformation. **Figure 4** illustrated that the features of adversarial samples are very similar to those of normal samples. However, deep feature-based methods were limited in terms of providing protection[24] .

The author delved into why deep learning medical systems can produce incorrect results when confronted with adversarial examples, as well as the challenges that arise from creating and recog-nizing such adversary samples on medical images which are much more intricate than natural images (**Figure 5**) . This can encourage the development of more efficient defensive measures to enhance the resilience of healthcare systems to malicious attacks.

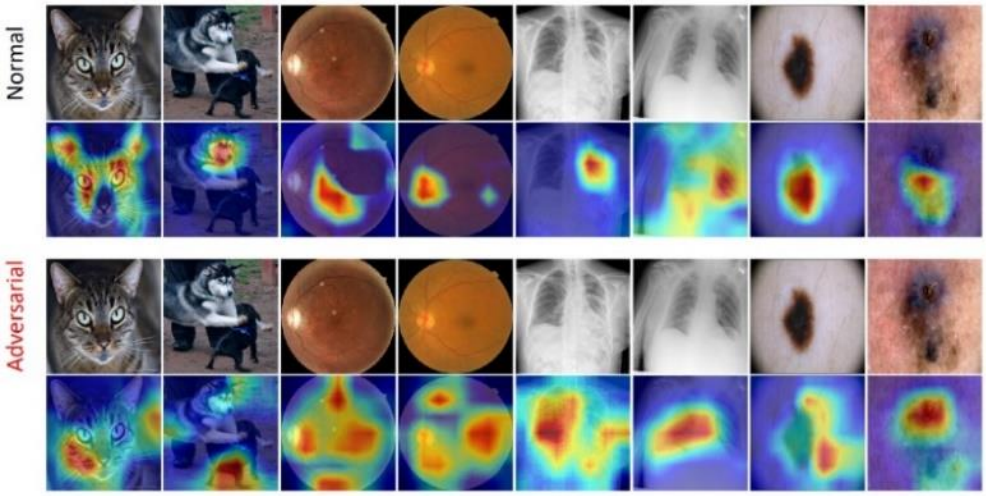

Figure 2. The network's focus on normal images (top row) is contrasted with its attention on adversarial images (bottom row), when using the Grad-CAM technique to calculate the attention maps[1].

Li et al.[2] put forward a detection approach for adversarial images that can effectively resist adversary samples on medical image classification models. This method employed features from the CNN classifier, allowing us to detect adversarial images that have been tampered with at the feature level. But when the author use the black-box setting, the substitute classifier used to create the adversarial image may be distinct from the original CNN classifier. Adding a detection system such as MGM to a white-box environment can have a significant impact on reducing repetition and rewriting data. It has also been found to provide more accurate classification results for hybrid clean and adversarial samples than if convolutional neural Networks classification model was tested on clean images alone, as the detection module eliminates all adversarial images. In order to reduce the rate of repetition, and to make a dramatic alteration to the text, it is necessary to ensure that the classification performance of the system remains uncompromised when using a clean dataset. This paper proposes a detection module for medical imaging classification systems that can accurately identify adversarial images based on high-level features from clean images. These features, which are typically located at the extremes of the distribution, are hard for a standard CNN classifier to recognize, but can be detected by the module. This way does not require any prior understanding of attack techniques or changes to the CNN design. Its effectiveness has been tested under white-box and black-box circumstances using a standard chest X-ray dataset. This method does not require expertise in attack methods or changes to CNN architecture. It has been tested in both white-box and black-box scenarios on a regular chest X-ray dataset, demonstrating its usefulness. Additionally, it is adaptable enough to be integrated with other protection strategies and applied to various medical imaging scenarios with different image types. It is anticipated that the implementation of this method will significantly improve the security of medical imaging classification systems relying on

deep learning.

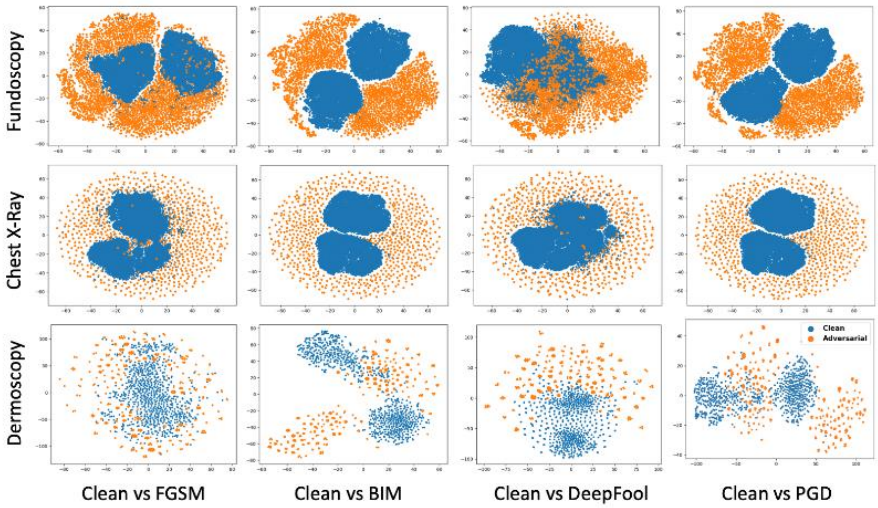

Figure 3. A graph showing a 2D representation of unaltered and changed features from the second-to-last densely connected layer of a DNN design. Every row is a set of data, each column is an attack, with blue/orange providing a visual cue for distinguishing between normal/malicious samples[1].

