Expert Systems With Applications 198 (2022) 116815

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

Adversarial attacks and defenses on AI in medical imaging informatics: A survey

Sara Kaviani, Ki Jin Han, Insoo Sohn *

Division of Electronics & Electrical Engineering, Dongguk University, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Complex systems
Artificial neural networks
Optimization

ABSTRACT

In recent years, medical images have significantly improved and facilitated diagnosis in versatile tasks including classification of lung diseases, detection of nodules, brain tumor segmentation, and body organs recognition. On the other hand, the superior performance of machine learning (ML) techniques, specifically deep learning networks (DNNs), in various domains has lead to the application of deep learning approaches in medical image classification and segmentation. Due to the security and vital issues involved, healthcare systems are considered quite challenging and their performance accuracy is of great importance. Previous studies have shown lingering doubts about medical DNNs and their vulnerability to adversarial attacks. Although various defense methods have been proposed, there are still concerns about the application of medical deep learning approaches. This is due to some of medical imaging weaknesses, such as lack of sufficient amount of high quality images and labeled data, compared to various high-quality natural image datasets. This paper reviews recently proposed adversarial attack methods to medical imaging DNNs and defense techniques against these attacks. It also discusses different aspects of these methods and provides future directions for improving neural network's robustness.

1. Introduction

Medical imaging has a revolutionary impact on medicine with a great ability to diagnose disease via imaging the human body organs, cells and pathological specimens. Medical imaging informatics is associated to all medical imaging tasks from image processing to image storage, analysis, retrieval and understanding (Bui & Taira, 2009; Kulikowski, 1997; Panayides, et al., 2020). Therefore, the goal of medical imaging informatics is to improve the accuracy, precision, efficiency and reliability of medical services.

Deep neural network (DNN), as one of the most efficient methods in artificial intelligence (AI), have become prominent recently in medical imaging, to improve diagnosis and assist medical staff to accelerate decision making in clinical tasks. DNNs have achieved a milestone in data processing and information acquisition from big data in many disciplines. There are various examples of medical image DNNs including early diagnosis of skin cancer classification from photographic images (Esteva, et al., 2017), classifying diabetic retinopathy from optical coherence tomography (OCT) images (Kermany, et al., 2018), pneumonia detection from chest X-ray (Kermany, et al., 2018), and nodule segmentation from CT images (Qin, Zheng, Huang, Yang, & Zhu, 2019). The analysis indicated that medical DNNs have shown near-human performance and their diagnosis are equivalent to professional medical staff (Liu, et al., 2019).

Learning systems are extremely cheap with high accuracy and remarkable results comparable to standard clinical practice and have already been approved by the United states food and drug administration (FDA). Nevertheless, the security and reliability of medical DNNs are of great importance for the scientists. Recent studies on both classification and segmentation tasks of medical imaging has shown that even state-of-the-art DNNs are significantly vulnerable to adversarial attacks. Medical imaging DNNs are even more vulnerable to adversaries than the DNNs with natural images as their input (Ma, et al., 2021). These vulnerabilities allow small crafted perturbations in image samples, which are imperceptible to human eye, excessively affect the DNN's performance. These harmful adversaries have become one of the most important challenges in medical deep learning systems. To generate adversarial attacks, various methods have been proposed such as fast gradient sign method (FGSM) (Goodfellow, Shlens, & Szegedy, 2014) and its stronger variants such as projected gradient descent (PGD) (Madry, Makelov, Schmidt, Tsipras, & Vladu, 2017), and Carlini and Wagner (C&W) method (Carlini & Wagner, 2017). In Finlayson, Chung, Kohane, and Beam (2018), the authors explained different incentives to attack medical learning systems mostly stem from monetary issues while physicians and nursing pay is one of the most expensive services in US (Papanicolas, Woskie, & Jha, 2018). Therefore, medical

* Corresponding author.
E-mail addresses: s.kaviani@dongguk.edu (S. Kaviani), kjhan@dongguk.edu (K.J. Han), isohn@dongguk.edu (I. Sohn).

https://doi.org/10.1016/j.eswa.2022.116815
Received 15 February 2021; Received in revised form 5 February 2022; Accepted 1 March 2022
Available online 19 March 2022

DNNs cannot replace physicians and medical specialists. Due to the medical image characteristics and small amount of annotated data, even with the presence of human experts (i.e., engineers, physicians, and radiologists), these neural networks have shown to be prone to rapidly updated and improved adversarial examples.

To defend against adversarial attacks, various mitigation and detection techniques have been proposed. One of the most popular methods are those which are based on adversarial training. This method augments the adversarial samples to the training dataset and improve the neural networks robustness against adversarial attacks. Despite the efficiency of defense approaches and previously mentioned adversarial attacks, all of these methods have shown better performance on DNNs with natural image datasets (i.e., CIFAR-10) than medical DNNs. It has been proven that lack of sufficient amount of high quality images and labeled data is one of the main reasons for these weaknesses.

In this paper, we survey different types of newly invented adversarial attacks against medical imaging informatics and various defense techniques. In Section 2, a brief introduction to deep learning based medical imaging and its variants is discussed. In 3, we explain adversarial attacks and different types of generating adversarial examples. In Sections 4 and 5, the most recent adversarial attack and defense approaches to medical imaging DNNs has been summarized. Finally in Sections 6, we discuss different aspects and deficiencies of the methods, possible future challenges, and provide concluding remarks.

2. Deep learning based medical image analysis

According to the DNN development and success in various domains of computer vision and high quality annotated medical images, deep learning (DL) based medical image analysis have significantly enhanced the precision of medical decision support systems. These intelligent systems which play the role of computer-aided diagnosis systems, can provide assessments of the intrinsic disease processes such as many disease progression like Alzheimer's and different types of cancer. Recently, convolutional neural networks (CNNs) have been actively applied in medical imaging systems due to their outstanding performance and their ability to become engaged with GPUs (Greenspan, Ginneken, & Summers, 2016; Havaei, Guizard, Larochelle, & Jodoin, 2016). CNNs use convolutional layers to extract features automatically from the input data and provide classification such as cancer diagnosis and segmentation tasks including brain tumor segmentation (Casamitjana, Puch, Aduriz, & Vilaplana, 2016; Hwang & Park, 2017; Litjens, et al., 2017). There are different types of CNNs which have been used in medical imaging research community such as ResNet (He, Zhang, Ren, & Sun, 2015), GoogLeNet (Inception V1–V4) (Szegedy, Ioffe, Vanhoucke, & Alemi, 2017; Szegedy, et al., 2015; Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016), VGG (Simonyan & Zisserman, 2014), YOLO (Redmon & Farhadi, 2017), DenseNet (Huang, Liu, Maaten, & Weinberger, 2017) and UNet (Ronneberger, Fischer, & Brox, 2015). Nevertheless, there are some limitations to these DNN-based medical systems which mainly stem from small or incomplete training datasets. Manually annotating medical images needs abundant time, cost and medical experts. Moreover, these images are also vague and annotating them is absolutely subjective and highly variable even among expert physicians. This data scarcity will result in overfitting issues and generalization problems. In the remainder of this section, we will discuss classification and segmentation procedures in different medical application.

2.1. Classification

CNNs provide the possibility to automate feature extraction from the input images and disease diagnosis with classification procedure (Kleesiek, et al., 2016; Nie, Zhang, Adeli, Liu, & Shen, 2016) and have resulted in outstanding performance. Classification has been used in various medical issues including CNN based methods for classification of lung diseases based on computed tomography (CT) images (Anthimopoulos, Christodoulidis, Ebner, Christe, & Mougiakakou, 2016) and lung nodules from chest X-ray (Lo, Lou, Lin, Freedman, Chien, & Mun, 1995; Shen, Zhou, Yang, Yang, & Tian, 2015), melanoma classification (Haenssle, et al., 2018), and diabetic retinopathy detection (Abràmoff, Lavin, Birch, Shah, & Folk, 2018). Most of the medical classification tasks have been done by fine-tuning previously trained networks with ImageNet. For the first time Shin, et al. (2016) applied CNN models, that have been trained on natural images, in medical classification tasks and showed that highly accurate performance can be achieved. The superiority of pre-trained neural networks have also been proved in different researches (McKinney, et al., 2020; Tajbakhsh, et al., 2016).

