Adversarial attack vulnerability of medical image analysis systems: Unexplored factors

Bortsova, G.; González-Gonzalo, C.; Wetstein, S.C.; Dubost, F.; Katramados, I.; Hogeweg, L.; Liefers, B.; van Ginneken, B.; Pluim, J.P.W.; Veta, M.; Sánchez, C.I.; de Bruijne, M.

DOI
10.1016/j.media.2021.102141

Publication date
2021

Document Version
Final published version

Published in
Medical Image Analysis

Medical Image Analysis 73 (2021) 102141

Contents lists available at ScienceDirect

Medical Image Analysis

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

Adversarial attack vulnerability of medical image analysis systems: Unexplored factors

Gerda Bortsova[a,1,*], Cristina González-Gonzalo[b,c,1], Suzanne C. Wetstein[d,1], Florian Dubost[a], Ioannis Katramados[e], Laurens Hogeweg[e], Bart Liefers[b,c], Bram van Ginneken[f], Josien P.W. Pluim[d], Mitko Veta[d,2], Clara I. Sánchez[b,c,g,2], Marleen de Bruijne[a,h,2]

[a] Biomedical Imaging Group Rotterdam, Department of Radiology and Nuclear Medicine, Erasmus MC, The Netherlands
[b] A-Eye Research Group, Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboudumc, Nijmegen, The Netherlands
[c] Donders Institute for Brain, Cognition and Behaviour, Radboudumc, Nijmegen, The Netherlands
[d] Medical Image Analysis Group, Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands
[e] Intel Corporation, The Netherlands
[f] Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboudumc, Nijmegen, The Netherlands
[g] Department of Ophthalmology Radboudumc, Nijmegen, The Netherlands
[h] Department of Computer Science, University of Copenhagen, Denmark

ARTICLE INFO

Article history:
Received 17 October 2020
Revised 10 June 2021
Accepted 17 June 2021
Available online 18 June 2021

Keywords:
Adversarial attacks
Medical imaging
Deep learning
Cybersecurity

ABSTRACT

Adversarial attacks are considered a potentially serious security threat for machine learning systems. Medical image analysis (MedIA) systems have recently been argued to be vulnerable to adversarial attacks due to strong financial incentives and the associated technological infrastructure.

In this paper, we study previously unexplored factors affecting adversarial attack vulnerability of deep learning MedIA systems in three medical domains: ophthalmology, radiology, and pathology. We focus on adversarial black-box settings, in which the attacker does not have full access to the target model and usually uses another model, commonly referred to as surrogate model, to craft adversarial examples that are then transferred to the target model. We consider this to be the most realistic scenario for MedIA systems. Firstly, we study the effect of weight initialization (pre-training on ImageNet or random initialization) on the transferability of adversarial attacks from the surrogate model to the target model, i.e., how effective attacks crafted using the surrogate model are on the target model. Secondly, we study the influence of differences in development (training and validation) data between target and surrogate models. We further study the interaction of weight initialization and data differences with differences in model architecture. All experiments were done with a perturbation degree tuned to ensure maximal transferability at minimal visual perceptibility of the attacks.

Our experiments show that pre-training may dramatically increase the transferability of adversarial examples, even when the target and surrogate's architectures are different: the larger the performance gain using pre-training, the larger the transferability. Differences in the development data between target and surrogate models considerably decrease the performance of the attack; this decrease is further amplified by difference in the model architecture. We believe these factors should be considered when developing security-critical MedIA systems planned to be deployed in clinical practice. We recommend avoiding using only standard components, such as pre-trained architectures and publicly available datasets, as well as disclosure of design specifications, in addition to using adversarial defense methods. When evaluating the vulnerability of MedIA systems to adversarial attacks, various attack scenarios and target-surrogate differences should be simulated to achieve realistic robustness estimates. The code and all trained models used in our experiments are publicly available.[3]

* Corresponding author.
E-mail address: gerdabortsova@gmail.com (G. Bortsova).
[1] The first three authors contributed equally to this work.
[2] The last three authors contributed equally to this work.

https://doi.org/10.1016/j.media.2021.102141

1. Introduction

Deep learning (DL) has been shown to achieve close or even superior performance to that of experts in many medical image analysis (MedIA) applications, including in ophthalmology (Gulshan et al., 2016; Ting et al., 2017), radiology (Rajpurkar et al., 2017), and pathology (Bejnordi et al., 2017; Bulten et al., 2020; Wetstein et al., 2020). This has created an opportunity to automate certain medical tasks and integrate DL systems in clinical settings (Abràmoff et al., 2018; Murphy et al., 2020; GE Reports, 2019). However, a threat to DL systems is posed by so-called *adversarial attacks* (Szegedy et al., 2013). Such attacks apply a carefully engineered, subtle perturbation to the input of the target model to cause misclassification. Those perturbed inputs, referred to as *adversarial examples*, have been shown effective in forcing state-of-the-art systems to produce incorrect predictions (Goodfellow et al., 2014; Madry et al., 2017).[3]

Adversarial attacks are not the only kind of malicious manipulation of input to DL models that changes their predictions. Adversarial attacks are manipulations that aim to preserve the semantic contents of a given image, e.g., whether it is healthy or diseased, while changing the prediction of the network for the image. Apart from this type of attack, images can also be manipulated to change their content: for example, signs of disease can be removed from a diseased image or added to a healthy image (Xia et al., 2020; Sun et al., 2020; Baumgartner et al., 2018; Becker et al., 2019), which, in turn, can change network predictions. However, developing these synthetically changed images remains challenging (Xia et al., 2020), as it is hard to guarantee they look realistic, hard to control which image structures are changed, and these algorithms may be difficult to train and require large training datasets. In contrast, adversarial examples generated by adding noise of bounded, small magnitude are guaranteed to look realistic and do not induce any unpredictable changes in the image. Therefore, we consider adversarial attacks to be a more feasible, and thus more likely type of attack on MedIA systems, which is why we have limited our scope to adversarial attacks.

1.1. Context of adversarial attacks in MedIA

A recent market report has predicted that through 2022, 30% of all cyberattacks against systems powered by artificial intelligence (AI) will leverage training-data poisoning, AI model theft, or adversarial examples (Cearley et al., 2019). This results especially alarming for the healthcare industry, considering that it is predicted to suffer two to three times more cyberattacks than the average amount for other industries (Cisco and Cybersecurity Ventures, 2019, 2019). Limited resources and fragmented governance on cybersecurity (Martin et al., 2017; Ghafur et al., 2019), and larger consequences at both financial (IBM, 2020) and human levels (Martin et al., 2017) make healthcare particularly vulnerable to cyberattacks.

Adversarial attacks may therefore pose a large threat in the medical domain (Finlayson et al., 2019; 2018). This is due to two main factors: financial interests and technical sources of vulnerability.

Firstly, some parties involved in healthcare systems have a financial interest in manipulating patient diagnosis and prognosis. Healthcare fraud has been shown to be committed by large companies as well as individuals (Rudman et al., 2009; Kalb, 1999). When expressed as a proportion of the global healthcare expenditure estimated by the World Health Organisation in 2013 ($7.35 trillion or € 5.65 trillion), the global average healthcare fraud and error loss equates to 6.19% ($455 billion or € 350 billion) (Gee and Button, 2015). In the future, adversarial attacks could be used as a tool to manipulate MedIA systems supporting insurance, clinical, or drug/device approval decisions. Adversarial attacks can boost existing fraudulent behavior in fee-for-service healthcare systems, such as the one in the United States, where healthcare providers and insurance companies manipulate diagnostic codes of patients to affect reimbursement decisions. Fraudulent behavior involving adversarial attacks could potentially be more difficult to detect compared to manipulating diagnostic codes directly. Adversarial attacks can also be used to bias patient diagnosis towards false referrals or unnecessary prescriptions of medication or treatment. Similarly, companies could bias trial outcomes and gain the favor of regulatory bodies, such as the United States Food and Drug Administration, by showing the desired effect of a drug/device to be approved. It is important to emphasize that these attacks would be facilitated because the attacker would be already inside the healthcare infrastructure. These situations can result in deteriorated quality of healthcare, financial loss, decreased trust in MedIA systems and hence impediments to their integration into clinical practice.

The second factor that may facilitate adversarial attacks in the medical domain concerns technical sources of vulnerability. These include domain-specific characteristics of medical images, such as highly-standardized image acquisition protocols, and the security of technological infrastructure into which MedIA systems will be embedded (Ma et al., 2021; Finlayson et al., 2019). In this case, the attacks would be performed most commonly from outside the healthcare infrastructure, by means of a breach. In a recent investigation, more than 45 million medical images and their patient metadata were found to be exposed and freely accessible, without hacking tools required, on over 2,000 unprotected medical servers across 67 countries, including the United States, United Kingdom, France, and Germany (CybelAngel, 2020). A survey from 2017 revealed that healthcare data breaches have affected one in four consumers in the United States (Accenture, 2017). The security risks of such breaches include blackmail and ransomware (Forbes, 2021), as well as malicious data manipulation. Among deployed connected medical devices, imaging systems (including systems for image acquisition, viewers, workstations, and servers) have been found to have the most security issues, mainly derived from user practice and outdated infrastructure (Healthcare Innovation, 2018). This last aspect is strongly related to widely used software and protocols, such as DICOM, which were developed before cybersecurity was a concern and leave serious security gaps (Eichelberg et al., 2020; Stites and Pianykh, 2016).