In an effort to guarantee the accuracy of medical diagnosis, it is essential to measure the strength of medical DNNs in the face of adversarial interference. It is important to assess the efficacy of these tasks, since DNN are commonly used in medical image classification to offer auxiliary information in clinical diagnosis. Earlier works have investigated basic adversarial samples. The loophole of deep neural networks to powerful and aggressive attacks, like universal adversarial perturbation (UAP), a single adversrary noise capable of disrupting DNN performance in almost all Categorize tasks, has been investigated. Hirano et al.[25] performed experiments on three DNN-based medical image classifications: skin cancer identification through photography, lung nodule detection from CT scans, and brain tumor recognition from MRI scans. Investigate the susceptibility of skin malignancy, diabetic eye condition, and pneumonic categorizations to the seven model architectures of UAPs in order to reduce the occurrence rate. DNNs are demonstrated to be susceptible to both nontargeted UAPs, resulting in an incorrect class assignment for an input, and targeted UAPs, causing the DNN to assign the input to a particular class, as depicted in **Figure 7**. The UAPs were nearly undetectable but still achieved high success rates, above 80%, when attempting both nontargeted and targeted attacks. The model architecture had little influence on the susceptibility to UAPs. For instance, the author observed that, while adversarial retraining is considered a successful defense against adversarial attacks, its utility in reinforcing the resilience of deep neural networks against UAPs is limited to certain scenarios. The results show that medical diagnosis based on natural neural network is more vulnerable to deceptive attacks than was previously believed. Such attacks are capable of causing incorrect diagnoses and can be executed at a cheap cost. The effects of adversarial defenses could be far-reaching, thus necessitating a thorough assessment when designing Deep Neural Networks for medical imaging and its related usage. Paul et al.[3] tackle the issue of adversarial manipulation of medical images in order to predict malignancy of lung nodules, with a five-year survival rate of 18% for lung cancer being the most common cancer across the globe. They compared two adversarial attack strategies, FGSM and one pixel attack, as illustrated in **Figure 6**. With regard to diminishing the accuracy of the categorize tasks, the FGSM approach was found to be more effective than the one pixel attack. Their ensemble of Convolutional Neural Networks (CNNs) proved to be more precise than a single classifier for malignancy prediction, even when adversarial images were used for training. Using it

leads to a 13% reduction in accuracy for fast gradient sign method and 10% for one pixel attack compared to an ensemble on unaltered data, representing a significant improvement from the more than 30% drop seen without protection.

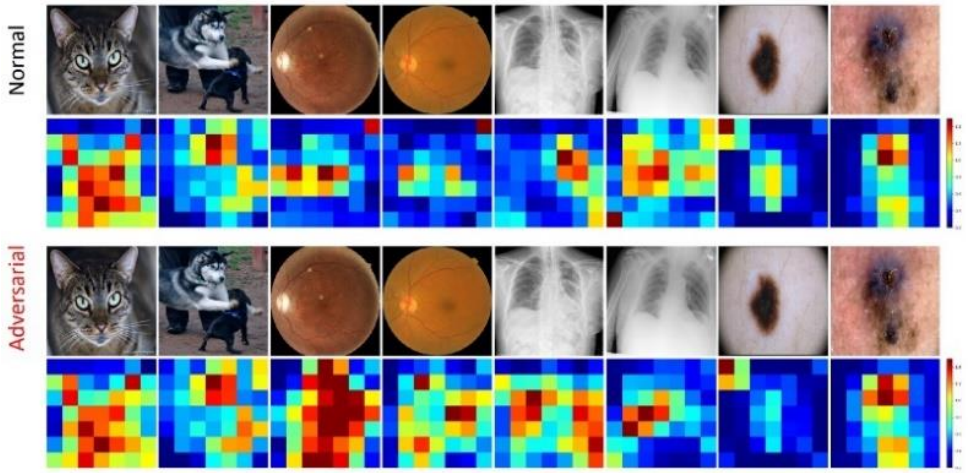

Figure 4. The ResNet-50 models at the 'res5b_relu' layer (channel averages included) have effectively minimized the duplication rate of representations of regular and adversarial images[1].

2.2. Segmentation

Recent studies have been utilizing deep learning models rapidly in order to address issues related to images for medical data sets. Ozbulak et al.[27] conducted an investigation into adversarial examples in regards to medical image segmentation issues. The author assessed the performance of the attack on two datasets: the Glaucoma Optic Disc Segmentation dataset[28] and the ISIC Skin Lesion Segmentation dataset[29]. The results are displayed in **Figure 7**. Taking into consideration the distinctions mentioned above and use the insights gained from researching adversarial examples in classification issues. The author suggests Adaptive Segmentation Mask Attack (ASMA) as a novel method to create targeted adversarial samples that focus on DNN-based image segmentation models, such as those used to analyze skin lesions or glaucoma optic discs. The flow of the ASMA method is shown in **Figure 14**. This work demonstrates the vulnerability of these models to such attacks.

Recent investigations have been utilizing deep learning models with rapid speed in order to address image-related issues with medical datasets. Adversarial examples have been used to show that DNNS are not invulnerable to attacks using gradient[30]. This was further demonstrated by the Dense Adversary Generation (DAG) algorithm[31], which is designed to cause DNNS to misclassify all pixels. The authors examine the issue of adversarial examples in relation to medical image segmentation issues and present an innovative approach for crafting targeted adversarial examples that are specific to this type of problem. Authors developed a novel algorithm to generate specifically tailored adversarial examples for image segmentation, aiming to produce the desired output shape. These modifications to the original image are nearly imperceptible to the untrained eye, yet still allow for accurate attainment of the intended results. Shao et al.[32] highlighted the possibility of a major security risk with image segmentation models through the target attack, which involves the use of gradients of various sizes. The process of creating adversarial examples is demonstrated in **Figure 11**.

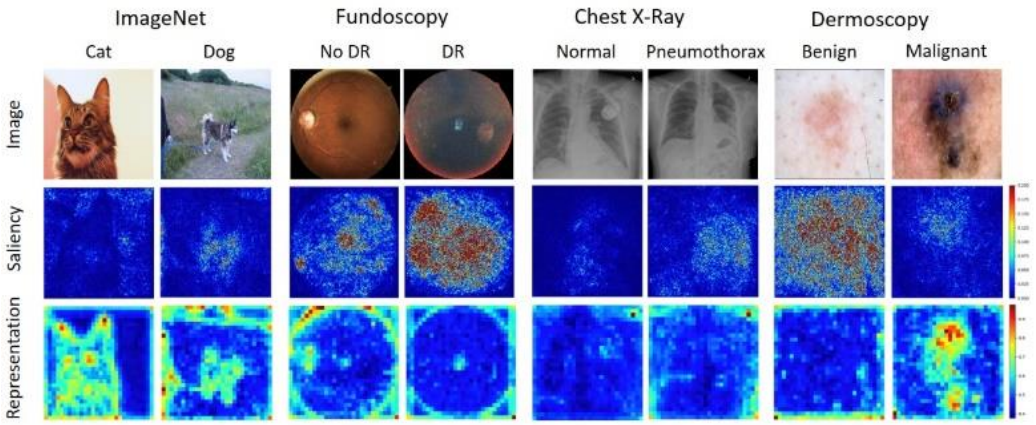

Figure 5. The top row has the standard images, the middle one has respective saliency maps, and the bottom row shows representations of the images found at the 'res3a_relu' (averaged across channels) layer of the networks[1].