2.2. Segmentation

Since the first successes in CNN-based medical classification and the revolutionary results of AlexNet, UNet, Boltzmann machines and autoencoders, DL-based segmentation of body organs in pathology has shown outstanding breakthrough. Segmentation procedure arrange the partitioning of the input image into separate parts by a pre-defined criterion including intrinsic color, texture and contrast (Qayyum, Qadir, Bilal, & Al-Fuqaha, 2020). Although obtaining dense predictions started from classification neural networks, due to some limitations, specific neural networks for segmentation approaches have designed and released such as 2D and 3D UNet (Ronneberger et al., 2015), which has been one of the best architectures for medical image segmentation in recent years. UNet is an encoder–encoder neural network that contains skip connections from encoding to decoding layers. These connections make it possible to train the neural network with small amount of training data with highly accurate segmentation performance. Usually, CNN-based medical image segmentation has been used in quantitative analysis of abnormalities in terms of clinical parameters such as measuring the shape and volume of brain tumors, skin cancer, and abdominal organs disease that results in early diagnosis of these abnormalities and preventing their progress (Hesamian, Jia, He, & Kennedy, 2019).

3. Adversarial attacks

In conventional machine learning systems, adversarial examples are inputs that have been manipulated and prepared to force the system to make incorrect classification and make difficulties in semantic segmentation. This type of attack, that causes problems in various domains such as spam filters, intrusion detection systems and biometric authentication (i.e. presentation attack), have been discussed for a few decades. The problem was first formulated by Dalvi, Domingos, Sanghai, and Verma (2004) and was updated in the context of deep computer vision systems by Szegedy, et al. (2013) and Goodfellow et al. (2014). As is shown in Fig. 1, the main objective of adversaries, by crafting these examples, is to extremely affect the systems performance while the perturbations are imperceptible to human eye. An overview of adversarial attacks history can be found in Yuan, He, Zhu, and Li (2019).

3.1. Adversaries knowledge

3.1.1. Black-box

In these attacks, it is assumed that the attacker does not have any knowledge about or to the trained model, training dataset, model parameters, and any information more than what is accessible to a normal user. Black-box attack is known to be a formidable task and is common when attacking an online ML service.

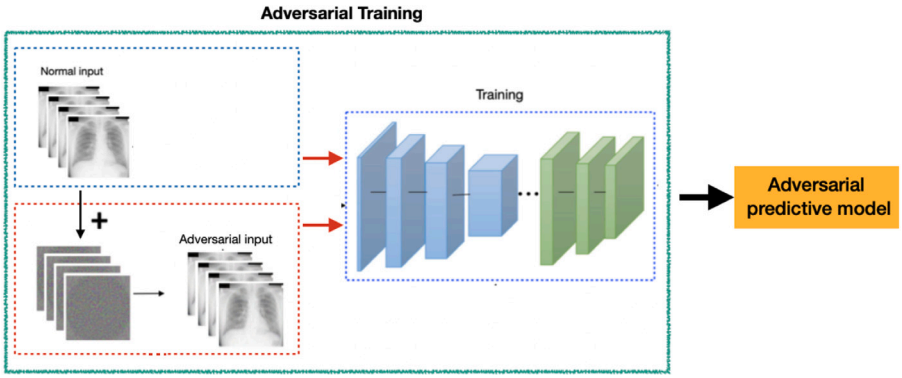

Fig. 1. The pipeline of adversarial training of DNNs.

3.1.2. White-box

It is when the adversaries have the complete information and accessibility to the trained model, network structure, training data, hyper parameters, weights and whatever is known to the network's trainer service. Most of the adversarial examples are generated with white-box access to the model and many of them are generated by calculating model gradients (Yuan et al., 2019).

3.2. Methods for generating adversarial examples

Gradient-based adversarial examples are the mostly used attack methods for generating adversarial examples. Their main objective is to generate minimum amount of perturbation ϵ to mis-classify original input images. With a trained model H, an adversarial example x^* can be generated, for original input image x with output label y, by solving an optimization problem

$$\min_x \|x^* - x\| \tag{1}$$

subject to

$$H(x) = y,$$
$$H(x^*) = y^*,$$
$$y^* \neq y$$

Therefore, this optimization minimizes the amount of perturbation α in $x^* = x + \epsilon$ while fooling the model H. For the first time Szegedy, et al. (2013) introduced L-BFGS attack model. For an input image x, their method finds a different image x^* which is similar to x under L_2 distance, but it is labeled differently by the classifier. They consider the problem as a constrained minimization problem and a binary search was used to find the important parameters. Since L-BGFS method was time consuming and impractical, Goodfellow et al. (2014), proposed another method called fast gradient sign method (FGSM) to generate adversarial examples. This model is optimized for the L_∞ distance metric, and second, designed to be fast instead of producing very close adversarial examples. Another attack that was optimized under L_0 distance was proposed by Papernot, McDaniel, Jha, Fredrikson, Celik, and Swami (2016) known as the Jacobian-based saliency map attack (JSMA). Computing the Jacobian matrix of the given sample x, the impact that each pixel has on the resulting classification can be modeled. By the saliency map, the most important pixel is picked and will be modified to increase the likelihood of a specific class. Carlini and Wagner (2017), modified the JSMA approach and design C&W attack which is effective against most of the adversarial detecting defense methods. To find the closest distance between the clean input to the decision boundary of adversarial examples, Moosavi-Dezfooli, Fawzi, and Frossard (2016), proposed DeepFool attack model. DeepFool is an untargeted attack model that is optimized for the L_2 distance metric. It is more efficient than L-BFGS and produces closer adversarial examples. Using this method, Moosavi-Dezfooli, Fawzi, Fawzi, and Frossard (2017) introduced another method called universal perturbation (UP) through which a group of images can be fooled. It showed that it can be efficiently generalized in popular DNN architectures such as ResNet and VGG. Furthermore, transferability over DNNs is one of the major challenges between the attackers. Transferability refers to the adversaries ability to attack a new model with adversarial examples that are generated by a different model. In this regard, Liu, Chen, Liu, and Song (2016) proposed a model-based ensembling attack for targeted adversarial examples since it has been shown to be harder to transfer targeted attacks on deep models than non-targeted adversarial attacks. Through Model-based ensembling attack, transferable adversarial examples can be generated to attack a black-box model. The above mentioned attacks are the most important methods for producing adversarial attacks.

4. Attacks to medical learning algorithms

Recent studies has shown that even state-of-the-art DNNs are significantly vulnerable to adversarial attacks on classification, segmentation of medical imaging and regression tasks.

4.1. Classification

Existing deep learning model's evaluation delve into generalizability and overfitting but insufficiently look deeply into model sensibility and vulnerability to variations of the input. For the first time, Paschali, Conjeti, Navarro, and Navab (2018) challenges the robustness of deep learning networks in medical imaging and investigate these state-of-the-art network vulnerability by utilizing adversarial examples. They also leverage these adversarial examples to benchmark model performance on clean, noisy and adversarially crafted data. The core idea is that both generalizability and robustness evaluation should take into account for evaluating a model. To this end, the authors compare a variety of architectures such as Inception V3 (IV3), Inception V4 (IV4), and MobileNet (MN) (Howard, et al., 2017) for skin lesion classification and UNet for whole brain segmentation. FGSM, DeepFool (DF) (Moosavi-Dezfooli et al., 2016) and saliency map attacks (SMA) (Papernot et al., 2016) methods for classification and dense adversarial generation (DAG) (Xie, Wang, Zhang, Zhou, Xie, & Yuille, 2017) method with varying degrees of perturbation and difficulty for semantic segmentation have been utilized for crafting adversarial examples. First of all, for classification task, it has been deduced that images distorted with noise are classified close to clean images while adversarial examples are pushed further towards other classes. Actually, addition of Gaussian noise only reduced classification confidence while almost all adversarial samples were incorrectly classified with high confidence. Therefore, adversarial examples are better suited for model robustness evaluation than noisy test images. Furthermore, they observed that for models with comparable performance on clean data

(IV4 and MN), significant differences in underlying data manifold result in a contrary trend in accuracy, sensitivity and robustness when attacked with FGSM. Finally, they conclude that despite all deviations in accuracy when different attacks applied, IV4 is preferred because of comparable generalizability and robustness amongst others. Hence, model depth seems to increase the robustness of classification models. For segmentation, they aim to evaluate the importance of skip connections in robustness. The results showed that DenseNet is the strongest model among others. Therefore, it has been deduced that dense blocks and skip connections improve both generalizability and robustness in segmentation tasks.