1.2. Adversarial attacks and defenses

Multiple methods to generate adversarial attacks have been proposed in the literature and can be categorized following different taxonomies (Yuan et al., 2019; Akhtar and Mian, 2018; Biggio and Roli, 2018). As an example, some methods perform one-shot attacks (Szegedy et al., 2013; Goodfellow et al., 2014), whereas other methods optimize the attack in an iterative way (Madry et al., 2017; Kurakin et al., 2016; Papernot et al., 2016b; Moosavi-Dezfooli et al., 2016; Carlini and Wagner, 2017b; Su et al., 2019; Moosavi-Dezfooli et al., 2017). Similarly, there are methods that generate a specific perturbation for each input (Szegedy et al., 2013; Goodfellow et al., 2014; Madry et al., 2017; Kurakin et al., 2016; Papernot et al., 2016b; Moosavi-Dezfooli et al., 2016; Carlini and Wagner, 2017b; Su et al., 2019) and methods that generate universal perturbations that can be applied to any image (Moosavi-Dezfooli et al., 2017; Brown et al., 2017).

Furthermore, adversarial attack methods can be applied in scenarios with different degrees of knowledge of the target system: from having full knowledge (*white-box attacks*) (Goodfellow et al.,

[3] https://github.com/Gerda92/adversarial_transfer_factors,https://doi.org/10.5281/zenodo.4792375

2014) to being agnostic to the (hyper)parameters of the target model (*black-box attacks*) (Papernot et al., 2017). The latter usually use another model, commonly referred to as *surrogate model*, to craft adversarial examples that are then transferred to the target model. The effectiveness of a black-box attack is determined by its *transferability* between the surrogate model and the target model (Papernot et al., 2017).

Several studies have investigated the impact of adversarial attacks on MedIA systems specifically. This has been studied for classification and segmentation problems in different imaging modalities, including color fundus imaging (Finlayson et al., 2018; Ma et al., 2021; Ozbulak et al., 2019), chest X-ray (Finlayson et al., 2018; Taghanaki et al., 2018; Ma et al., 2021), dermoscopy (Finlayson et al., 2018; Ma et al., 2021; Paschali et al., 2018; Ozbulak et al., 2019), and brain MRI (Paschali et al., 2018). In these studies, adversarial attacks were proven effective in both white- and black-box settings.

Correspondingly, numerous defense methods (Yuan et al., 2019; Papernot et al., 2017; Biggio and Roli, 2018) have been proposed to protect DL systems from adversarial attacks by training networks so as to robustify them against adversarial attacks (Goodfellow et al., 2014; Papernot et al., 2016c) or by detecting adversarial examples or neutralizing adversarial noise (Lu et al., 2017; Song et al., 2017). Defense methods have also been considered to protect MedIA systems from adversarial attacks (Ma et al., 2021). Nevertheless, almost all proposed countermeasures have been shown to be only effective against some attacks (Yuan et al., 2019), hardly work against infinitesimal perturbations (Papernot et al., 2017), or can easily be made ineffective if the attacker is aware of them (Uesato et al., 2018; Athalye et al., 2018; Carlini and Wagner, 2017a).

1.3. Adversarial vulnerability of MedIA systems

A better understanding of factors affecting the vulnerability of MedIA systems is therefore crucial to inform and improve the evaluation of their robustness against adversarial attacks, as well as the design of new MedIA systems. There are several factors related to the design of the target model, such as network architecture (Szegedy et al., 2013; Su et al., 2018), and the attack scenario, such as disparity in the development data, i.e., difference in the data used for training and validation, between the target and the attacker (Szegedy et al., 2013), that affect the transferability of adversarial attacks and thus the vulnerability of the systems. Although factors such as network architecture disparity (i.e. having different network architectures) (Paschali et al., 2018; Taghanaki et al., 2018) are sometimes considered when evaluating vulnerability of MedIA systems against adversarial attacks, the impact of other crucial aspects of real-world MedIA scenarios has not been explored yet.

In this paper, we focus on two unexplored factors that can potentially influence adversarial attack transferability in MedIA systems: ImageNet pre-training and development data disparity. The key contributions of our paper are:

- We study the effect of ImageNet pre-training on adversarial attack transferability. Since systems pre-trained on natural images have shown to achieve improved performance in shorter training times in several medical applications (Gulshan et al., 2016; Wang et al., 2017), pre-training on ImageNet has become a common design choice for development of MedIA systems (Litjens et al., 2017). Pre-trained models may be more similar to each other compared to randomly initialized models due to retaining information learned from ImageNet. However, to the best of our knowledge, no studies (of MedIA or any other DL systems) have compared transferability of adversarial attacks between pre-trained models to that between randomly initialized models.
- We study the effect of disparity in the data used for development of the target and surrogate models. With increasing availability of high-quality, large public datasets, it becomes more likely that MedIA systems will, at least partly, rely on these easily accessible data in order to fulfill the requirement of large datasets for DL development. Simultaneously, MedIA systems in the deployment stage might also make use of larger amounts of private data (Abràmoff et al., 2016; González-Gonzalo et al., 2020; Murphy et al., 2020). Comparing adversarial transferability in scenarios of development data parity and disparity may provide further insight on how vulnerable MedIA systems are. Additionally, we study adversarial robustness of ImageNet pre-trained and randomly initialized networks trained using smaller development sets under an attack scenario of data disparity, simulating target models developed with small, private datasets.
- We investigate these factors in three popular medical applications: detection of referable diabetic retinopathy in color fundus images, classification of pathologies in chest X-Ray, and breast cancer metastasis detection in histological lymph node sections.

We used the following methodology to study the effect of ImageNet pre-training and development data disparity on adversarial transferability. We implemented different adversarial attack methods and applied them to different state-of-the-art network architectures, which allows us to additionally evaluate the effect of network architecture disparity: i.e., the effect of target and surrogate models having a different architecture as compared to them having the same architecture. We perform our experiments in varying black-box settings, which we consider to be the most realistic attack scenario for MedIA systems. In contrast to previous studies, we analyze and adjust the perturbation degree used in our experiments so as to ensure optimal transferability at minimal visual perceptibility of the adversarial attacks, considering human input is often required in MedIA settings. We thoroughly examine the implications of our results on the design of MedIA systems, as well as provide recommendations for evaluating their robustness against adversarial attacks.

2. Related work

Black-box attacks can have varying degrees of interaction with the target model: from having no interaction at all to unlimited querying of the model and using its predictions in crafting adversarial perturbations (for example, one-pixel attacks by Su et al. (2019), or oracle attacks such as the one proposed by Papernot et al. (2016a)). The non-query-based type of black-box attacks is the focus of this work and is, perhaps, the most commonly studied (Akhtar and Mian, 2018; Yuan et al., 2019), including in the MedIA field (Finlayson et al., 2018; Paschali et al., 2018; Taghanaki et al., 2018).

Black-box attacks that do not allow querying the target model typically rely on the transferability of adversarial perturbations from a surrogate model to the target model. Adversarial examples have been shown to be transferable between highly distinct models (Szegedy et al., 2013; Liu et al., 2016; Moosavi-Dezfooli et al., 2017). The transferability of adversarial examples between different models can be explained by the similarity of their decision boundaries (Tramèr et al., 2017b) and depends on how similar their design and training are (Uesato et al., 2018). Perhaps, the most well-studied factor affecting adversarial transferability is disparity in model architecture (Su et al., 2018). Relatively few studies have investigated the influence of other kinds of target-surrogate differences on the success of adversarial attacks: most studies trained their target and surrogate models on exactly the same subset of the same dataset, and use the same pre-processing, data augmen-

tation, weight initialization, training loss function, and other training parameters.

In this study, we focus on the effects of two previously unexplored factors in MedIA settings on the transferability of black-box attacks: pre-training on ImageNet and disparity in the development data between target and surrogate models. We also study the interaction of both factors with network architecture disparity. Below we provide an overview of the literature related to these factors:

Pre-training on ImageNet. In the MedIA field, DL methods commonly use pre-training on natural images to improve performance (Litjens et al., 2017). Pre-trained networks have also often been used in studies on adversarial robustness (Finlayson et al., 2018; Paschali et al., 2018; Ma et al., 2021). However, all studies either considered target and surrogate models that were both pre-trained or both randomly initialized. To our knowledge, no studies have compared adversarial attack transferability between DL networks pre-trained on ImageNet (or any other dataset used for performance boosting) to that between randomly initialized networks. We hypothesize that the transferability of adversarial examples between pre-trained target and surrogate models may be larger than that between randomly initialized models, since pre-training might increase the similarity between models due to retaining information learned from ImageNet.