The aim of the model is to find a series of parameters θ so that the output mask M is as precise as can be when given an image x. This is done by defining an objective function using θ and x that measures how accurate the predicted mask M is. The specific expression is as follows:

$$min \parallel x - x' \parallel_2, s.t. \operatorname{argmax}\big(U(\theta, \mathrm{x}')\big) = \mathrm{M}^{\mathrm{t}}$$

This attack uses gradients to create adversarial examples which resemble the original clean images, yet lead the segmentation model to produce a different output mask. By calculating the loss function with gradients, the input image is adjusted, resulting in the model predicting a mask that matches the target. The targeted attack in segmentation seeks to raise the prediction probability of the chosen pre-background knowledge of the target adversarial mask while decreasing the chances of all pixels that are not indicated in the mask. The SSM approach uses a mathematical equation to define the perturbation it adds to achieve this goal. Their approach reduces the amount of disturbance and resulting adversarial examples from Multiscale Attack (MSA) method, making them appear more visually pleasing compared to those created by the Adaptive Segmentation Mask Attack (ASMA) method. The MSA method produces adversarial examples that look better than those created by the ASMA method. This research looks into how adversarial examples affect a medical image segmentation model. Multiscale Attack (MSA) was proposed as a method of using multiscale gradients to generate a segmentation mask for a medical image segmentation model. Iteratively applying various levels of loss to the original image is used to interfere it so that the resulting adversarial example segmentation mask is close to the desired target mask. In the future, researchers will investigate the extent to which adversarial samples are likely to be transferred to other segmentation DNNs, and also explore ways to protect medical image segmentation DNNS from adversarial samples in order to increase the reliability of medical image DNNs.

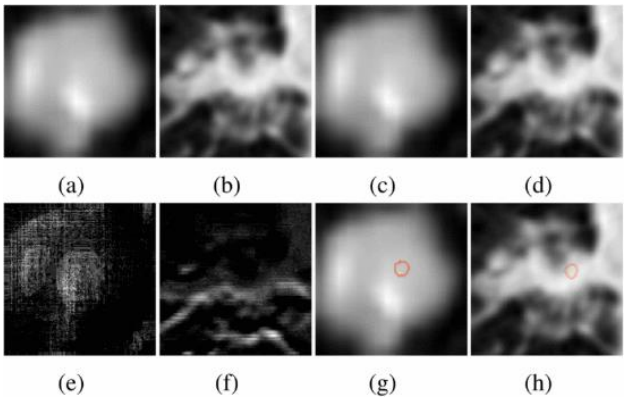

Figure 6. (a, b): Original nodules; (c, d): FGSM images; (e, f): Difference of original and FGSM images; (g, h): 1-pixel attack images (the pixel value changed is shown in red)[33].

Convolutional Neural Networks are a highly-regarded option for image classification and segmentation tasks because of their proficiency in extracting pertinent features from images and their generally good performance. Chen et al.[34] strived to create a new technique for producing adversarial examples that can target Convolutional Neural Networks models used for image segmentation. Creating adversarial examples to undermine semantic image segmentation models is difficult, as

one needs to assign a label to each pixel, rather than one label for the whole image, which is the typical approach used with adversarial attacks in computer vision. A successful adversarial attack for a classification model can cause the entire image to be mislabeled, whereas a successful adversarial attack for a segmentation DNNs need not result in an error prediction result for each pixel, but rather alter the model's segmentation output. Adversarial attacks usually only make slight changes to image brightness, but in medical imaging it's more useful to use deformations to attack segmentation models. Regardless of the segmentation model, unseen poses or shapes of organs can still pose a problem. A CT scan dataset with 150 subjects was tested, each with marked organs like the liver, kidneys, spleen, and pancreas identified by human professionals. Subjects were divided into three groups: training (60), validation (15), and testing (75). The authors implemented a 2D U-Net[35] to segment abdominal organs, using image patches for training and image slices for testing. In this study, a trained U-Net was used as a fixed Convolutional Neural Networks and exposed to adversarial samples. The authors had advanced a new method for creating adversarial examples to try and subvert a CNN model that been utilized for medical image segmenting. These generated adversarial samples feature both geometric changes to represent differences in anatomy, as well as variations in intensity that model how something looks. These examples demonstrated how Convolutional Neural Networks based segmentation models, such as a U-Net[35] , can have their Dice score decreased by a predetermined amount through training process, with no predefined guidelines. This study explored the potential of using the proposed method to create extra training images in order that a CNN-based segmentation model can be made more resilient to attacks. If the examples an adversary presents are realistic and plausible, then the Convolutional Neural Network model is not strong enough to resist attack.

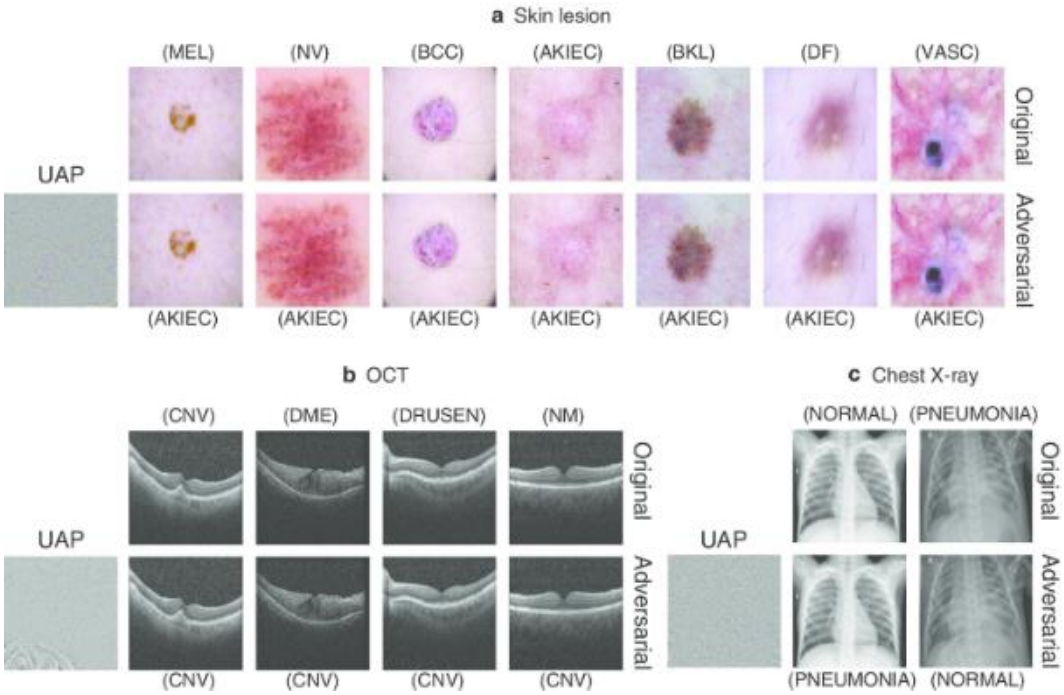

Figure 7. Utilize non-specific UAPs with a p value of two to challenge Inception V3 models, as well as their adversarial pictures, for skin lesions, OCT (Optical Coherence Tomography), and chest X-ray datasets(c). Furthermore, = 4% for a and c, and = 6% for b. Brackets next to the images denote the classifications predicted. The original (untouched) images are correctly labeled. UAPs are highlighted to ensure understanding; each one is graded on a scale of 0 to 1[25].