In addition to technical results obtained in Paschali et al. (2018), the highly effective performance of the adversarial attacks raised security concerns about medical DNNs. In a more extensive effort by Finlayson et al. (2018), they tried to show and declare the possible vulnerabilities in medical DNN models. The authors outlined the incentives for fraud against medical neural networks and the way these attacks can be accomplished. It is explained that the huge healthcare economy and the role of learning algorithms in medical reimbursement decisions and pharmaceutical and device approvals in near future are the most important stimulants for fraud. In addition they tried to show that deep neural networks are vulnerable to both black-box and white-box adversarial attacks. Both PGD and a naive patch attack on three baseline models were applied to classify diabetic retinopathy from retinal fundoscopy, pneumothorax from chest X-ray and melanoma from dermoscopic photographs. The results showed that both types of attacks are likely to be feasible, human imperceptible, and successful even in state-of-the-art medical classifiers (ResNet-50 model) disregarding the attackers amount of access to the network.

Now one important question is whether medical DNN models have different degrees of robustness compared with models for natural images. If so, where does this difference comes from? In a remarkable research Ma, et al. (2021) tried to answer these questions. They provide a comprehensive understanding of medical image adversarial attacks for the first time. FGSM, basic iterative method (BIM) (Kurakin, Goodfellow, & Bengio, 2016), PGD and C&W attacks on the same medical domain as in Finlayson et al. (2018) (i.e., fundoscopy, chest X-ray, and dermoscopy), with 2-class and multi-class datasets were applied. As it is shown in Fig. 2, the authors have also illustrated the concentration of DNNs in the attraction maps of the normal and adversarial examples. They showed that, in 2-class dataset classifications, with a small amount of perturbation ($\epsilon < \frac{1.0}{255}$) a medical DNN is more easily attacked compared with natural images such as those in CIFAR-10 and ImageNet that require $\epsilon > \frac{8.0}{255}$ perturbation for a successful attack. By increasing the classes in datasets the network is more vulnerable against adversarial attacks. The authors also referred to two reasons to explain this high amount of vulnerability against adversarial attacks including the complex biological texture of medical images and state-of-the-art DNNs which can be overparametrized for medical imaging tasks.

Pursuing comprehensive vulnerability analysis of deep learning approaches for classifying chest X-ray images into various disease categories, Taghanaki, Das, and Hamarneh (2018) extensively analyzed two of these deep neural networks when attacked with 10 different adversarial attacks. In contrast to previous methods with a single gradient-based attack, the authors applied various models of gradient-based, score-based and decision-based attacks on Inception-ResNet-v2 (Szegedy et al., 2017) and NasNet-large (Zoph, Vasudevan, Shlens, & Le, 2018) to evaluate their performance on chest X-ray images. They showed that, white-box gradient-based attacks were the most successful in fooling both machine and human while score-based and decision-based attacks were easily detectable by the human eyes and unsuccessful in white-box and partially successful in black-box situations. This may be the reason why Gradient-based attacks such as FGSM are preferred to assess the network's vulnerability. Moreover, it has been shown that average-pooling captures more global features that make the neural network more robust against attacks compared with max-pooling. Therefore, both state-of-the-art structures have proven to be non-resilient against gradient-based attacks and what makes difference is the pooling method.

For further assessment of CNN vulnerabilities, for the first time, Yilmaz (2020) investigated the security gap of mammographic image classifier against adversarial attacks. The author analyzed the similarity between benign and malicious images using structural similarity index method (SSIM) (i.e. a perception-based model used for measuring the similarity between two images) and applying FGSM attacks to trained CNNs. He studied the original image changes due to different perturbation coefficients as well. The generated adversarial samples have shown to mislead the model into incorrect prediction. By increasing the perturbation coefficient amount the CNN's accuracy reduces to less than 35% which is an alert to raise the awareness of radiologists and doctors.

In all the previous studies, adversarial attacks were input-dependent which means that a specific adversarial perturbation is used for each image mis-classification. Moreover most of the adversarial attacks are hard to implement and needs high computation costs. Recently, more realistic and strong attacks, with image agnostic perturbations have been introduced (Moosavi-Dezfooli et al., 2017) called universal adversarial perturbations (UAP). In this method, simple iterative algorithm were used to add small perturbation by an adversarial method such as FGSM for an input image. These newly invented attacks are more straightforward to be applied by the adversaries and difficult to detect. To evaluate the vulnerability of DNNs to UAPs, Hirano, Minagi, and Takemoto (2021), introduced a single perturbation UAP to induce performance degradation to classification networks. The authors focused on skin cancer, referable diabetic retinopathy, and pneumonia DNN-based medical image classification tasks. They showed that adversaries can deceive DNNs with small UAP attacks more easily and with lower cost. These types of attacks are imperceptible to human eye, approximately structure independent and have universal features for attacking DNNs. Both targeted and non-targeted UAPs achieved $> 80\%$ attack success rate. Moreover, the influence of adversarial training as a defense method has shown to be limited to non-targeted attacks and most of the recent defense methods have been failed. Therefore, it seems to be difficult to mitigate UAPs on medical DNNs.

As the chest X-ray images reveal various types of disease, automated diagnosis with deep learning networks have been of great interest between physicians and radiologists. Therefore, security of these models is considered as a crucial task. Rao, et al. (2020) studied different types of attacks and defenses for Thorax disease classification in chest X-rays. In this comparison study, the authors applied five types of attacks including FGSM, PGD, MIFGSM, DAA, and DII-FGSM. They conducted their experiments by attacking a single model and ensemble models and compared the DNNs performance against different adversaries. For single models, it has been shown that the area under curve (i.e., AUC is an attack evaluation metric from 0 to 1.) of FGSM gained the highest value compared to other models in white-box and black-box attacks, and FGSM outperformed all others most of the time. For ensemble models, FGSM has also shown highest AUC among others. Moreover, too much noise has been shown to weaken the adversarial examples transferability and reduces the success rate of black-box attacks.

Recently, deep learning algorithms have been successfully used to diagnose COVID-19 patients from the data obtained by medical IoT devices. These devices provide CT scanning or X-ray images, thermal cameras and face detection outputs. Existing researches show that DL networks used for COVID-19 prediction are vulnerable to adversarial attacks. For the first time Rahman, Hossain, Alrajeh, and Alsolami (2020) studied adversarial perturbations to these types of deep learning networks. They investigated six different DL applications applied to diagnose COVID-19 and presented multi-modal AEs on diversified COVID-19 diagnostic systems. The authors considered white-box, gray-box, and black-box attacks including FGSM, Deepfool, C&W and six

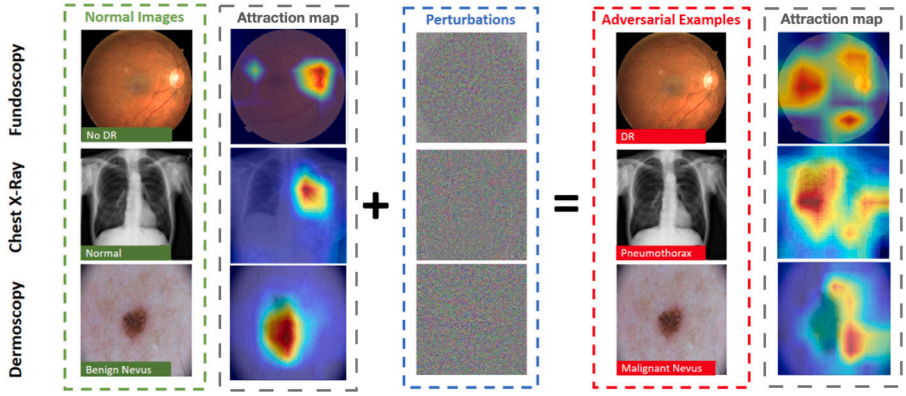

Fig. 2. Examples of adversarial attacks generated by PGD to fool DNNs trained on medical images. Each row shows normal images, attraction map and resulting adversarial examples of medical image data. The first row is related to diabetic retinopathy, the second row shows chest X-ray, and the third row shows the skin lesion images (Ma, et al., 2021).

other attack types on ResNet-101 with different kernel sizes (i.e., 1, 30, 300) and other state-of-the-art structures. The results showed that these DL networks are vulnerable to all types of attacks such as train and test data poisoning, model stealing, and evasion while there is no defensive approach.