Although the effect of ImageNet pre-training was not studied in the black-box attack scenario, its effect on white-box adversarial robustness was studied by Hendrycks et al. (2019) for networks that were adversarially fine-tuned: i.e., trained using adversarial training (Madry et al., 2017) on the target data after pre-training. In their study, regular ImageNet pre-training had no positive effect on the white-box robustness of networks adversarially fine-tuned on CIFAR. However, adversarial ImageNet pre-training increased the robustness substantially. The effect on robustness of adversarial pre-training for networks that are fine-tuned normally (not adversarially) was not reported by Hendrycks et al. (2019) or others.

Disparity in development data. Szegedy et al. (2013) reported that adversarial examples crafted using a surrogate model trained on a different (similarly sized) data subset as the target model are substantially less transferable than those crafted using the same training data for the target and surrogate model. However, they only demonstrated this for simple fully-connected models trained on MNIST. No further studies have focused on the effect of training data disparity, including in the MedIA field: all studies of black-box attacks on MedIA DL assumed perfect data parity (Finlayson et al., 2018; Paschali et al., 2018; Taghanaki et al., 2018). This factor is particularly important to study in the context of MedIA systems, where some systems are trained on easily accessible public data, whereas others rely on private data. In the case of using only public data for development, we can assume that surrogate models can be trained with the same dataset as the target (data parity), while in the case of using private data, this is not possible (data disparity). We believe it is important to consider these different scenarios and study the influence of data (dis)parity on transferability of adversarial examples in MedIA systems.

Disparity in model architecture. Su et al. (2018) studied the adversarial robustness of 18 well-known image classification models trained on ImageNet. Their findings suggest that adversarial examples crafted from one model can only be transferred within the same family (e.g. VGGs or Densenets). They also found that deeper models within the same family are slightly more robust than shallower models, but differences in model architecture were found to affect transferability more than differences in model size. There have been no similarly comprehensive studies on architecture disparity or adversarial example transferability between different architectures for MedIA systems. However, some studies reported attack performance under both architecture parity and disparity (Paschali et al., 2018) or under disparity only (Taghanaki et al., 2018). Szegedy et al. (2013) found that having architecture disparity in addition to development data disparity further reduced the transferability of attacks. In this study, we investigate the interaction of architecture (dis)parity with weight initialization (pre-training on ImageNet or random initialization) and development data (dis)parity.

3. Methods

3.1. Threat model

The security of any system is measured in relation to the capabilities and goals of its potential adversaries. The limits to the attackers capabilities, including their knowledge, and their goals are captured by the concept of a *threat model*. In the context of evaluating adversarial robustness of machine learning systems, explicitly specifying the considered threat model helps to clearly delineate the scope of attacks against which robustness is studied and thus allows for falsifiable claims (Carlini et al., 2019). The threat model considered in this study is the following:

Goal. We assume the attackers goal is to cause general misclassification, which is usually called an *untargeted* adversarial attack. In an untargeted adversarial attack the goal is to modify the input in a way that it will be classified as any class but the ground-truth class, whereas in a targeted adversarial attack the goal is to modify the input in a way that it will be classified as a specific class.

Capability. We assume the attacker's capabilities are:

- The attacker can only manipulate the input to the target system (we assume this input is directly fed into DL networks) and only at inference time.
- The attacker is allowed to modify the input images in a way that appears very subtle or even imperceptible to the human eye.
- The attacker cannot query the target model.

Knowledge. We simulate scenarios of the attacker lacking knowledge of the following features of the target model: weight initialization (pre-trained on ImageNet or randomly initialized), data used for development, and network architecture. The weights of the target model cannot be accessed by the attacker in all attack scenarios we consider.

3.2. Adversarial attacks

In this study, we used two adversarial attack methods that were most commonly and effectively used in the literature: fast gradient sign method (FGSM) (Goodfellow et al., 2014) and projected gradient descent (PGD) (Madry et al., 2017).

Fast gradient sign method. FGSM is a one-shot attack method in which the adversarial perturbation is computed as the sign of the gradient of the loss with respect to the input image. The sign of the gradient in every pixel determines whether ϵ, the parameter regulating the maximum amount of perturbation, is added or subtracted from every pixel in the target image x to create an adversarial example:

$$x_{adv} = x + \epsilon \cdot sign\big(\nabla_x \mathcal{L}(f(x;\theta), y)\big), \tag{1}$$

where $\mathcal{L}$ represents the loss, f the selected network architecture, θ the corresponding parameters, and y the image label.

Projected gradient descent. PGD is an iterative version of FGSM, in which several steps for computing the perturbation and adding it to the input are performed:

$$x_{adv}^{(i+1)} = clip_x^{\epsilon}\Big\{x^{(i)} + \alpha \cdot sign\big(\nabla_x \mathcal{L}(f(x^{(i)};\theta), y)\big)\Big\}, \tag{2}$$

Table 1
Size of development and testing subsets for each dataset. In order to study the effect of development data disparity on adversarial attack transferability, development sets were divided into two equal-sized parts: *d1* and *d2; d1* was subsequently subsampled to obtain small datasets which contained 10% of the data *d1/10; d2* was subsequently subsampled to obtain a half-sized subset *d2/2*.

		Ophthalmology		Radiology		Pathology	
Development	d1	79,058 (88%)	39,030	86,524 (80%)	43,657	294,912 (90%)	147,456
	d2		39,028		42,867		147,456
	d1/10		3,906		3,989		14,892
	d2/2		19,514		23,666		73,728
Test		10,644 (12%)		25,596 (20%)		32,768 (10%)	
Total		88,702 (100%)		112,120 (100%)		327,680 (100%)	

where α controls the step size; ϵ is the parameter regulating the maximum degree of perturbation added to every pixel; $clip_x^{\epsilon}$ function clips its input so that it does not deviate from x more than ϵ as measured by ℓ_∞ norm.

In the black-box setting, $f'(\cdot, \theta')$, where f' is the surrogate network architecture and θ' are the corresponding parameters, is used to compute the attack, which is then transferred to the target model.

3.3. Network architectures, training, and data

We selected Inception-v3 (Szegedy et al., 2016) and Densenet-121 (Huang et al., 2017) as the base architectures for our experiments. Both architectures were previously applied in the selected medical applications and achieved good performance (Gulshan et al., 2016; Rajpurkar et al., 2017; Guendel et al., 2018; Veeling et al., 2018). All networks were trained until convergence on a validation set using Adam optimization with learning rate decay and binary cross-entropy loss.

For the dataset used in each application, a development and a test set were defined. The development set was used for training and validation. The independent test set was used to measure the performance of each model on clean and adversarial examples. We randomly divided all development sets, at patient-level, into two non-overlapping, equal-sized parts — *d1* and *d2* — to be able to study the influence of data parity on attack transferability. Two more sets, *d2/2* and *d1/10*, were created by randomly sampling at patient level half of *d2* and 10% of *d1*, respectively. This was done to study the influence of dataset size. The description of each dataset and dataset-specific network parameters is stated below. Table 1 provides an overview of data partitioning for each dataset.

Ophthalmology. We used the Kaggle dataset for diabetic retinopathy detection (Kaggle, 2015), which contains 88,702 color fundus images with manually-labeled diabetic retinopathy severity. In order to have more images available for development, as proposed in Finlayson et al. (Finlayson et al., 2018), we merged the original training (35,126 images) and test sets (53,576 images) and split the images randomly at patient-level subsets for development (88%) and testing (12%).

Pre-processing included extracting the field of view and rescaling to 512×512 pixels. The networks were trained to distinguish between non-referable (stages 0 to 1) and referable diabetic retinopathy (stages 2 to 4) using batch class balancing. For data augmentation, we used flipping and rotation.

Radiology. We used the ChestX-Ray14 dataset (Wang et al., 2017), consisting of 112,120 frontal-view X-rays annotated with 14 non-mutually-exclusive pathology labels. The official data split (80%-20%) was used to define our development and test sets.

Pre-processing included downsampling images to 256×256 resolution. The architectures were trained using binary cross-entropy loss to predict 14 pathology classes and one "no finding" class. For data augmentation, we used translation and horizontal flipping.

Pathology. We used the PatchCamelyon (PCam) (Veeling et al., 2018) dataset, which contains 327,680 patches extracted from histopathology whole-slide images of lymph node sections. The official data split (90%-10%) was used to define our development and test sets.

The networks were trained to distinguish between the presence or absence of metastatic tissue in the patch center. For data augmentation, we used horizontal and vertical flipping and random color augmentations.

4. Experimental setup

In all experimental setups, the performance of the target models on the test set of each dataset was measured using the area under the receiver operating characteristic curve (AUC) or mean AUC for the multi-class case.