Figure 8. Summary of medical image adversary attack in recent years.

2.3 Overview of recent medical image attacks

From **Figure 8**, we can see that the neural network model trained on the medical data set is easy to be successful against attack, and the success rate is higher than 90%. Both gradient-based attacks and query-based attacks can make the model identify errors with high probability. Whether it is a white-box attack or a black-box attack, that is, no matter whether the known conditions required for the attack are more or less, the attack will have a greater probability of success. Of course, the less known preconditions for an attack, the more iterations it will take and the more time it will take. That is, the more known conditions, the easier it will be to attack. The less known conditions, the harder it will be to attack. The result is that the attack rate will be slightly lower. From **Figure 8**, we can see that although it is more difficult to attack medical image segmentation and medical image detection than to attack medical image classification, the attack methods in recent years show that it can perform very beneficial performance against samples for all medical image tasks, which also means that the medical image model has nothing to do with the task. As long as the deep neural network is used, it will be in a very insecure position because of the targeted attack against the adversary sample. So the defense of medical image model is very important because of the existence of adversary sample.

From **Figure 9** the statistics of the number of adversary attack and defense published in PubMed in the past five years, the spear and shield in the field of medicine is an accompanying progress. The progress and innovation of attack technology of medical image drive the update of corresponding defense technology. However, from the number of attack and defense papers per year, the number of attack related papers is always more than the number of defense papers. This also shows that defense is more difficult than attack from the side, but it should attract the attention of the academic community. More efforts should be devoted to the research of defense methods, so as to promote the robustness of medical image models and improve their use security. From **Figure 10**, we can see the number of attack and defense papers of natural images is almost an order of magnitude less than that of medical images. This must arouse the alert of the academic community. The safety of medical image models is related to the lives of patients. We should put more attention to the safety of medical models, which has great medical practical significance.

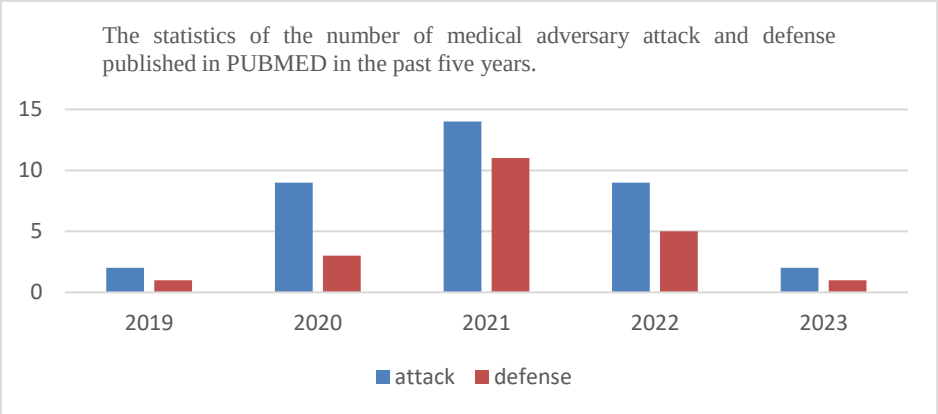

Figure 9. Statistics of the number of medical adversary attacks and defenses on PubMed.

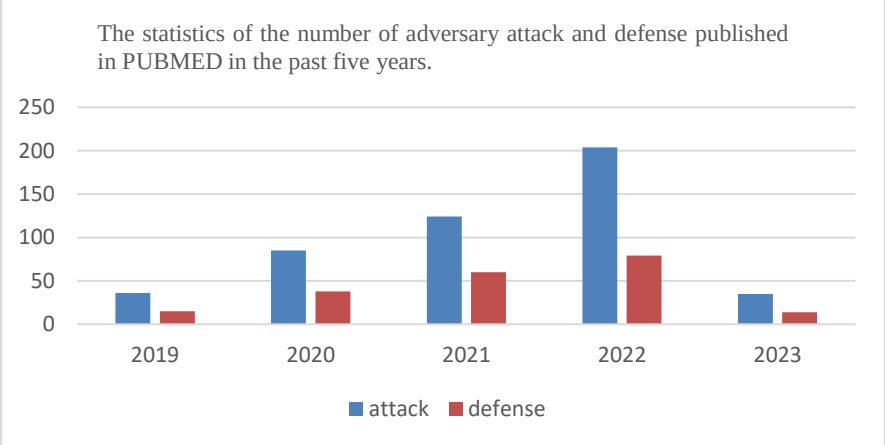

Figure 10. Statistics of the number of adversary attacks and defenses on PubMed.

3. Defense methods against attacks to medical image deep learning

The utilization of biomedical image analysis for computer-aided diagnosis and medical plan formulation had sparked increasing interest. The accuracy of CNN-based biomedical image segmen-tation surpassed that of traditional techniques. The precision of medical image segmentation is significant, but its dependability and reliability are important for it to be employed in clinical practice without mistakes. To this end, He X et al.[36] had created a non-local context encoder (NLCE), shown in **Figure 12**, which is designed to be resilient to adversarial attacks. NLCEs considered both close and far spatial associations and sharpened the qualities of feature maps by using channel-wise feature map attention based on the encoded overall contexts. NLCE can be broken down into two parts, as displayed in the **Figure 13**. By combining global spatial dependencies and global contextual information, NLCE's resilience to malicious attacks is increased. The authors presented their new NLCEN model in an effort to improve segmentation accuracy and capture more distinct boundaries. This framework captures high-level and powerful feature representations from different scales, then fused them to produce the final output (H, W, C).