Gongye, et al. (2020) have also studied the effects of adversarial attack as an active attack to deep learning algorithms for COVID-19 diagnosis from chest X-ray images. They tried PGD and FGSM attacks and showed that by FGSM, which are imperceptible by the human eye, the effective accuracy of the neural network reduces about 93% compared with a clean baseline model. The PGD attack degrades the accuracy even more (94.7%) and showed to be very successful.

On the other hand, rather than adversaries there are other ways of AE appearance in medical imaging that have not attracted much attention in literature. Vatian, et al. (2019) discussed about the possibility of AE appearance due to the inherent noise in the formation of high-tech medical images. The authors used a combination of UNet and the region proposal (RPN) with lung image database and Brain MRI dataset containing clinical data of glioma patients. The results showed that in all experiments, there is a quite large portion of images which are not recognized correctly by the CNN. Therefore, the probability of appearing AE when classifying by CNN of high-tech images arises which is because of the inherent noise in the image formation approaches. They also proposed an effective defense method that will be explained in Section 5.

As we mentioned above, sometimes generating adversarial examples are the results of real-world noise or ambiguous characteristics of the medical images. Another example of these types of mis-predictions occurs via the low quality fundos images while the goal is to diagnose diabetic retinopathy (DR). The major reason of the low quality in these images is the uneven exposure. Therefore, Cheng, et al. (2020) study this problem from the adversarial attack point of view and introduce a new attack called adversarial exposure attack. According to their new method, the adversarial images were generated by tuning image exposure to fool the DNN. They start with a method called multiplicative-perturbation-based exposure attack and improve it to generate more natural images by adversarial bracketed exposure fusion (BEF). The latter method concerns the exposure attack as an element-wise bracketed exposure fusion problem in the laplacian-pyramid space. In the next step, to make the attack transferable, they proposed the convolutional bracketed exposure fusion (CBEF) where the multiplicative fusion is extended to the element-wise convolution. BEF and CBEF are compared with six advanced additive-perturbation-based adversarial attacks as baseline methods. The results showed that BEF outperform others in image quality with 0.97 SSIM and CBEF outperform BEF and baseline attacks in transferability. They showed that their significantly successful attack with high quality images and significant transferability reveals the serious vulnerability of DNNs in DR automated diagnosis.

4.2. Segmentation

Hitherto, attacking classification models has been our main focus while attacking to segmentation DLs is of great importance. Generating adversarial examples to attack image segmentation models is harder than attacking classification models due to some inherent characteristics such as labeling each single pixel in segmentation instead of labeling the whole image in classification, complicated evaluation of the attack success and susceptibility of segmentation models to both image deformations and the image intensity variations. Therefore, there is a need to develop different approaches to generate adversarial examples that are capable of attacking segmentation models. To this end, Chen, Bentley, et al. (2019) introduced a novel method to attack segmentation CNNs using adversarial learning. They proposed to mix variational auto-encoder (VAE) and Generative adversarial networks (GAN) to generate images with deformations and appearance changes for attacking medical segmentation models. The authors apply their novel attack approach on CNN models such as UNet for abdominal organ segmentation in 2D CT images. The attack success is evaluated by a significant reduction in dice score (Bertels, Eelbode, Berman, Vandermeulen, Maes, Bisschops, & Blaschko, 2019) (i.e., commonly used metric for the evaluation of segmentation tasks in medical imaging) compared to ground truth segmentation. They showed that the attacking model results in 30% decrease in dice score on every organ. Although, attacking segmentation on the pancreas and kidneys has shown to be more difficult than the liver and spleen. On the other hand, as the intensity variations introduces shadows and artifacts, the segmentation model is more sensitive to this quantity. Therefore, their proposed model can be used to verify the CNN robustness if the generated adversarial examples are reasonable.

Another application of segmentation is to detect and localize brain tumor regions on X-ray or MRI images by brain tumor segmentation models. This segmentation algorithms assist physicians to identify abnormal regions faster which is critical for early tumor recognition. Many state-of-the-art CNNs have been developed for brain tumor segmentation such as V-Net and U-Net and MRI images have shown to be the most useful datasets for these networks. To label brain tumors, doctors and medical personnel use MRI images with different pixel intensity (i.e., different modalities) due to the unique characteristics of brain tumors in each patient. As the security of these medical neural networks are vital to the patients, Cheng and Ji (2020) studied the effects of a universal adversarial perturbation on brain tumor segmentation models and four different modalities. The authors utilized the MICCAI BraTS which is the largest publicly available dataset with MRI brain tumor images on U-Net model. Perturbations are generated according to Gaussian distribution. Therefore, the results showed that those modalities with the intensity distribution more similar to the

Table 1
Details of adversarial attack methods to medical imaging DNNs.

Reference	Attack	Dataset	Architecture	NN. type
Paschali et al.	FGSM, DF, SMA DF,SMA	skin lesion (Dermofit lib.) brain (OASIS)	IV3, IV4, MN SN, UN, DN	classification segmentation
Finlayson et al.	PGD, adv. patch	DR (Fundos) pneumothorax (X-ray) melanoma (skin images)	ResNet	classification
Ma et al.	FGSM, BIM, PGD, C&W	DR (Fundos) pneumothorax (X-ray) melanoma (skin images)	ResNet	classification
Taghanaki et al.	Gradient-based(5 types) Score-based(2 types) Decision-based(3 types)	pneumothorax (X-ray)	IV2-ResNet, Nasnet-Large	classification
Yilmaz et al.	FGSM	breast cancer (DDSM)	CNN	classification
Hirano et al.	UAP	skin lesion(ISIC2018) DR (OCT) pneumothorax (X-ray)	IV3, VGG, ResNet, IV2-ResNet, DenseNet	classification
Rahman et al.	FGSM, MI-FGSM, DF, L-BFGS, C&W, BIM, FB, PGD, JSMA, BD, MS, poisoning	COVID-19 (IoT outputs)	ResNet, YOLO, DarkNet, GRAD-CAM	classification
Gongye et al.	FGSM, PGD	COVID-19(X-ray)	ResNet	classification
Vatian et al.	Inherent AE Inherent AE	lung cancer (LIDC-IDRI) brain tumor (MRI)	UNET+ RPN	classification
Cheng et al.	Exposure AE (BEF, CBEF) FGSM, IFGSM, MIFGSM, TI versions	DR (Fundos)	ResNet, MN, Efficient Net	classification
Chen et al.	New method (intensity variance+ deformation)	Abdominal organs (CT)	UNet	segmentation
Cheng et al.	New method	brain tumor (MRI)	Ensemble NNs	segmentation
Li et al.	L_0, L_2(C&W) & L_∞(FGSM)	age prediction (MRI)	CNN, hybrid	regression

Gaussian distribution are more affected by the adversarial attack. It has been shown that the most severe performance degradation is observed when all the 4 modalities are attacked while if one modality is attacked the performance is not affected strongly.

4.3. Regression

In addition to classification and segmentation, which are the most applied deep learning algorithms in medical domain, regression models are also utilized and investigating their robustness against adversaries can be useful for the medical systems security and reliability. In this regards, Li, Zhang, et al. (2020) presented the first investigation about the vulnerabilities of regression-based predictions in medical image processing against adversarial attacks. The authors study the influence of adversarial attacks to CNNs and a hybrid DL model which predict an individual's age based on a 3D MRI brain image. They generate adversarial examples with l_0, l_2 and l_∞ constraints on the magnitude of perturbation. It has been shown that, both image-specific and a universal adversarial perturbation which is a single perturbation that can be effective on a large batch of images, are extremely effective on reducing the deep learning age prediction. Therefore, there is still significant concerns about robustness of DL to adversarial perturbations since a single perturbation may introduce significant bias into predictions. In Table 1, the details of the above mentioned attacks are summarized.

5. Defense methods against attacks to medical learning algorithms

Defense methods against adversarial attacks to medical imaging DNNs are discussed in this section and summarized in Table 2. These countermeasures include both mitigation and detection techniques.