4.1. Perturbation degree

Firstly, we analyzed the effect of perturbation degree on the adversarial attacks to ensure maximal transferability at minimal visual perceptibility in further experiments. To our knowledge, only one study has systematically analyzed the effect of perturbation degree in MedIA settings (Ma et al., 2021), although it was only done for white-box attacks. We believe perturbation degree is a parameter that should be further investigated to yield more accurate estimations of robustness against adversarial attacks. In this study, we analyzed the performance of FGSM and PGD attacks and the visual perceptibility under different degrees of perturbation, controlled by ϵ: 0.01, 0.02, 0.03, 0.04, 0.05, and 0.06. These values were applied to image intensities rescaled between -1 and 1. We assessed visual perceptibility of attacks bounded by different epsilons in two different ways. Firstly, the first authors used their own visual perception to judge how subtle adversarial perturbations appear when adversarial and original images were viewed in juxtaposition. Due to impracticality of assessing every adversarial input to our models, this was evaluated in a subset of images of each modality and for each epsilon. Secondly, we computed mean Structural Similarity Index Measure (SSIM) (Wang et al., 2004) between adversarial and original versions of all images for each modality and epsilon. SSIM is based on a hypothesized characteristic of the human visual system to be sensitive to structural information in images and was previously shown to be a robust measure of perceptual quality of images (Wang et al., 2004).

For the PGD attacks, we used step size $\alpha = 0.01$ and 20 iterations. In this experiment, all models were randomly initialized and trained on the same partition of the development set, *d1*.

To ensure that the decrease in target model performance after an adversarial attack is due to the adversarial nature of the perturbation and not solely due to added noise, we additionally computed "control" noise. While existing works chose standard noise distributions such as Gaussian for this purpose (Paschali et al., 2018), we chose to compare adversarial perturbations with their randomly spatially shuffled versions to ensure the same degree of

Table 2
Effects of perturbation degree on attack transferability. Average performance (AUC) over two model architectures is shown when using FGSM, PGD or control noise (spatially shuffled black-box adversarial perturbations) with varying perturbation degrees. The target and surrogate model were both trained with the same dataset, *d*1. The lowest AUC value (highest attack transferability) in each application is shown in bold.

Data	Noise	FGSM					PGD					
	$\epsilon =$	0.01	0.02	0.03	0.04	0.05	0.06	0.01	0.02	0.03	0.04	0.05	0.06
Ophthalmology	None	0.86											
Ophthalmology	Adversarial	0.56	0.44	0.37	**0.32**	**0.32**	0.33	0.72	0.56	0.44	0.37	0.35	0.34
Ophthalmology	Control	0.85	0.85	0.84	0.79	0.76	0.73	0.86	0.85	0.85	0.84	0.84	0.84
Radiology	None	0.75											
Radiology	Adversarial	0.61	0.55	0.52	0.51	0.51	0.52	0.65	0.57	0.52	0.49	0.47	**0.45**
Radiology	Control	0.75	0.75	0.75	0.74	0.73	0.72	0.75	0.75	0.75	0.75	0.74	0.74
Pathology	None	0.87											
Pathology	Adversarial	0.70	0.56	0.45	0.38	0.35	**0.33**	0.73	0.56	0.47	0.41	0.38	0.36
Pathology	Control	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87	0.87

perturbation in adversarial and "control" examples. Spatial shuffling was performed by randomly permuting perturbation values for all pixels.

4.2. Pre-training on ImageNet

In this set of experiments, we measured the attack effectiveness when target and surrogate were both pre-trained on ImageNet, both randomly initialized, or had different initializations (pre-trained or random). We measured this for target-surrogate pairs with the same and different architectures separately. For this purpose, we trained four versions of each architecture (two pre-trained and two randomly initialized) to cover all possible target-surrogate combinations in black-box settings, using the same partition of the development set, *d1*.

4.3. Development data disparity

This experimental setup focused on the effect of disparity in the data used for the development of the target and surrogate models, as well as its interaction with architecture disparity. For the first part of this set of experiments, we trained four randomly initialized versions of each architecture: a target model trained and validated on *d1* and three surrogate models trained and validated on *d1, d2*, and *d2/2*, respectively. For every development dataset, the same split between training and validation images was used to train every model.

For the second part, we experimented with target models trained on small datasets attacked by surrogate models trained on larger, non-overlapping datasets. Since pre-training on ImageNet is often needed to reach good performance in models trained on small datasets, we have included it in this experiment. As target models, we trained pre-trained and randomly initialized versions of each architecture on *d1/10*; as surrogate models, we trained pre-trained and randomly initialized versions of each architecture on *d2*. For every development dataset, the split between training and validation images was the same as in previous experiments.

5. Results

5.1. Perturbation degree

The results of our experiments with different attack methods (FGSM and PGD) at different perturbation degrees can be found in Table 2. The results for individual models are included in the Supplementary Material. Increasing adversarial perturbation degree decreased the target model's performance in most cases. The experiments with control noise (spatially shuffled noise) showed that in the ophthalmology and radiology datasets the decrease in the performance of the target could be partially attributed to image corruption. However, this effect was quite small, except for the FGSM attack in the ophthalmology dataset. FGSM and PGD attacks performed similarly for the radiology and pathology dataset. For the ophthalmology dataset, the FGSM attack decreased the performance of the target model more than the PGD attack. We chose to use both attacks in our subsequent experiments and report average results.

Fig. 1 shows original images and their adversarial counterparts computed using FGSM attacks at different perturbation degrees. Fig. 2 shows mean SSIM values across all images for FGSM and PGD attacks. SSIM values for individual models are included in the Supplementary Material. As can be seen, applying the same amount of perturbation to different imaging modalities has a different effect on human visual perceptibility and the measured SSIM. Adversarial perturbations were the most noticeable in the radiology images, with $\epsilon = 0.02$ yielding an already visible, albeit quite subtle perturbation. For the ophthalmology and pathology images, at the same perturbation degree, perturbations were almost imperceptible and became noticeable with higher epsilon values. Perturbations computed by FGSM had lower SSIM than those computed by PGD in all three datasets. This is an expected result, since PGD optimizes perturbations according to both their impact on model predictions and their size.

For our further experiments, we chose to report attacks using $\epsilon = 0.02$, as this was the highest perturbation degree that was still visually subtle for all applications and attack methods, and it had substantially better transferability than an epsilon of 0.01 in most of the studied applications.

5.2. Pre-training on ImageNet

Table 3 summarizes our experiments on the effect of pre-training on adversarial attack transferability and its interaction with model architecture parity. The results for individual models and different attack methods can be found in the Supplementary Material. In the ophthalmology and radiology datasets, the attack transferability between pre-trained models was substantially higher than that between randomly initialized models. In both datasets, the effect was consistent: for all eight combinations of attack method and target and surrogate pairs (including pairs having a different architecture), pre-trained targets had lower performance when attacked by pre-trained surrogates, compared to their randomly initialized counterparts. In the pathology dataset, however, the opposite effect was observed with similar consistency. It is noteworthy that the effect of pre-training on transferability seemed to correlate to the performance increase resulting from pre-training: in the ophthalmology dataset, both the performance boost obtained by using pre-training and the transferability