He X, et al.[36] proposed the integration of a Non-Local Context Encoder (NLCE) into a Non-Local Context Encoding Network (NL-CEN) to enhance the robustness and accuracy of biomedical im-age segmentation in the face of adversarial

attacks. The proposed technique was evaluated on both lung and skin lesion segmentation datasets, which demonstrated its ability to reduce adversarial perturbations and protect against them while not sacrificing segmentation accuracy. The NLCE modules developed by the authors can enhance the resilience of other biomedical image segmentation techniques against malicious attempts. Xin L, et al.[5], points out that DNNs have demonstrated impressive results in a number of medical imaging applications, for instance, pneumonia detection from X-ray images[13] and early diagnosis of prostate cancer from MRI scans[37] . Khowaja SA, et al. [13] explored the application of optical coherence tomography (OCT) for classifying diseases of the retina, while[38] looked into segmentation of pulmonary nodules from CT scans. To evaluate the effectiveness of DNNs against adversarial samples, Ozbulak U, et al.[27] proposed the adaptive segmentation mask attack (ASMA) to craft a mask to mislead the model.

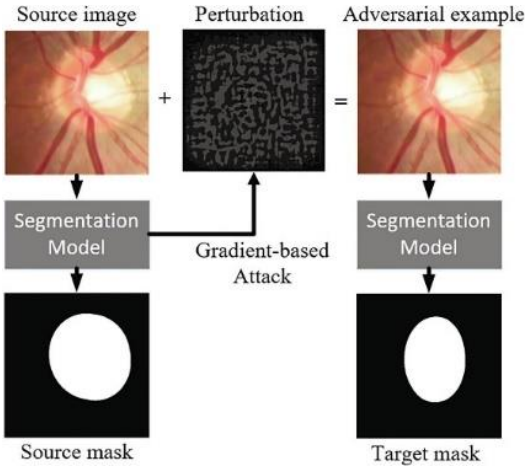

Figure 11. The working flow of generating adversarial example[32].

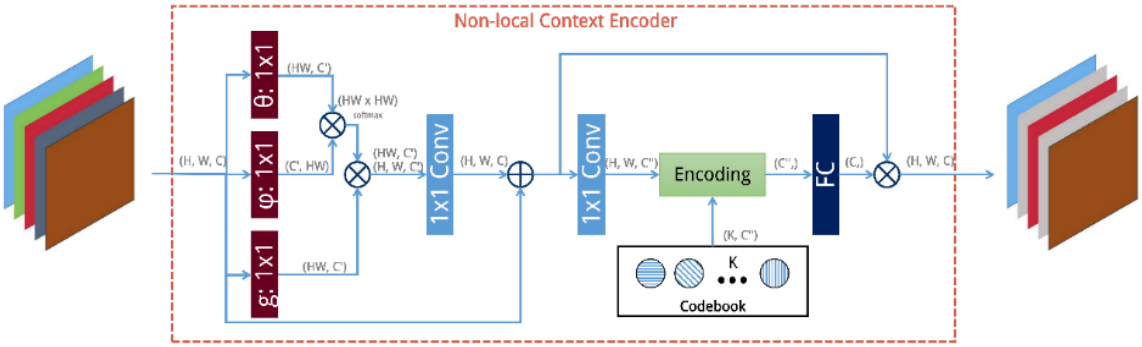

Figure 12. NLCE design starts by amplifying and cleaning up the feature map by taking into consideration the global spatial connections and then utilizes the encoded global context from a learned codebook to apply channel-wise feature map attention[36].

Concerns have arisen about the use of medical image classification model on a wide scale due to the loophole of deep neural network to adversarial samples. The use of deep neural networks (DNNs) for medical image is difficult, as they are mainly used to operate with natural images and require a sizable set of training data[39], while medical data set is often restricted in terms of labeled samples. As Li X, et al. [5], and Taghanaki SA, et al.[40] demonstrated, an alternative method for safeguarding against malicious attacks is to train models to distinguish between normal and adversarial samples. This hybrid approach combines SSAT and UAD to bolster the performance of DNNs in defense. The authors make use of labeled and unlabeled data to create synthetic labels for SSAT to heighten the accuracy of categorization. This technique is particularly intended for medical imaging datasets which have a limited amount of labels and is equipped to manage various unseen adversarial conditions.

Studies have indicated that tiny changes in data too subtle for the human eye to detect can lead to DNNs misclassifying, which has ignited much attention in deep learning[30,41]. Research on various tasks, such as classification, have highlighted the importance of accurately gauging robustness[42], object detection[43] and semantic segmentation[31,44,45]. Various approaches to utilizing computer technology for medical diagnosis have been developed to provide more precise segmentation of specific anatomical structures[46,47,48,49]. These methods are customized to cover the broad array

of structures, imaging modalities, and accuracy measures available. Evaluating the resistance of a model against adversarial examples has proven to be a difficult challenge[50], with a variety of strategies being suggested[41,51].

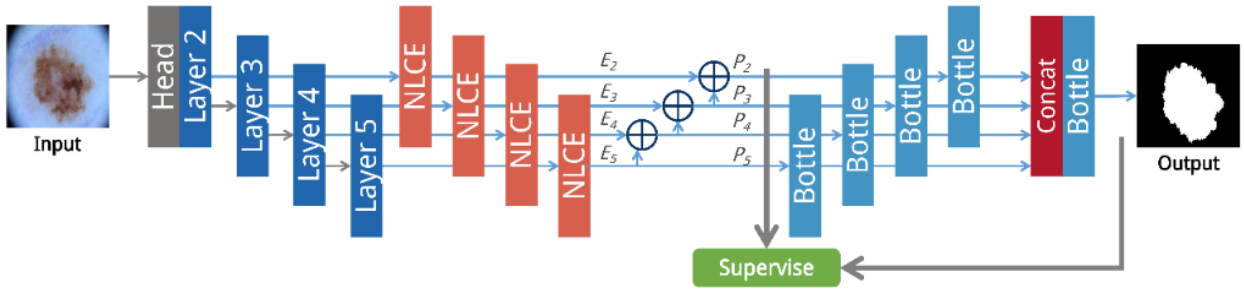

Figure 13. The author's blueprint for the NLCEN seeks to decrease the repetition rate and was crafted with great vigor. It is based on a ResNet backbone and feature pyramid. A NLCE module is tacked on to the feature activations that originate from the bottom-up path, after which lateral links are formed at distinct levels, with independent monitoring being operated. To reduce the occurrence of duplication, the details from each pyramidal component are utilized to reinforce the prediction and create segmentation at every level[36].