5.1. Adversarial training

Adversarial training is one of the attack methods that can deceive the neural network with significantly high attack success rates (Kaviani & Sohn, 2021). Madry et al. (2017) found out that training a neural network with adversarial examples will also make it robust against first-order attack methods. As Goodfellow et al. (2014) suggested, in adversarial training clean images and adversarial examples are combined and the total loss can be defined as:

$$\tilde{J}(\theta, x, y) = \alpha J(\theta, x, y) + (1-\alpha) J(\theta, x^*, y) \tag{2}$$

in which x is the input to the model, y is the target associated with x, θ represents the parameters of the model, $J(\theta, x, y)$ is the cost used to train the neural network, and

$$x^* = x + \epsilon sign(\Delta_x J(\theta, x, y)). \tag{3}$$

PGD adversarial training is one of the most successful defense methods against adversarial attack that generates adversarial examples per epoch and it keeps the clean data accuracy high. As we explained in Section 4, Vatian, et al. (2019) investigate how instances of real high-tech medical images may generate adversarial examples due to their inherent noise. The authors explained three different ways of decreasing incorrectly recognized images in these networks where the most successful way is recognized to be adversarial training. It has been shown that by properly choosing activation functions for the layers (i.e., Bounded ReLU instead of ReLU) the number of incorrectly recognized images by the model will be decreased by $\sim 70\%$. Moreover, with augmentation of training dataset by Gaussian noise images this number will be decreased by $\sim 90\%$. The greatest reduction will occur by using adversarial training method by $\sim 95\%$. Therefore, it can be inferred that adversarial training techniques such as FGSM and JSM can provide the best robustness in high-tech medical image classification.

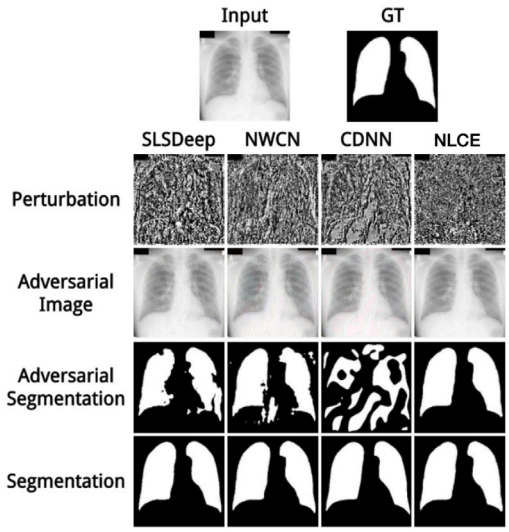

Fig. 3. Adversarial attacks on SLSDeep, NWCN, CDNN, and NLCE. The first row shows the normal input chest X-ray and its segmentation. Perturbation in the second row is generated by FGSM attack method and the third row shows adversarial examples. The fourth and fifth rows show segmentation of adversarial and input images. Experiments on both lung and skin lesion segmentation datasets have demonstrated that NLCE-Net outperforms other state-of-the-art biomedical image segmentation methods against adversarial attacks. (He, Yang, Li, Li, Chang, & Yu, 2019).

5.2. Pixel deflecting transform and adversarial training (PDT & adv_train)

In Section 4, we explained that Rao et al. studied different types of attacks on chest X-rays for Thorax disease diagnosis. In addition, the authors study two types of defense methods, including PGD adversarial training and pixel deflecting transform (PDT) (Prakash, Moran, Garber, DiLillo, & Storer, 2018). PDT randomly samples a pixel from adversarial examples, and replaces the pixel with another pixel that was selected from a small square neighborhood. When the average AUC of clean images is 0.87633, using PGD adversarial training keeps the AUC of all untargeted attacks over 0.8. When PDT is applied, AUC decreases to over 0.68 which means that it is robust against both black-box and white-box attacks but it is weaker than PGD on clean data. On the other hand, in their newly proposed method, the authors combined PGD adversarial training and PDT. The results showed that for low amount of perturbation PGD has the highest AUC among others, but by increasing the perturbation size, AUC sharply drops while AUCs of PDT and PDT & adv_train stay steady. However, PDT & adv_train outperforms PDT for all amount of perturbation. They concluded that the existing defense methods has poor performance for large perturbations while their new method solved this problem.

5.3. Non-local context encoder (NLCE)

For the first time He et al. (2019) discovered that all the CNN-based biomedical image segmentation models are sensitive to attack and introduced two factors which can improve defending against adversarial attacks: 1) global spatial dependencies and 2) global contextual information. Based on these findings, they proposed to add a robust module to the network called non-local context encoder (NLCE) module to model spatial dependencies and encode global contexts. The authors also designed a NLCE network (NLCE-Net) that is based on feature pyramid network (FPN) (Lin, Dollár, Girshick, He, Hariharan, & Belongie, 2017) and the NLCE module. The experiments on lung and skin lesion segmentation datasets showed that NLCE-Net is the most robust method among state-of-the-art CNNs against adversarial attacks with different amount of perturbation. NLCE-Net achieves high segmentation accuracy as well. As it is shown in Fig. 3, NLCE-Net was compared with neural networks such as SLSDeep (i.e., a skin lesion segmentation model (Sarker, et al., 2018)), NWCN (Hwang & Park, 2017), and UNET. Moreover, it turned out that NLCE module can be applied on all other CNNs and the results showed that models with NLCE module achieves significantly higher accuracy and robustness. This method seems to be perfectly robust against adversarial attacks specially when there is a lung segmentation model.

5.4. Detection methods

5.4.1. KD, LID, Qfeat and Dfeat

Although medical imaging DNNs are more vulnerable to adversarial attacks than real-world DNNs, medical adversarial examples can be detected more easily. In Ma, et al. (2021), the authors applied four detection methods on medical DNNs called Kernel density (KD) (Feinman, Curtin, Shintre, & Gardner, 2017), local intrinsic dimensionality (LID) (Ma, et al., 2018), Deep features (DFeat) and quantized features (QFeat) (Lu, Issaranon, & Forsyth, 2017). In KD, it is assumed that the adversarial examples are placed in a more sparse regions of the data submanifold while clean samples from the same class are placed on the data manifold. In LID, the dimensional characteristics of adversarial subspaces in the neighborhood of adversarial examples are provided as a measurement. They applied four attacking methods including FGSM, PGD, BIM, and C&W to generate adversarial examples. Results showed that all 4 detection methods provide a very robust performance while KD achieves the highest AUC of 99% against all attacks for three studied datasets. On the other hand, these detection methods obtain less than 80% AUC for real-world images. To answer why medical adversarial examples are more detectable, they claimed that adversarial features of medical images are almost linearly separable from normal features while in real images these two types of features are very similar. They proved their claim by visualizing t-SNE 2D embedding of the adversarial and normal features (Fig. 4). Although the performance is quite high, since their method rely on estimating the density of adversarial examples, the effectiveness of the method is limited to the certain types of popular attacks.

5.4.2. Unsupervised anomaly detection

In another effort to detect adversarial attacks on medical images in Li and Zhu (2020), the authors proposed an unsupervised learning approach. The authors claimed that their new method can be utilized as a separate module in any deep learning-based medical imaging system and improve the network's robustness. Firstly, to learn the detection module, the CNN classifier is trained with clean images to extract high-level features. When a new image is inserted to the trained CNN classifier, the features will be extracted as the input of the detection module and if it is adversary it will be detected and stopped to enter the classification layer. The authors specifically used uni-modal multivariate Gaussian model (MGM) as the detector. Since this method is based on unsupervised learning, it can be applied against diverse adversarial attacks. To evaluate their method, it has been used against four types of attacks including FGSM, PGD, MIM (Dong, et al., 2018) and BIM attacks in both black-box and white-box situations and compared with isolation forest (ISO) (Wang, Peng, Lu, Lu, Bagheri, & Summers, 2017) and one-class SVM (OCSVM) (ISIC, 2019) detection techniques on X-ray image classifiers. The results showed that although in the black-box and white-box settings all the detection methods demonstrated robust performance against attacks, MGM has the best performance recognized by the highest area under ROC (AUROC) curve values (i.e., determine the optimal cut of values for classification decision based on the class probabilities). The interesting point in their method is that this robustness is achieved where the architecture of the CNN is not recognized by the attacker. Moreover, in this method, in contrast with most of the previous approaches, the neural network's performance with a mixture of clean and adversarial examples is higher than the classifier on clean dataset. The key point is that the detection module can detect both adversarial examples and those clean samples that can be problematic for the image classification.