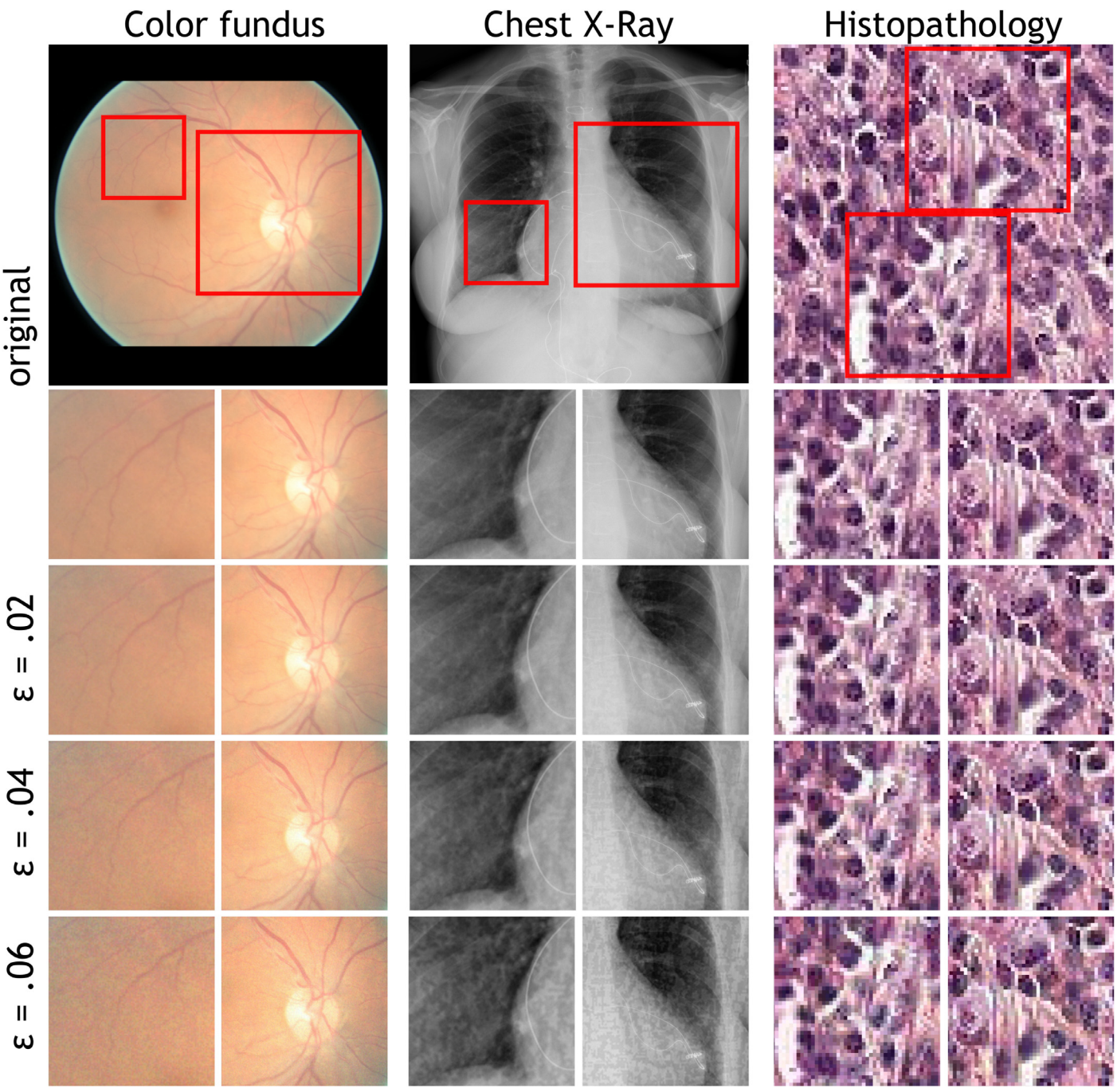

Fig. 1. Original and adversarial images created with fast gradient sign method attacks using different perturbation degrees (ϵ). The images in the top row are the original images. The red squares indicate the location of the patches that we show in the rest of the figure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of adversarial examples between pre-trained networks were high; in the radiology dataset, the performance boost was smaller and the transferability was also smaller; in the pathology dataset, pre-training yielded no benefit and the effect on transferability was reversed.

Fig. 3 includes two examples from the ophthalmology dataset that illustrate attack transferability when both target and surrogate are pre-trained on ImageNet and when both are randomly initialized.

All the aforementioned effects held similarly for the scenarios where the target and surrogate model had the same or different architecture.

5.3. Development data disparity

The effects of data disparity on adversarial attack transferability and its interaction with model architecture disparity can be seen in Table 4. The results for individual models and different attack methods are included in the Supplementary Material. For all datasets, networks were substantially less susceptible to attacks crafted using surrogates with the same architecture but trained on a different data subset (*d2* or *d2/2*). This held for both target architectures and both attack methods. Decreasing the surrogate training set size (from *d2* to *d2/2*) led to a further drop in the attack transferability for the ophthalmology and radiology datasets.

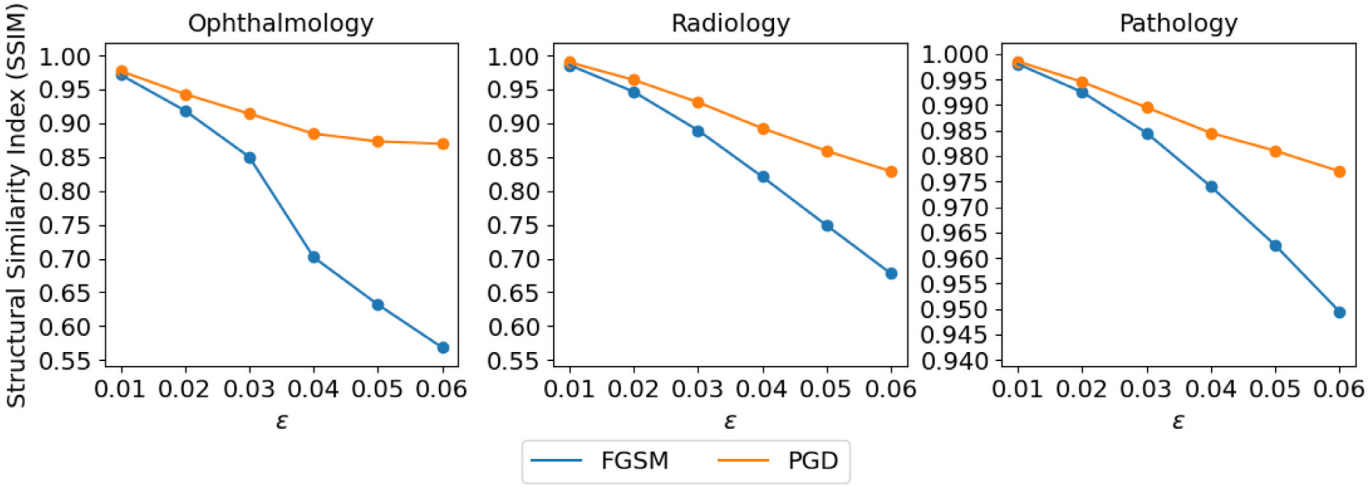

Fig. 2. Mean Structural Similarity Index Measure (SSIM) computed between original images in the test sets and adversarial examples generated using FGSM or PGD with varying perturbation degree ϵ. Points represent SSIM values averaged over two model architectures (Inception-v3 and Densenet-121)..

Table 3
Effects of pre-training and the interaction between pre-training and model architecture parity on attack transferability. Average performance (AUC) over FGSM and PGD (ϵ=0.02) and two model architectures (Inception-v3 and Densenet-121) is shown. The target and surrogate model were both trained with the same dataset, $d1$. Average relative performance with respect to the no-attack setting is shown in brackets. The lowest AUC value (highest attack transferability) in each application is shown in bold. .

Architecture	Target	Surrogate	Ophthalmology	Radiology	Pathology
No attack	Imagenet	-	0.94 (100%)	0.78 (100%)	0.87 (100%)
No attack	Random	-	0.86 (100%)	0.75 (100%)	0.87 (100%)
Same	Imagenet	Imagenet	**0.00 (0%)**	**0.31 (40%)**	0.61 (70%)
Same	Random	Random	0.44 (51%)	0.48 (64%)	**0.41 (47%)**
Same	Random	Imagenet	0.63 (74%)	0.63 (83%)	0.60 (69%)
Same	Imagenet	Random	0.80 (85%)	0.55 (71%)	0.71 (82%)
Different	Imagenet	Imagenet	0.24 (25%)	0.50 (65%)	0.75 (86%)
Different	Random	Random	0.55 (64%)	0.64 (86%)	0.71 (82%)
Different	Random	Imagenet	0.71 (83%)	0.65 (86%)	0.69 (80%)
Different	Imagenet	Random	0.86 (92%)	0.59 (76%)	0.75 (86%)

Table 4
Effects of data and model architecture parity on attack transferability. Average performance (AUC) over FGSM and PGD (ϵ=0.02) and two model architectures is shown, with surrogate models trained on different sets while the target model is trained on $d1$. Average relative performance with respect to the no-attack setting is shown in brackets. The lowest AUC value (highest attack transferability) in each application is shown in bold.

Architecture	Training set	Ophthalmology	Radiology	Pathology
No attack	-	0.86 (100%)	0.75 (100%)	0.87 (100%)
Same	d1	**0.44 (51%)**	**0.48 (64%)**	**0.41 (47%)**
Same	d2	0.56 (65%)	0.56 (75%)	0.67 (77%)
Same	d2/2	0.75 (88%)	0.59 (79%)	0.65 (75%)
Different	d1	0.55 (64%)	0.64 (86%)	0.71 (82%)
Different	d2	0.66 (77%)	0.65 (87%)	0.74 (85%)
Different	d2/2	0.80 (93%)	0.69 (91%)	0.71 (81%)

When the architecture of the surrogate was different, however, additional data disparity between the target and surrogate substantially decreased the attack performance only for the ophthalmology dataset. Disparity in the model architecture had greater effect on attack performance than disparity in data for the radiology and pathology datasets; for the ophthalmology dataset, data and model architecture disparity had similar effects.

The transferability of attacks on models trained on small datasets in a data disparity scenario is reported in Table 5. The Supplementary Material contains the results for individual models and different attack methods. For the ophthalmology and radiology datasets, the pre-trained models clearly outperformed their randomly initialized counterparts on clean images. For the pathology dataset, pre-trained models performed slightly worse than randomly initialized ones. These results are similar to ones we observed for models trained on larger sets (see Table 3). On adversarial images, pre-trained models performed worse than their randomly initialized counterparts in all three datasets, both in absolute terms and relative to their performance on clean images. These results were mostly similar to the results for networks trained on larger data (Table 3). For the ophthalmology and radiology datasets, adversarial attack transferability between pre-trained models was higher than that between randomly initialized models, and this effect was stronger in the ophthalmology dataset. There was an interesting difference, however: for the pathology dataset, pre-training increased transferability, whereas in our experiments with networks trained on larger data it was the other way around. Attacks on randomly initialized models trained on small datasets hardly have any effect (Table 5), while attacks on randomly initialized models trained on larger sets can lead to performance decreases of up to 35% (Table 4).