AutoAttack is a technique used to attack models that allocate a probability to each sample. Its authors aim to assess DNNs models that allocate a probability to every spatial site and measure the reduction in performance when noise is added. They assess this by inspecting the decrease in dice score, which is determined by averaging the performance on clean images and halving it. The authors modify the attack to enable the calculation of voxelwise functions over all spatial locations while still upholding the definition of "Robust Accuracy" as the worst result of the four attacks on an individual case. They measure the total number of FLOPs by taking the average of results obtained through all methods using an input size of 96 × 96 × 96, as previously research[52]. This experiment shows that ROG is a good choice for analyzing 3D segmentation in medical image processing, considering the high computational cost of other methods which only yield 79.32 GFLOPs. This is beneficial since robustness to adversarial attacks is a computationally demanding task.

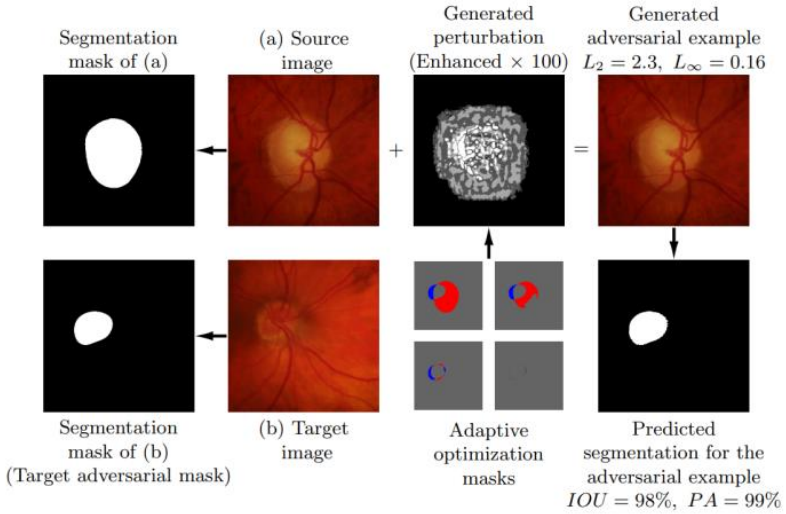

Figure 14. An illustration of how ASMA can optimize an adversarial example for segmentation can be seen best in color[27].

4. Discussion

Currently, white box attacks are becoming

more frequent, while black box attacks on medical image recognition models are decreasing. In white box attacks, both medical image classification and segmentation tasks have attained high success rates. Once the medical image recognition model and the network structure parameters are securely protected, access to the model is restricted, resulting in difficulties for white box attacks with high attack success rate and black box attacks which require a large number of queries.

However, the robustness of the model is always enhanced along with the enhancement of attacks which is also called while the priest climbs a foot, the devil climbs ten. Therefore, the academia should spend more time on black box attacks, especially those black box attacks with low query times and less information required. For example, attackers can successfully attack only by knowing the output result of the model, or even without the confidence rate of the model output. In addition, the training data set of medical image model is not as easy to obtain as the natural image data set, which will cause a lot of difficulties in order to generate black box attacks. In the future, we can solve the difficulties that medical data sets are not easy to obtain by improving the migration of confrontation samples from natural images to medical images. We can also look at creating simulated fake medical datasets that come close to real data, without the need of real medical data, in order to address the issue of a lack of medical training datasets.

In terms of the defense of medical recognition models, the introduction of confrontation training is undoubtedly a very efficient defense strategy at present. However, there is no free lunch in the world. The use of confrontation training generally increases the robustness of the model, but can adversely impact the accuracy of the medical model with pristine data. Therefore, how to improve the confrontation training in the future to achieve a balance between high robustness and high accuracy is the focus of research. In addition, it may be a good choice to put the medical adversary sample detection before the defense model after adversary training. Specifically, let the medical adversary sample detection module first judge whether the input is a clean sample or an adversary sample. If it is a clean sample, the original model without adversary training will be used for identification. Otherwise, if it is an adversary sample, the model with adversary training may be used for identification with slightly reduced accuracy. In this way, it may be possible to give consideration to both the high accuracy of clean samples and the strong defense ability against adversary samples.

The results of medical models are essential for the wellbeing of patients, so it is important for the academic community to focus on enhancing the reliability of these models. Their accuracy is also a factor in ensuring patient health. If the manufacturer of medical recognition model equipment pays attention to providing long-term software update support for medical equipment, such as timely applying the latest medical robust recognition model methods to medical model equipment being used by hospitals, the medical recognition model will remain immune to the latest attack methods. After all, medical model security is an endless battle of attack and defense, only by constantly updating the defense methods of the equipment can the medical model equipment be under a dynamic security. On the other hand, training medical workers in medical image sensitivity can minimize the security risk of medical confrontation.

Generally, when the adversary samples and original data are negligible, the machine recognition model will output wrong results. However, if the medical institutions do not rely entirely on the machine medical recognition model to judge, they can make manual secondary judgments on the medical images that have been recognized by the machine, so that they can make correct judgments as far as possible, so as to maximize support for and patient's condition by providing accurate judgment results.

Today, the demand for the medical metaverse

is growing, because chronic diseases, led by cancer, diabetes and cardiovascular diseases, have become a major threat to human health. However, few experts can accurately judge these diseases, especially in grassroots hospitals where medical treatment is not developed and high-end equipment coverage is low. For this reason, the vigorous development of smart medicine and the medical metaverse needs to attract attention from all parties. Smart medicine and the medical metaverse can break through the current situation of extremely uneven distribution of medical resources in time, space, and computing power. Using 5G and cloud computing to achieve unlimited time and location diagnosis can greatly provide timely and accurate diagnosis to these patients with poor medical conditions around them, thereby obtaining timely preventive diagnosis and treatment. It can greatly save the money and medical resources consumed in later treatment, and save more lives.

Although intelligent medicine and the medical metaverse can bring great convenience to treatment, they also bring new problems. Medical image information, such as CT, MRI, and color Doppler ultrasound, is processed in a large amount in the medical metaverse. However, once the medical image information is tampered with by hackers into medical image adversary samples, it will cause medical model to output erroneous diagnoses. If the medical model of intelligent medicine and medical metaverse is not defensible and robust, the entire intelligent medicine and metaverse will be at great risk.

Therefore, when building a medical metaverse and intelligent medicine, we must take into account the safety and defense of medical identification models. As a key infrastructure for the medical metaverse and intelligent medicine, the medical robustness identification model, if it is safe and robust, will also make the entire medical metaverse and intelligent medicine safe and trustworthy in general.