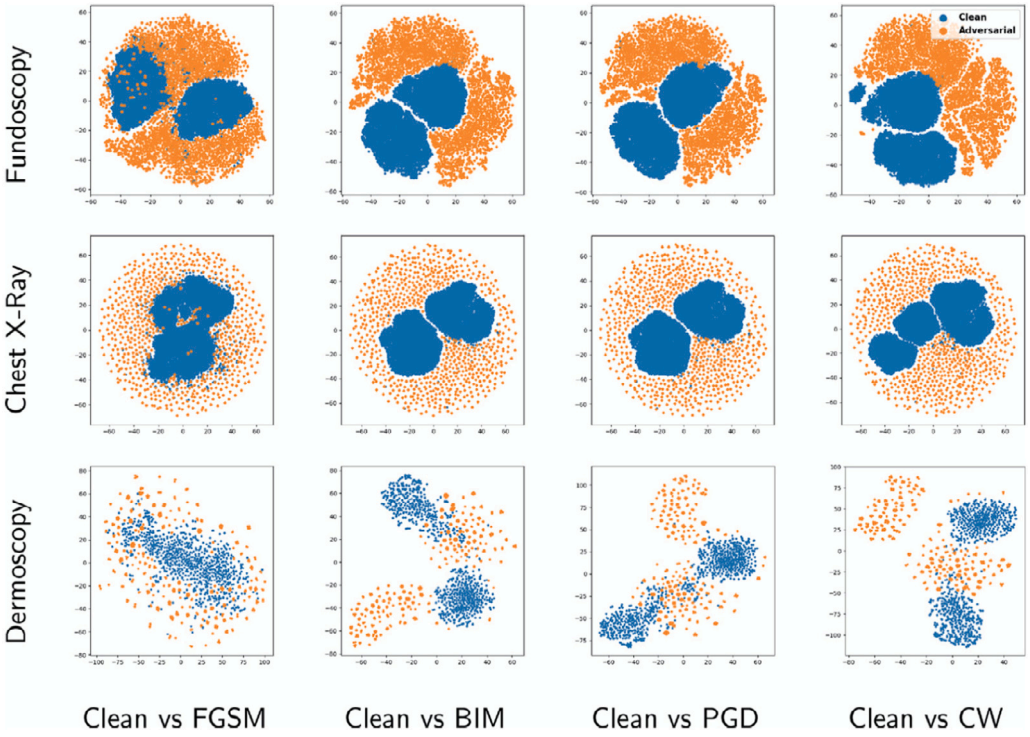

Fig. 4. Visualization of t-SNE 2D embeddings of adversarial and normal features, extracted from the second last dense layer of the DNN models. Each row is a dataset, each column is an attack. Adversarial features of medical images (orange) are almost linearly separable from normal features (blue) while in real images these two types of features are very similar (Ma, et al., 2021).

Table 2
Details of defense methods against adversarial attacks to medical imaging DNNs.

Defense method	Attack type	Dataset	Model architecture	NN. type	Reference
Adversarial training	FGSM, JSMA	lung images (CT) brain MRI	UNet+ RPN	classification	Vatian et al.
PDT & adv_train	FGSM, PGD, MIFGSM, DAA, DII-FGSM	pneumothorax (X-ray)	DenseNet, ResNet, VGG, IV3	classification	Rao et al.
NLCE	FGSM	lung (JSRT) skin lesion(ISBI2016)	SLSDeep, NWCN, UNet, InverNet, CDNN, ResNet	segmentation	He et al.
KD & LID	FGSM, PGD, BIM C&W	DR (Fundos) pneumothorax (X-ray) melanoma (skin images)	ResNet	classification	Ma et al.
Unsupervised anomaly detection	FGSM, BIM, MIM, PGD	pneumothorax (X-ray)	DenseNet, ResNet	classification	Li et al.
SSAT & UAD	FGSM, PGD, C&W	DR (OCT)	ResNet	classification	Li et al.

5.5. Semi-supervised adversarial training and unsupervised adversarial detection (SSAT & UAD)

Li, Pan, and Zhu (2020), proposed a hybrid approach to present a robust medical imaging AI framework based on semi-supervised adversarial training and unsupervised adversarial detection and a newly invented measure for the systems adversarial risk. Their method is based on two important challenges in detecting adversarial samples which are having small set of labeled images in medical tasks and having non-efficient detection methods against unseen new attacks. In this method both labeled and unlabeled data have been used for SSAT to improve robustness. They also proposed a new adversarial risk measure based on the classification situation and being accepted or rejected by UAD. They evaluated their approach using OCT imaging dataset and compared it with other defense methods such as natural training (Chen, Liang, et al., 2019) and adversarial training with cross entropy loss while they are attacked with FGSM, PGD, and C&W. They showed that SSAT method can significantly outperform other defense methods while maintaining a similar or better accuracy compared with clean data classification and UAD can correctly filter out a great amount of OOD adversarial samples. According to their new adversarial risk measure, SSAT against weak attacks gives rise to a low adversarial risk. On the other hand, for stronger attacks, the adversarial risk profoundly decreases by adding UAD . This method seems to provide a robust model against strong, heterogeneous attacks with low amount of labeled medical dataset.

6. Discussion

Adversarial attacks have shown significant potential to be feasible even for state-of-the-art DNNs, regardless of the attackers amount of accessibility to the model and being perceptible to human eye. Compared to other domains of computer vision, medical DNNs are very fragile against adversarial attacks. Adversarial samples with a limited amount of perturbation may fool state-of-the-art medical systems which show great performance on clean data. One of the main questions that the

previous researchers tried to answer is that *what makes these medical DNNs significantly weak against adversarial attacks?* and *what make them robust?*

6.1. Medical image learning vulnerabilities

Label scarcity. Medical image dataset for training medical DNNs in contrast with real-world images are very rare due to important reasons such as privacy concerns and having no universally shared mechanisms for sharing medical data. Another reason is that assigning labels to medical images is a very time and energy consuming task while, most of the time, the ground truth images in medical datasets are ambiguous and controversial even with physicians and radiologists. Therefore, due to lack of enough amount of data, memorizing and overfitting occurs in medical DNNs which significantly affect generalizability and making the network vulnerable against adversarial attacks.

Image characteristics. One of the most important sources of vulnerability to adversarial attacks in medical image learning networks are the unique characteristics of medical images which are utilized to train the neural network. Some of these security threatening items are listed below.

- Fore- and backgrounds similarities in medical image make it difficult for the network to learn discriminating features in regards to clean and adversarial samples classification. This is due to highly standardized images with well-established exposure and high quality. This will also eliminate the need for the attackers to change their attack standards and parameters since there are similar characteristics in all images such as lighting and positional status of each body organ. This potentially makes imaging AI systems more susceptible to even simple attacks with low amount of perturbation.
- Camera exposure effects in retinal fundos images have shown to be potential to make the benign images resembling adversarial samples. These perturbations mislead the DNNs with significantly high transferability. Although it has not been exactly proven but it reveals potential threats to DNN-based medical imaging systems.
- Although it has been shown that noisy data does not show the same effect on the network as adversarial samples, inherent noises of high-tech medical images are going to be serious threats for AI based predictions. These noises have shown complex behavior in various equipment such as CT and MRI images based on different parameters such as equipment parameters, patient parameters and experiment parameters. These noises have shown to work as AE when being analyzed with neural networks.
- It has been highly proven that even small image intensity changes can be regarded as adversarial attacks to segmentation CNNs. This makes an opportunity for the attackers to design specific attacks that can fool segmentation models. It is shown that even deformation does not have similar effects on CNNs performance.
- In medical images, due to their complex textures, the DNNs attention may be payed more on areas of the image that are not important for the classification or diagnosis. Therefore, the inherent characteristics of medical images make them more vulnerable against adversarial attacks compared with real-world images (Ma, et al., 2021) .

Lack of diversity. As we mentioned previously, there are insufficient shared amount of labeled medical images that have been utilized to train medical neural networks. There are also limited amount of learning models which are similar between almost all medical computer vision tasks with the same architecture. Therefore, lack of popular architectural and dataset diversity make it easier to threaten the neural networks security since designing universal or transferable adversarial attacks has made it straight forward to attack any medical system.

State-of-the-art DNNs. Highly efficient deep neural networks are basically designed for large-scale real-world image analysis. Therefore, this will cause the neural network to become overparametrized that affect the generalizablity and consequently make the neural network more vulnerable against adversarial attacks.

Segmentation vs. classification models. Segmentation methods are more vulnerable to attack compared with classification models due to these reasons. First of all, in semantic segmentation a label is assigned to each pixel in contrast with classification in which a whole image have a certain label. Second, evaluating the attack success rate in segmentation models is not straight forward. Third, a segmentation image can be easily fooled with a small change in the image intensity, and forth, these models usually does not have high-quality image samples which result in overfitting. Despite all these facts, there is not a solid research on segmentation models security against adversarial attacks.