6. Discussion

In this study, we have demonstrated that ImageNet pre-training may substantially affect transferability of adversarial examples, even between networks with different architecture. This effect varied substantially across the applications and appeared to be re-

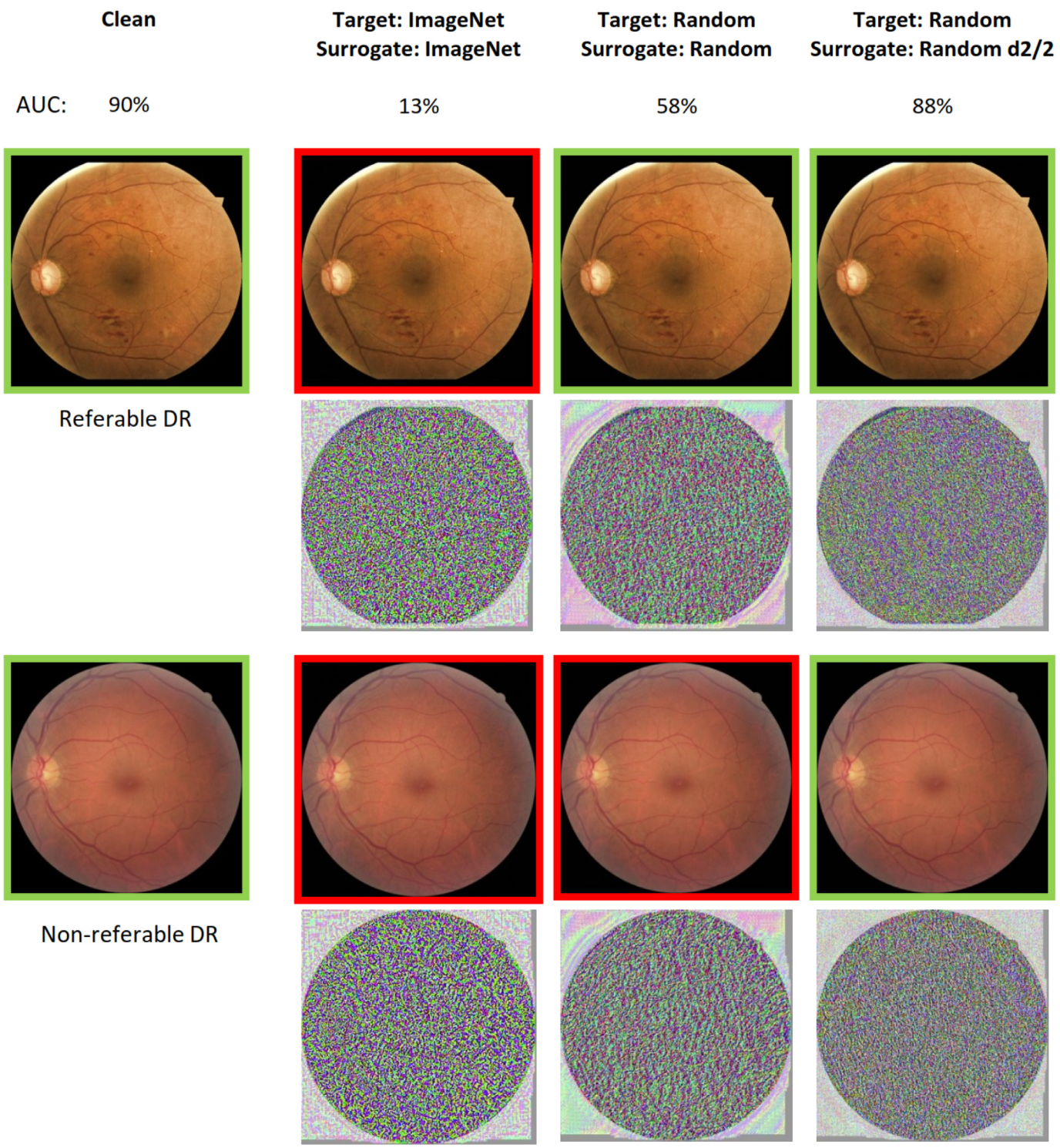

Fig. 3. Original images, adversarial images and corresponding adversarial noise created with FGSM (ϵ=0.02) in different black-box settings: target and surrogate pre-trained on ImageNet; target and surrogate randomly initialized; target and surrogate randomly initialized plus surrogate developed using a different and reduced dataset (d2/2). The average area under the receiver operating characteristic curve (AUC) is indicated above of each configuration for the clean and the black-box settings. Green frame indicates correct classification of referable or non-referable diabetic retinopathy (DR); red frame, incorrect classification. The adversarial noise shown is equivalent to the difference between the original and the adversarial image.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

lated to the gain in performance resulting from pre-training. We have also shown that differences in development data between target and surrogate models reduce transferability substantially, even when development sets are equally sized and sampled from the same distribution. This effect was in some cases comparable to that of architecture disparity. All experiments were performed using a perturbation degree tuned to be visually subtle and perform optimally in the black-box attack setting.

In this section, we discuss the importance of perturbation degree tuning and the influence of pre-training and data disparity on transferability of adversarial attacks. Based on the results of our study, we make recommendations for developers of MedIA sys-

Table 5
Transferability of attacks on models trained on small datasets in a data disparity scenario. Average performance (AUC) over FGSM and PGD (ϵ=0.02) and two model architectures is shown, with surrogate models trained on $d2$ and target models trained on $d1/10$. Average relative performance with respect to the no-attack setting is shown in brackets. The lowest AUC value (highest attack transferability) in each application is shown in bold.

Architecture	Target	Surrogate	Ophthalmology	Radiology	Pathology
No attack	Imagenet	-	0.88 (100%)	0.69 (100%)	0.79 (100%)
No attack	Random	-	0.61 (100%)	0.64 (100%)	0.81 (100%)
Same	Imagenet	Imagenet	**0.13 (15%)**	**0.58 (84%)**	**0.73 (93%)**
Same	Random	Random	0.60 (99%)	0.63 (99%)	0.78 (96%)
Different	Imagenet	Imagenet	0.45 (51%)	0.62 (90%)	0.75 (96%)
Different	Random	Random	0.60 (100%)	0.63 (99%)	0.79 (98%)

tems, as well as for future evaluation of adversarial robustness of these systems.

6.1. Perturbation degree

Our experiments confirmed that perturbation degree is an important attack parameter to take into account to obtain more accurate estimates of adversarial robustness of DL systems. Using a lower-than-optimal perturbation degree may lead to an underperforming attack and hence an overestimated robustness; using a higher-than optimal perturbation degree may make the adversarial perturbation visually perceptible. We observed differences in visual perceptibility of adversarial perturbations in different imaging modalities as estimated by both our visual perception and SSIM. This could occur because of differences in color, homogeneity, contrast, and resolution between the imaging modalities. Since these characteristics may affect visual perceptibility of adversarial attacks, it is important to optimize the perturbation degree to the image type and task, as well as for the considered attack scenario (e.g., whether adversarial examples are likely to be inspected by a human). Quantitative measures, such as SSIM, could also be used to ensure minimal visual perceptibility of adversarial perturbations. However, since there is not an accepted threshold value to determine whether a perturbation is imperceptible for SSIM or other quantitative perceptibility measures, an optimal threshold value would still need to be found and ensured to agree with human visual perception. Moreover, to our knowledge, no quantitative metric can perfectly capture human visual perceptibility (Chandler, 2013). Therefore, we think the best way to assess visual perceptibility of different degrees of adversarial perturbations would be a blinded observer study involving medical experts. Such a study is beyond the scope of this paper.

Ma et al. (2021) experimented with different perturbation degrees in the white-box attack setting and concluded that MedIA systems are “easier to attack” than systems trained on natural images, based on their observation that for MedIA systems far smaller perturbations were needed to reach near-maximal attack performance. In our study, we considered the more realistic black-box setting, in which perturbations became visually perceptible before yielding high attack performance. This suggests that, firstly, in black-box settings, MedIA systems may not be very easy to attack. Secondly, it suggests it is harder to compare the difficulty of attacking systems in different applications: for example, in applications where a given perturbation degree yields better attack effectiveness, the perturbations may also be more perceptible.

6.2. Pre-training on ImageNet

In the ophthalmology and radiology applications, we observed that transferability between pre-trained models, including the ones with different architectures, was substantially larger than that between randomly initialized models: 20–50% difference in AUC was observed. These results motivate caution in generalizing performance of black-box adversarial attacks from pre-trained networks to randomly initialized ones and vice versa. For example, an attack that was only shown effective on pre-trained targets and surrogates may be substantially less effective when applied to randomly initialized targets and surrogates in the same application or to networks in applications that do not benefit from pre-training.