5. Conclusion

Although machine learning algorithms have high accuracy and performance in intelligent medicine and the medical metaverse, they are found to be vulnerable to subtle perturbations, resulting in disastrous consequences in security related environments, especially in the field of medical image deep learning networks. This paper aims to summarize and sort out the attack and defense methods of confrontation samples in the field of medical image classification and segmentation. The limitations of current research methods and the future research directions are been discussed, and also offered some useful advice to enhance the robustness of medical models with a view to promoting the construction of a more reliable and more robust medical deep learning model. It is hoped that this review of on adversarial attack and defense of deep learning models for medical image recognition can provide a reference for the academic community to build a more robust and secure intelligent medical and medical metaverse, and attract the attention of relevant personnel. We are looking forward to building a safe and robust intelligent medicine, as well as a trustworthy, efficient and safe medical metaverse, so that patients around the world can receive more convenient, safe, and accurate diagnosis and treatment.

Conflict of interest

Authors declare that there is no conflict of interest.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62072126, and No. 61772164), the Fundamental Research Projects Jointly Funded by Guangzhou Council and Municipal Universities (No. 202102010439).

References

1. Ma X, Niu Y, Gu L, et al. Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognition 2021; 110: 107332.

2. Li X and Zhu D. Robust detection of adversarial attacks on medical images. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI); 2020 Apr. 3-7; Lowa City, IA, USA. New York: IEEE; 2020. p. 1154–1158. doi: 10.1109/ISBI45749.2020.9098628.
3. Paul R, Schabath M, Gillies R, et al. Mitigating adversarial attacks on medical image understanding systems. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI); 2020 Apr. 3-7; Lowa City, IA, USA. New York: IEEE; 2020. p. 1517–1521. doi: 10.1109/ISBI45749.2020.9098740.
4. Park H, Bayat A, Sabokrou M, et al. Robustification of segmentation models against adversarial perturbations in medical imaging. In: Rekik I, Adeli E, Park SH, et al. (editors). Predictive Intelligence in Medicine. PRIME 2020. Lecture Notes in Computer Science (vol. 12329). Cham: Springer; 2020. p. 46–57. doi: 10.1007/978-3-030-59354-4_5.
5. Li X, Pan D, Zhu D. Defending against adversarial attacks on medical imaging AI system, classification or detection? In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI); 2021 Apr. 13-16; Nice, France; New York: IEEE; 2021. p. 1677–1681.
6. Minagi A, Hirano H, Takemoto K. Natural images allow universal adversarial attacks on medical image classification using deep neural networks with transfer learning. Journal of Imaging 2022; 8(2): 38. doi: 10.3390/jimaging8020038.
7. Apostolidis KD, Papakostas GA. Digital watermarking as an adversarial attack on medical image analysis with deep learning. Journal of Imaging 2022; 8(6): 155. doi: 10.3390/jimaging8060155.
8. Winter TC. Malicious adversarial attacks on medical image analysis. American Journal of Roentgenology 2020; 215(5): W55–W55. doi: 10.2214/AJR.20.23250.
9. Desjardins B, Ritenour ER. Reply to "Malicious Adversarial Attacks on Medical Image Analysis". American Journal of Roentgenology 2020; 215(5): W56–W56.
10. Yao Q, He Z, Zhou SK. Medical aegis: Robust adversarial protectors for medical images. arXiv: 2111.10969v1. 2021.
11. Zhou Q, Zuley M, Guo Y, et al. A machine and human reader study on AI diagnosis model safety under attacks of adversarial images. Nature Communications 2021; 12(1): 1–11.
12. Selvakkumar A, Pal S, Jadidi Z. Addressing adversarial machine learning attacks in smart healthcare perspectives. In: Suryadevara NK, George B, Jayasundera KP, et al. (editors). Sensing Technology. Lecture notes in electrical engineering 2022; 886: 269–282. doi: 10.1007/978-3-030-98886-9_21.
13. Khowaja SA, Lee IH, Dev K, et al. Get your foes fooled: Proximal gradient split learning for defense against model inversion attacks on IoMT data. arXiv: 2201.04569 v3. 2022. doi: 10.48550/arXiv.2201.04569.
14. Rodriguez D, Nayak T, Chen Y, et al. On the role of deep learning model complexity in adversarial robustness for medical images. BMC Medical Informatics and Decision Making 2022; 22(2): 160. doi: 10.1186/s12911-022-01891-w.
15. Jin R, Li X. Backdoor attack is a devil in federated GAN-based medical image synthesis. In: Zhao C, Svoboda D, Wolterink JM, et al. (editors). Simulation and Synthesis in Medical Imaging. SASHIMI 2022. Lecture Notes in Computer Science; Cham: Springer; 2022. p. 154–165. doi: 10.1007/978-3-031-16980-9_15.
16. Kos J, Song D. Delving into adversarial attacks on deep policies. arXiv:1705.06452v1. 2017. doi: 10.48550/arXiv.1705.06452.
17. Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum. In: 2018 IEEE/CVF Conference on sComputer Vision and Pattern Recognition; 2018 Jun. 18-23; Salt Lake City, UT, USA; 2018. p. 9185–9193. doi: 10.1109/CVPR.2018.00957.
18. Naseer M, Khan SH, Porikli F. Local gradients smoothing: Defense against localized adversarial attacks. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV); 2019 Jan. 7-11. New York: IEEE; 2019. p. 1300–1307. doi: 10.1109/WACV.2019.00143.
19. Dai H, Li H, Tian T, et al. Adversarial attack on graph structured data. In: Proceedings of the 35th International Conference on Machine Learning; 2018 Jul. 10-15; Stockholm, Sweden. Priscilla Rasmussen: Curran Associates, Inc.; 2018; 80: 1115–1124.
20. Lin Z, Shi Y, Xue Z. IDSGAN: Generative adversarial networks for attack generation against intrusion detection. In: Gama J, Li T, Yu Y, et al. (editors). Advances in Knowledge Discovery and Data Mining. PAKDD 2022. Lecture Notes in Computer Science 2022. Cham: Springer; 2022; 13282: 79–91. doi: 10.1007/987-3-031-05981-0_7.
21. Qiu S, Liu Q, Zhou S, et al. Review of artificial intelligence adversarial attack and defense technologies. Applied Sciences 2019; 9(5): 909. doi: 10.3390/app9050909.
22. Dong Y, Pang T, Su H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun. 15-19; Long Beach, CA, USA; New York: IEEE; 2019. p. 4312–4321. doi: 10.1109/CVPR.2019.00444.
23. Morris JX, Lifland E, Yoo JY, et al. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. arXiv:

2005.05909v4. 2020.
doi: 10.48550/arXiv.2005.05909.
24. Feinman R, Curtin RR, Shintre S, et al. Detecting adversarial samples from artifacts. arXiv:1703.00410v3. 2017. doi: 10.48550/arXiv. 1703.001410.
25. Hirano H, Minagi A, Takemoto K. Universal adversarial attacks on deep neural networks for medical image classification. BMC Medical Imaging 2021; (21)1: 1–13.
26. Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun. 18-23; Salt Lake City, UT, USA; New York: IEEE; 2018. p. 9185–9193. doi: 10.1109/CVPR.2018.00957.
27. Ozbulak U, Messem AV, Neve WD. Impact of adversarial examples on deep learning models for biomedical image segmentation. In: Medical Image Computing and Computer Assisted Intervention-MICCAI 2019. 2019 Oct. 13-17; Shenzhen, China; Lecture Notes in Computer Science (vol. 11765); Cham: Sprigner; 2019. p. 300–308. doi: 10.1007/978-3-030-32245-8_34.
28. Pena-Betancor C, Gonzalez-Hernandez M, Fumero-Batista F, et al. Estimation of the relative amount of hemoglobin in the cup and neuroretinal rim using stereoscopic color fundus images. Investigative Ophthalmology & Visual Science 2015; 56(3): 1562–1568. doi: 10.1167/iovs.14-15592.
29. Codella NCF, Gutman D, Celebi ME, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018); 2018 Apr. 4-7; Washington, DC, USA; New York: IEEE; 2018.p. 168–172. doi: 10.1109/ISBI.2018.8363547.
30. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. arXiv:1312.6199. 2013. doi: 10.48550/arXiv.1312.6199.
31. Xie C, Wang J, Zhang Z, et al. Adversarial examples for semantic segmentation and object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct. 22-29; Venice, Italy; New York: IEEE; 2017. p. 1369–1378. doi: 10.1109/ICCV.2017.153.
32. Shao M, Zhang G, Zuo W, et al. Target attack on biomedical image segmentation model based on multi-scale gradients. Information Sciences 2021; 554: 33–46. doi: 10.1016/j.ins.2020.12.013.
33. Eladawi N, Elmogy MM, Ghazal M, et al. Classification of retinal diseases based on OCT images. Front Biosci (Landmark Ed) 2018; 23(2): 247–264. doi: 10.2741/4589.PMID:28930545.
34. Chen H, Liang J, Chang S, et al. Improving adversarial robustness via guided complement entropy. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Apr. 13-16; Seoul, Korea (South); 2019. p. 4881–4889. doi: 10.1109/ICCV.2019.00498.
35. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, et al. (editors). Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015. 2015 Oct. 5-9; Lecture Notes in Computer Science 2015; Munich, Germany; 2015. p. 234–241. doi: 10.1007/978-3-319-24574-4_28.
36. He X, Yang S, Li G, et al. Non-local context encoder: Robust biomedical image segmentation against adversarial attacks. In: Proceedings of the AAAI Conference on Artificial Intelligence; Palo Alto, California, USA; California: AAAI Press; 2019; 33(01):8417–8424. doi: 10.1609/aaai.v33i01.33018417.
37. Reda I, Ayinde BO, Elmogy M, et al. A new CNN-based system for early diagnosis of prostate cancer. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018); 2018 Apr. 4-7; Washington; DC; USA; New York: IEEE; 2018. p. 207–210. doi: 10.1109/ISBI.2018.8363556.
38. Qin Y, Zheng H, Huang X, et al. Pul-monary nodule segmentation with CT sample synthesis using adversarial networks. Medical Physics 2019; 46(3): 1218–1229. doi: 10.1002/mp.13349.
39. Stanforth R, Fawzi A, Kohli P, et al. Are labels required for improving adversarial robustness? In: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); 2019 Dec. 8-14; Vancouver; Canada; 2019. p. 1–10.
40. Taghanaki SA, Abhishek K, Azizi S, et al. A kernelized manifold mapping to diminish the effect of adversarial perturbations. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun. 15-20; Long Beach, CA, USA; New York: IEEE; 2019. p. 11332–11341. doi: 10.1109/CVPR.2019.01160.
41. Carlini N and Wagner D. Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP); 2017 May 22-24; San Jose; CA; USA; New York: IEEE; 2017.p. 39–57. doi: 10.1109/SP.2017.49.
42. Madry A, Makelov A, Schmidt L, et al. Towards deep Learning models resistant to adversarial attacks. arXiv: 1706.06083v4. 2017. doi: 10.48550/arXiv.1706.06083.
43. Zhang H, Wang J. Towards adversarially robust object detection. arXiv: 1907.10310v1. 2019. doi: 10.48550/arXiv.1907.10310.
44. Arnab A, Miksik O, Torr PHS. On the robustness of semantic segmentation models to adversarial attacks. arXiv:1711.09856v1. 2018. doi: 10.48550/arXiv.1711.09856.

45. Cisse M, Adi Y, Neverova N, et al. Houdini: Fooling deep structured prediction models. arXiv:1707.05373.; 2017. doi: 10.48550/arXiv.1707.05373.
46. Isensee F, Jaeger PF, Full PM, et al. nnU-Net for brain tumor segmentation. arXiv:2011.00848. 2020. doi: 10.48550/arXiv.2011.00848.
47. Milletari F, Navab F, Ahmadi SA. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D vision (3DV); 2016 Oct. 25-28; Stanford; CA;USA; New York: IEEE; 2016. p. 565–571. doi: 10.1109/3DV.2016.79.
48. Tang H, Zhang C, and Xie C. Nodulenet: Decoupled false positive reduction for pulmonary nodule detection and segmentation. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2019. Lecture Notes in Computer Science; 2019 Oct. 13-17; Shenzhen, China; Cham: Springer; 2019. p. 11769. doi: 10.1007/978-3-030-32226-7_30.
49. Zhu Z, Xia Y, Shen W, et al. A 3D coarse-to-fine framework for volumetric medical image segmentation. In: 2018 International Conference on 3D Vision (3DV); 2018 Sep. 5-8; Verona; Italy; New York: IEEE; 2018. p. 682–690. doi: 10.1109/3DV.2018.00083.
50. Carlini N, Athalye A, Papernot N, et al. On evaluating adversarial robustness. arXiv:1902.06705. 2019.
51. Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. arXiv: 1511.04599v3. 2016. doi: 10.48550/arXiv.1511.04599.
52. Yu Q, Yang D, Roth H, et al. C2FNAS: Coarse-to-fine neural architecture search for 3D medical image segmentation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun. 13-19; Seattle, WA, USA; New York: IEEE; 2019. p. 4125–4134. doi: 10.1109/CVPR42600.2020.00418.