6.2. Attack transferability and universality

Transferability and universality have shown to be among two important factors which result in more effective attacks with higher risk of DNNs mis-classification. Several previous studies have focused on the network's vulnerability against naive normal adversarial attacks. These attacks have limited applications in specific DNN models and can be detected by recently invented defense methods. Recently, the main challenge of researchers in computer vision area is to design adversarial attacks which are both model and image agnostic and can be utilized on almost all medical learning tasks. They are more realistic high risk attacks that can induce performance failure to DNNs. In addition defense methods and detection techniques against universal attacks are still in their initial stages, although the DNNs vulnerability to adversarial samples alerts for new defense strategies against universal attacks since these attacks deceive DNNs more easily and at lower costs. Evaluating transferability has shown to be an important measure in previous studies while it determines the attacks potential to affect other models in a black-box status. According to Cheng, et al. (2020), attacks with high amount of transferability should sacrifice the adversarial image quality while the attack success rate is kept sufficiently high (100%). Nevertheless, applying unsupervised learning has shown to be useful for improving networks transferability. Therefore, transferability and universality of the attacks seems to be two important concerns and challenges among researchers in this area.

6.3. Defense methods deficiencies

Although most of the defense methods increase the robustness of medical learning models with different approaches, there are still big concerns remain about using computer vision methods instead of human. Some shortcomings of recent defense methods are listed below.

- Although some detection methods, on specific models and datasets, even improve the performance of the system on clean data, in most of the previously proposed defense techniques, performance maybe compromised by the highly reported trade-off between accuracy and robustness.
- Considering that recently invented attacks are effective on all DNN models with various datasets and different tasks, there is not sufficient research on defending medical systems against powerful universal perturbations. While previous defense methods are mostly limited on certain types of attacks, UAPs and transferable unseen attacks are going to be serious threats for medical systems.
- Despite the fact that adversarial training has been one of the most powerful defense methods against adversarial attacks, it has some limitations when dealing with medical systems. In this method, the training dataset is augmented with adversarial images while this large amount of medical image data can significantly impact the classification accuracy. The problem is that DNNs are basically designed for real-world images and they require large amount of labeled data whereas medical datasets have usually small amount of data.

6.4. Effective parameters on DNNs robustness

According to the studied papers, we can infer that there are some important factors that can be tuned to improve the network's robustness against adversarial attacks. These effective factors can be related to image characteristics, structural properties of the DNNs or datasets and the learning algorithms:

- *Global spacial dependencies:* Global spatial dependencies have proven to be crucial in defending against adversaries in segmentation models. Capturing this quantity in an image of a segmentation model means to find all highly correlated pixels in the whole image, that affect the prediction at a single pixel. In other words, a wrong label that is given to a pixel result in the spread of the incorrect loss to all related pixels by back-propagation. Unfortunately, even popular CNNs are incapable of capturing long-range dependencies despite the fact that stacked convolution operations can capture long-range dependencies but result in gradient vanishing (He et al., 2019).
- *Global contextual information:* This quantity is another important parameter in medical segmentation models which can improve the network's robustness since the human body configuration is symmetric and stable. Therefore, the same amount of perturbations have to be inserted to the geometrically related organs which result in enhancing perturbation. Most of the known CNNs do not make use of this parameter.
- *Degrees of perturbation:* It has always been a challenge for the attackers to insert a sufficient amount of perturbation into the images in a way that it fools the neural network with high attack success rate while the image quality is not affected as much as it is perceptible to human eyes. Therefore, with low amount of perturbation the attack is less likely to be successful. Nevertheless, it has been shown that medical imaging DNNs are vulnerable to even very small amount of perturbations (Ma, et al., 2021). Moreover, in Chen, Bentley, et al. (2019), the authors concluded that the segmentation models are more vulnerable to intensity variations than image deformation.
- *Combine adversarial training with other methods:* Although adversarial training itself has some shortcomings in defending medical DNNs due to small datasets, as can be seen in previous defense methods, it can be modified and become more efficient by adding other methods such as pixel deflecting transform (Rao, et al., 2020) and using semi- or un-supervised training (Li, Pan, et al., 2020). This combinations has made adversarial training a strong tool for enhancing the networks robustness.
- *Non-similarity in perturbation and modalities intensity distribution in segmentation:* In Cheng and Ji (2020), it has been indicated that in MRI image modalities, those that have more similar distributions to the perturbation distribution, are more vulnerable against adversarial attacks. Therefore, having knowledge of the more popular perturbations among the adversaries, the image intensity distribution can be selected in a way that the probability of being affected by the attacks is decreased.
- *More labeled data:* By increasing the number of correctly annotated data by the experts, on one hand, DNNs are less prone to overfitting and accordingly more robust against adversarial attacks. On the other hand, adversarial training as a defense method can be more efficient in enhancing the network security.
- *Network parameters, structures and algorithms:* Other than the discussed effective robustness parameters which are mostly related to the input images and the quality of the adversaries techniques, the neural network characteristics and the learning algorithms also affects the DNN's performance against adversarial attacks. As an example, according to Taghanaki et al. (2018), modifying pooling operations in classification DNNs affects the neural network's robustness. It has been shown that average-pooling captures more global features that make the network more robust against attacks compared with max-pooling. On the other hand in Vatian, et al. (2019), changing the activation function has improved the number of correctly recognized images by the CNN. Therefore, network parameters are also seem crucial in their resilience against adversarial attacks.

6.5. Future directions

Adversarial attacks have proven to play a crucial role in evaluating the deep learning networks vulnerability. Therefore improving these attacks may help scientists to fix the shortcomings and plan for more efficient and secure medical learning systems. Following the recent trend, in near future, transferability of the attack methods and visibility of the perturbation while they strongly affect the network's predictions continues to be the main challenges. Moreover, more attack strategies on segmentation models and breaking their resilience are going to be considered. On the other hand, to provide more reliable medical learning systems, different ways of easily producing medical annotated images and producing larger datasets have to be maintained. Another option that can decrease the network's vulnerability is to change the neural network's architecture which has not been studied yet. In a recent paper (Kaviani & Sohn, 2020), a mitigation technique against backdoor attacks (Kaviani & Sohn, 2021) is proposed, in which the fully-connected layers of the perceptron are altered to scale-free connections. The results showed a significant improvement in the neural networks robustness against backdoor adversaries. These kinds of structure modification and the effect of scale-free connections in DNN's robustness against adversarial attacks has to be studied in the future as well.

7. Conclusion

In this paper we reviewed the recent methods of generating adversarial examples to attack medical imaging deep learning networks and the defense approaches to detect and mitigate these perturbations. Attacks and defenses on both classification and segmentation models have been considered and the neural network's effective parameters on resilience and vulnerabilities have been explored. Since medical images have specific characteristics and medical imaging system's security are of great importance in intelligent disease diagnosis, more investigations have to be done to improve medical DNN technology related to performance on accuracy, precision and reliability.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRFK) funded by the Ministry of Education (2018R1D1A1B07041981).

References

Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine*, *1*(1), 1–8.

Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A., & Mougiakakou, S. (2016). Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE Transactions on Medical Imaging*, *35*(5), 1207–1216.

Bertels, J., Eelbode, T., Berman, M., Vandermeulen, D., Maes, F., Bisschops, R., et al. (2019). Optimizing the dice score and jaccard index for medical image segmentation: Theory and practice. In *International conference on medical image computing and computer-assisted intervention* (pp. 92–100). Cham: Springer.

Bui, A. A., & Taira, R. K. (Eds.), (2009). *Medical imaging informatics*. Springer Science & Business Media.

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 Ieee symposium on security and privacy* (pp. 39–57).

Casamitjana, A., Puch, S., Aduriz, A., & Vilaplana, V. (2016). 3D convolutional neural networks for brain tumor segmentation: a comparison of multi-resolution architectures. In *International workshop on brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries* (pp. 150–161). Cham: Springer.

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., & Rueckert, D. (2019). Intelligent image synthesis to attack a segmentation CNN using adversarial learning. In *International workshop on simulation and synthesis in medical imaging* (pp. 90–99). Cham: Springer.

Chen, H. Y., Liang, J. H., Chang, S. C., Pan, J. Y., Chen, Y. T., Wei, W., et al. (2019). Improving adversarial robustness via guided complement entropy. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4881–4889).

Cheng, G., & Ji, H. (2020). Adversarial perturbation on MRI modalities in brain tumor segmentation. *IEEE Access*, *8*, Article 206009-206015.