We believe increased transferability between pre-trained models may be explained by increased closeness of their decision boundaries. Tramèr et al. (2017b) showed empirically that decision boundaries of DL models are on average closer to each other than to data points, which implies that adversarial perturbations causing data points to cross one model’s decision boundary would likely cause them to cross another model’s decision boundary as well. There are several possible mechanisms through which pre-training may increase closeness of decision boundaries of models. Firstly, pre-trained networks with the same architecture start with the same weight initialization (whereas randomly initialized networks in our experiments started with different initializations), which may increase the similarity of the features they learn. The fact that pre-training speeds up convergence may amplify this. Secondly, pre-trained networks may be more similar because they retain some features from ImageNet pre-training. As, in our experiments, pre-training also increased transferability between models with different architectures, same weight initialization is likely not the only cause of increased similarity between pre-trained networks. The correlation between the strength of the performance boost from pre-training and the increase in transferability also supports the second mechanism: the higher the performance gain from pre-training, the more the network retains from its ImageNet pre-training.

Our observations put into an interesting perspective the ones made in the study by Hendrycks et al. (2019) the only study on the effects of pre-training on adversarial robustness we are aware of. Hendrycks et al. (2019) found that adversarial pre-training on ImageNet can increase adversarial robustness for networks adversarially fine-tuned on the target data in the white-box attack setting. We found that regular ImageNet pre-training can decrease adversarial robustness in the black-box setting. Whether adversarial pre-training could instead improve robustness in the black-box setting remains an open question. On the one hand, adversarial training is substantially less successful in preventing attacks in the black-box than in the white-box setting (Tramèr et al., 2017a) and these results could be expected to extend to adversarial pre-training. Furthermore, if any kind of pre-training increases vulnerability to black-box attacks by similarly pre-trained networks, for example, by increasing similarity between the decision boundaries of the target and the surrogate, adversarial pre-training could be less beneficial or even detrimental to black-box robustness. On the other hand, even if adversarial pre-training facilitated transferability to some degree, it could still be overall beneficial due to the fact that the network would be trained to be adversarially robust

on a larger and more variable set of images. Future research could focus on answering these questions.

6.3. Development data disparity

Data parity is assumed, to our knowledge, by all studies on black-box adversarial attack robustness of MedIA systems (Finlayson et al., 2018; Taghanaki et al., 2018; Paschali et al., 2018). Our results, however, indicate that black-box attacks may be less effective when using a surrogate trained on a different dataset, even if it is a large dataset of the same size as the development data of the target and it is sampled from the same distribution. A 30–40% increase in AUC of the attacked models in the ophthalmology and pathology datasets was observed when the surrogate was trained on a disjoint subset. This data disparity effect may be as strong or even stronger than the effect of architecture disparity, as observed in the ophthalmology dataset.

The data disparity effect is even stronger when the target model is trained on a small dataset, in which case attacks are generally quite ineffective, especially when performed against randomly initialized target models. However, pre-trained models trained on small datasets can still be vulnerable to adversarial attacks. This vulnerability increases for applications where ImageNet pre-training provides a significant boost in clean performance, similar to what was observed in the experiments with models trained on larger datasets.

These results suggest that MedIA systems that use private development data are less susceptible to adversarial attacks than systems that use public development data (assuming attacks performed by an external party who cannot access the private data and assuming other properties of the systems are equal). Simulating data disparity between the target and surrogate model yields a more realistic estimate of adversarial robustness for such systems. Studies on adversarial robustness would therefore benefit from including different attack scenarios assuming data parity and disparity, including differences in development data sizes of target and surrogate networks, in their evaluation.

6.4. Adversarial robustness: Inception-v3 vs Densenet-121

Considering the results included in the Supplementary Material for each implemented model architecture, Inception-v3 and Densenet-121, we observed that, in the ophthalmology application, target models based on Inception-v3 tended to be more vulnerable when attacked by surrogate models with the same architecture, whereas target models based on Densenet-121 were slightly more vulnerable when attacked by surrogate models based on Inception-v3 (compared to Inception-v3 attacked by Densenet-121 models). In the radiology application, target models based on Inception-v3 were observed to be on average more vulnerable than those based on Densenet-121, although no substantial differences were observed for ImageNet pre-trained versions of the models. In the pathology application, target models based on Densenet-121 were found to be slightly more vulnerable in most scenarios. Furthermore, when there was development data disparity between target and surrogate models, only small differences in robustness between architectures were observed in all applications.

Su et al. (2018) studied transferability of adversarial attacks between popular architectures trained on ImageNet. Densenet-121 was observed to be more robust, often substantially, to FGSM and PGD attacks by Inception-v3 than the other way around. Simultaneously, there was almost perfect transfer between different variants of Densenet: Densenet-121, Densenet-161, and Densenet-169 (although transferability between the same version of architectures were not reported). Our results showed different trends when comparing Densenet-121 and Inception-v3 in different applications, for different weight initializations (ImageNet pre-training or random initialization), and for different target-surrogate development data configurations. For example, we observed perfect or high transferability between Densenet-121 models only for ophthalmology and radiology applications and only for the ImageNet-pretrained versions. It is thus difficult to conclude whether either of these architectures is innately more robust to black-box attacks than the other.

6.5. Recommendations for developers of MedIA systems

We recommend developers of all MedIA systems to be deployed in clinical practice to consider the environment their system will be used in and assess whether the following holds:

1. Users of these systems may have a motivation (financial or otherwise) to manipulate their output.
2. Users may have the capacity to manipulate their input without being detected.

For MedIA systems satisfying these criteria, especially those systems that significantly affect clinical or financial decision-making, we recommend taking proactive measures to mitigate the risk of successful adversarial attacks.

Many different methods have been proposed to defend DL systems from adversarial attacks (Yuan et al., 2019; Akhtar and Mian, 2018; Biggio and Roli, 2018). Although all defense methods proposed to date are only partially effective (Yuan et al., 2019), applying the most successful methods is likely to increase the difficulty of manipulating DL systems. We thus recommend developers of security-critical MedIA systems to consider employing some of these strategies. In addition to strategies purposefully designed to defend against adversarial attacks, quantifying uncertainty and using techniques for interpreting predictions may aid in detecting adversarial attacks (Li and Gal, 2017; Tao et al., 2018). It was shown that adversarial perturbations can increase the model's uncertainty (Li and Gal, 2017) and cause discrepancies in interpretations of the model's predictions (Tao et al., 2018). However, detection of adversarial examples based on uncertainty and interpretability also provides only partial protection against adversarial attacks (Carlini, 2019; Smith and Gal, 2018), and can be easily circumvented when taken into account in the attack method (Zhang et al., 2020).

Given that adversarial defense methods are not fully reliable, and given that increased transferability between similar models was observed in this and other studies (for example, Su et al. (2018)), we also recommend taking measures to increase the difficulty of training a surrogate model similar to the target. As one such measure, we recommend restricting the amount of information on the design of the system available to the public. This includes information on the methodology components of the system, such as network architecture and weight initialization. We also recommend avoiding disclosing extensive details on the systems data: for example, names and identifying details of used public data, detailed information on distribution of subjects, scanning modalities, and protocols. However, we do not recommend keeping secret the methods, procedures, and description of datasets used to evaluate the system, since this would make it harder to ensure the system is safe and has the desired performance level.

To further increase the difficulty of emulating the target model for an attacker, we recommend considering re-designing MedIA systems to reduce the use of standard components, such as popular network architectures, and components that facilitate transferability, such as pre-training, as well as reducing the reliance of these systems on publicly available development data. For example, standard architectures could be replaced by customized architectures and pre-training may be substituted by random initializa-

tion. However, we recommend this strategy only as a complement to more explicit defense strategies and only if it does not lead to a significant decrease in the system performance or substantially slow down its development.

We acknowledge that our recommendation to avoid using standard components, such as pre-training on ImageNet and publicly available development datasets might hamper performance. However, for MedIA systems planned to be deployed in clinical practice, robustness needs to be considered in addition to performance. The trade-off between performance and robustness has already been discussed by others (Zhang et al., 2019; Tsipras et al., 2018; Paschali et al., 2018). The decision on how much performance to sacrifice for robustness will differ per case depending on the likelihood of adversarial attacks against the given system and their potential consequences.

We believe a combination of several defense strategies would provide the most comprehensive security. Thus, we recommend combining multiple methods for detecting, neutralizing, or robustifying against adversarial perturbation, with measures that increase the difficulty of modeling the target system for a potential attacker.

We would like to emphasize that all recommendations above only apply to systems that are planned to be deployed in practice. However, we believe they are also relevant to researchers developing models at earlier stages, or performing research not specifically focused on adversarial attacks. We consider it important that MedIA researchers are aware of the effect that commonly used design components, such as pre-training on ImageNet or public datasets, have on attack transferability and the existing trade-off between performance and robustness of DL systems. This way, researchers will acknowledge the role of adversarial vulnerabilities in model development, with the capability of shifting what is currently standard in MedIA towards components that acknowledge these vulnerabilities as well.