Cheng, Y., Juefei-Xu, F., Guo, Q., Fu, H., Xie, X., Lin, S. W., et al. (2020). Adversarial exposure attack on diabetic retinopathy imagery. arXiv preprint arXiv:2009.09231.

Dalvi, N., Domingos, P., Sanghai, S., & Verma, D. (2004). Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 99–108). August.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., et al. (2018). Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 9185–9193).

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118.

Feinman, R., Curtin, R. R., Shintre, S., & Gardner, A. B. (2017). Detecting adversarial samples from artifacts. arXiv preprint arXiv:1703.00410.

Finlayson, S. G., Chung, H. W., Kohane, I. S., & Beam, A. L. (2018). Adversarial attacks against medical deep learning systems. arXiv preprint arXiv:1804.05296.

Gongye, C., Li, H., Zhang, X., Sabbagh, M., Yuan, G., Lin, X., et al. (2020). New passive and active attacks on deep neural networks in medical applications. In *Proceedings of the 39th international conference on computer-aided design* (pp. 1–9).

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Greenspan, H., Ginneken, B. Van., & Summers, R. M. (2016). Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, *35*(5), 1153–1159.

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., et al. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, *29*(8), 1836–1842.

Havaei, M., Guizard, N., Larochelle, H., & Jodoin, P. M. (2016). Deep learning trends for focal brain pathology segmentation in MRI. In *Machine learning for health informatics* (pp. 125–148). Cham: Springer.

He, X., Yang, S., Li, G., Li, H., Chang, H., & Yu, Y. (2019). Non-local context encoder: Robust biomedical image segmentation against adversarial attacks. In *Proceedings of the AAAI conference on artificial intelligence, vol. 33* (01), (pp. 8417–8424).

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026–1034).

Hesamian, M. H., Jia, W., He, X., & Kennedy, P. (2019). Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of Digital Imaging*, *32*(4), 582–596.

Hirano, H., Minagi, A., & Takemoto, K. (2021). Universal adversarial attacks on deep neural networks for medical image classification. *BMC Medical Imaging*, *21*(1), 1–13.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Huang, G., Liu, Z., Maaten, L. Van. Der., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4700–4708).

Hwang, S., & Park, S. (2017). Accurate lung segmentation via network-wise training of convolutional networks. In *Deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 92–99). Cham: Springer.

ISIC (2019). *The international skin imaging collaboration*. (https://www.isic-archive.com/).

Kaviani, S., & Sohn, I. (2020). Study of scale-free structures in feed-forward neural networks against backdoor attacks. *ICT Express*.

Kaviani, S., & Sohn, I. (2021). Defense against neural trojan attacks: A survey. *Neurocomputing*, *423*, 651–667.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C., Liang, H., Baxter, S. L., et al. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, *172*(5), 1122–1131.

Kleesiek, J., Urban, G., Hubert, A., Schwarz, D., Maier-Hein, K., Bendszus, M., et al. (2016). Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. *NeuroImage*, *129*, 460–469.

Kulikowski, C. A. (1997). Medical imaging informatics: challenges of definition and integration.

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.

Li, X., Pan, D., & Zhu, D. (2020). Defending against adversarial attacks on medical imaging AI system, classification or detection? arXiv preprint arXiv:2006.13555.

Li, Y., Zhang, H., Bermudez, C., Chen, Y., Landman, B. A., & Vorobeychik, Y. (2020). Anatomical context protects deep learning from adversarial perturbations in medical imaging. *Neurocomputing*, *379*, 370–378.

Li, X., & Zhu, D. (2020). Robust detection of adversarial attacks on medical images. In *2020 IEEE 17th international symposium on biomedical imaging* (pp. 1154–1158).

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117–2125).

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88.

Liu, Y., Chen, X., Liu, C., & Song, D. (2016). Delving into transferable adversarial examples and black-box attacks. arXiv preprint arXiv:1611.02770.

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., et al. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, *1*(6), e271–e297.

Lo, S. C., Lou, S. L., Lin, J. S., Freedman, M. T., Chien, M. V., & Mun, S. K. (1995). Artificial convolution neural network techniques and applications for lung nodule detection. *IEEE Transactions on Medical Imaging*, *14*(4), 711–718.

Lu, J., Issaranon, T., & Forsyth, D. (2017). Safetynet: Detecting and rejecting adversarial examples robustly. In *Proceedings of the IEEE international conference on computer vision* (pp. 446–454).

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S., Schoenebeck, G., et al. (2018). Characterizing adversarial subspaces using local intrinsic dimensionality. arXiv preprint arXiv:1801.02613.

Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., et al. (2021). Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, *110*, Article 107332.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89–94.

Moosavi-Dezfooli, S. M., Fawzi, A., Fawzi, O., & Frossard, P. (2017). Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1765–1773).

Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2574–2582).

Nie, D., Zhang, H., Adeli, E., Liu, L., & Shen, D. (2016). 3D deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. In *International conference on medical image computing and computer-assisted intervention* (pp. 212–220). Cham: Springer.

Panayides, A. S., Amini, A., Filipovic, N. D., Sharma, A., Tsaftaris, S. A., Young, A., et al. (2020). Ai in medical imaging informatics: Current challenges and future directions. *IEEE Journal of Biomedical and Health Informatics*, *24*(7), 1837–1857.

Papanicolas, I., Woskie, L. R., & Jha, A. K. (2018). Health care spending in the United States and other high-income countries. *Jama*, *319*(10), 1024–1039.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy* (pp. 372–387).

Paschali, M., Conjeti, S., Navarro, F., & Navab, N. (2018). Generalizability vs. robustness: investigating medical imaging networks using adversarial examples. In *International conference on medical image computing and computer-assisted intervention* (pp. 493–501). Cham: Springer, September.

Prakash, A., Moran, N., Garber, S., DiLillo, A., & Storer, J. (2018). Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8571–8580).

Qayyum, A., Qadir, J., Bilal, M., & Al-Fuqaha, A. (2020). Secure and robust machine learning for healthcare: A survey. arXiv preprint arXiv:2001.08103.

Qin, Y., Zheng, H., Huang, X., Yang, J., & Zhu, Y. M. (2019). Pulmonary nodule segmentation with CT sample synthesis using adversarial networks. *Medical Physics*, *46*(3), 1218–1229.

Rahman, A., Hossain, M. S., Alrajeh, N. A., & Alsolami, F. (2020). Adversarial examples-security threats to COVID-19 deep learning systems in medical IoT devices. *IEEE Internet of Things Journal*.

Rao, C., Cao, J., Zeng, R., Chen, Q., Fu, H., Xu, Y., et al. (2020). A thorough comparison study on adversarial attacks and defenses for common thorax disease classification in chest X-rays. arXiv preprint arXiv:2003.13969.

Redmon, J., & Farhadi, A. (2017). YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7263–7271).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Cham: Springer.

Sarker, M. M. K., Rashwan, H. A., Akram, F., Banu, S. F., Saleh, A., Singh, V. K., et al. (2018). Slsdeep: Skin lesion segmentation based on dilated residual and pyramid pooling networks. In *International conference on medical image computing and computer-assisted intervention* (pp. 21–29). Cham: Springer.

Shen, W., Zhou, M., Yang, F., Yang, C., & Tian, J. (2015). Multi-scale convolutional neural networks for lung nodule classification. In *International conference on information processing in medical imaging* (pp. 588–599). Cham: Springer.

Shin, H. C., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., et al. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, *35*(5), 1285–1298.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence, vol. 31*. (1).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *281*, In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2818–2826).

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., et al. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

Taghanaki, S. A., Das, A., & Hamarneh, G. (2018). Vulnerability analysis of chest x-ray image classification against adversarial attacks. In *Understanding and interpreting machine learning in medical image computing applications* (pp. 87–94). Cham: Springer.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., et al. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE Transactions on Medical Imaging*, *35*(5), 1299–1312.

Vatian, A., Gusarova, N., Dobrenko, N., Dudorov, S., Nigmatullin, N., Shalyto, A., et al. (2019). Impact of adversarial examples on the efficiency of interpretation and use of information from high-tech medical images. In *2019 24th Conference of open innovations association* (pp. 472–478). IEEE.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2097–2106).

Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., & Yuille, A. (2017). Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 1369–1378).

Yilmaz, I. (2020). Practical fast gradient sign attack against mammographic image classifier. arXiv preprint arXiv:2001.09610.

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*, *30*(9), 2805–2824.

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8697–8710).