6.6. Recommendations for evaluating adversarial robustness of MedIA systems

Carlini et al. (2019) presented a detailed discussion on best evaluation practices to conduct reproducible, falsifiable studies on adversarial robustness of DL systems. They place an emphasis on estimating the upper bound of adversarial robustness: that is, adversarial robustness measured against attacks of the maximally knowledgeable and capable attacker. Below is a condensed list of their general recommendations:

- State a precise threat model that the target system is supposed to be robust under.
- Perform adaptive attacks to estimate the upper bound of robustness: test attacks that have full access to the defense mechanisms the target system might use and adapt attacks to the target system so as to maximize their effectiveness. This includes carefully investigating the attack parameters to ensure optimal attack performance.
- Perform various sanity checks on the success rates of the attacks to ensure they are correctly implemented and their methodology is valid (for example, white-box iterative attacks should perform better than one-step attacks; attacks adapted to the studied system should perform at least as good as any other).
- Test diverse attacks (e.g. one-shot attacks and iterative attacks).
- Describe the attacks studied fully, including parameters.
- Compare against prior work and explain important differences.

Current studies on adversarial attacks on MedIA systems do not follow all of these practices. To the best of our knowledge, no published studies investigating robustness of MedIA systems formulate an explicit threat model, and thus do not clearly define the considered attack scenarios; many do not tune the parameters of their attacks, including perturbation degree (Paschali et al., 2018; Taghanaki et al., 2018; Finlayson et al., 2018); and some do not report all attack parameters (Paschali et al., 2018; Taghanaki et al., 2018).

Inspired by the results in our study, we have developed several additional recommendations and suggestions for evaluating adversarial robustness of DL systems. Note that while recommendations of Carlini et al. (2019) (particularly the recommendation on performing adaptive attacks) have as their aim estimating the upper bound of adversarial robustness, our recommendations have a different scope. We aim at investigating factors that may affect adversarial vulnerability of real-world DL systems, which are unlikely to be completely known by the attacker, as well as at obtaining realistic estimates of robustness of such systems.

- For image analysis (including MedIA) applications, we recommend tuning the perturbation degree (ϵ or another parameter controlling it) to the target image type or modality, so that the attack yields maximal performance while the perturbations still satisfy a chosen criterion for bounding perturbation degree, such as visual perceptibility. Such criterion should be explicitly defined and measured. For example, if the criterion is visual perceptibility, we suggest the studies to describe how perceptibility was judged and to provide fully-sized or zoomed-in versions of images that the reader can also examine.
- We encourage researchers to consider design components shared by both target and surrogate that may increase the similarity between them and study the effect of changing such settings on attack transferability. For example, pre-training on ImageNet, development data parity, and architecture parity could be considered as similarity-promoting components as in our study. Other similarity-promoting settings could focus on regularization techniques, which encourage networks to have specific properties (e.g. weight decay, deformation consistency regularization), loss function, pre-processing, data augmentation protocol, or popular network architectures other than the ones we used and their properties (such as ResNets and skip connections, found to increase adversarial vulnerability (Wu et al., 2020)).

The recommendations developed by Carlini et al. (2019) and by us are aimed at public scientific studies on adversarial robustness. However, we can envision a different setting for evaluating robustness of DL systems where most of this advice may also be useful: a private evaluation setting in which the robustness of a closed-source DL system is evaluated by the developing company or a different organization. In this setting, it may be of interest to estimate robustness under the most likely attack scenarios, which may exclude scenarios where the attacker has complete or very comprehensive knowledge of the target system. Therefore, recommendations aimed at obtaining realistic robustness estimates, as opposed to the upper bound estimates, may be the most applicable. Recommendations we would not advise to apply in this setting are those concerning public disclosure of robustness evaluation procedure, including tested attacks and their parameters.

7. Conclusion

In this paper, we studied the influence of two previously unexplored factors on the transferability of black-box adversarial attacks in three different MedIA applications. We observed that pre-training on ImageNet may dramatically increase the transferability of adversarial examples in MedIA systems; the larger the performance gain achieved by pre-training, the larger the transfer and thus the more vulnerable the pre-trained system is to attacks by

pre-trained surrogate models. We also showed that disparity in development data and model architecture between target and surrogate models can substantially decrease the success of attacks. We believe these factors should be considered in the design of security-critical MedIA systems, especially those planned to be deployed in clinical practice. In order to reduce the transferability of potential attacks, in addition to using techniques developed for defending DL models against adversarial attacks, we recommend restricting the disclosure of information on design specifications, as well as considering reducing the use of standard components (such as pre-training on ImageNet and popular network architectures) and publicly available datasets for development. Finally, we believe future studies on adversarial robustness of MedIA systems may benefit from simulating various attack scenarios and target-surrogate disparities. This may facilitate a better understanding of attack transferability and the factors that determine it, as well as more realistic robustness estimates for MedIA systems.

Declaration of Competing Interest

None.

Acknowledgments

This work was supported by the Deep Learning for Medical Image Analysis (DLMedIA) research program by The Dutch Research Council (project number P15-26), Intel Corporation (GB, IK, LH), and Philips Research (SCW, JPWP, MV).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2021.102141

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.

Abràmoff, M.D., Lou, Y., Erginay, A., Clarida, W., Amelon, R., Folk, J.C., Niemeijer, M., 2016. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Investigative ophthalmology & visual science 57 (13), 5200–5206.

Accenture 2017. One in four US consumers have had their healthcare data breached, Accenture survey reveals. https://newsroom.accenture.com/subjects/technology/one-in-four-us-consumers-have-had-their-healthcare-data-breached-accenture-survey-reveals.htm#:~:text=ORLANDO2C20Fla.3B20Feb.,today20at20HIMSS201720in20Orlando, Accessed: 2021, February 2.

Akhtar, N., Mian, A., 2018. Threat of adversarial attacks on deep learning in computer vision: a survey. IEEE Access 6, 14410–14430.

Athalye, A., Carlini, N., Wagner, D., 2018. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. arXiv preprint arXiv:1802.00420.

Baumgartner, C.F., Koch, L.M., Tezcan, K.C., Ang, J.X., Konukoglu, E., 2018. Visual feature attribution using wasserstein gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8309–8319.

Becker, A.S., Jendele, L., Skopek, O., Berger, N., Ghafoor, S., Marcon, M., Konukoglu, E., 2019. Injecting and removing suspicious features in breast imaging with cyclegan: a pilot study of automated adversarial attacks using neural networks on small images. Eur J Radiol 120, 108649.

Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318 (22), 2199–2210.

Biggio, B., Roli, F., 2018. Wild patterns: ten years after the rise of adversarial machine learning. Pattern Recognit 84, 317–331.

Brown, T.B., Man, D., Roy, A., Abadi, M., Gilmer, J., 2017. Adversarial patch. https://arxiv.org/abs/1712.09665.

Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., Litjens, G., 2020. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. The Lancet Oncology.

Carlini, N., 2019. Is ami (attacks meet interpretability) robust to adversarial examples? arXiv preprint arXiv:1902.02322.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., Kurakin, A., 2019. On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705.

Carlini, N., Wagner, D., 2017. Adversarial examples are not easily detected: Bypassing ten detection methods. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 3–14.

Carlini, N., Wagner, D., 2017. Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). IEEE, pp. 39–57.

Cearley 2019. Top 10 strategic technology trends for 2020: A gartner special report. https://www.gartner.com/en/doc/432920-top-10-strategic-technology-trends-for-2020, Accessed: 2021, February 15.

Chandler, D.M., 2013. Seven challenges in image quality assessment: past, present, and future research. Int Sch Res Notices 2013.

Cisco and Cybersecurity Ventures 2019. Press release: 2019/2020 cybersecurity almanac: 100 facts, figures, predictions and statistics. https://cybersecurityventures.com/cybersecurity-almanac-2019/,Cisco and Cybersecurity Ventures. Accessed: 2021, February 16.

CybelAngel 2020. 45M medical images accessible online. https://cybelangel.com/blog/medical-data-leaks/Accessed: 2021, February 2.

Eichelberg, M., Kleber, K., Kämmerer, M., 2020. Cybersecurity in PACS and medical imaging: an overview. J Digit Imaging 1–16.

Finlayson, S.G., Bowers, J.D., Ito, J., Zittrain, J.L., Beam, A.L., Kohane, I.S., 2019. Adversarial attacks on medical machine learning. Science 363 (6433), 1287–1289.

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.

Forbes 2021. Ransomware attacks on the healthcare sector are skyrocketing. https://www.forbes.com/sites/leemathews/2021/01/08/ransomware-attacks-on-the-healthcare-sector-are-skyrocketing/?sh=4e1cba62d250,Accessed: 2021, February 2.

GE Reports 2019. GE healthcare receives FDA clearance of first artificial intelligence algorithms embedded on-device to prioritize critical chest X-ray review. https://www.genewsroom.com/press-releases/ge-healthcare-receives-fda-clearance-first-artificial-intelligence-algorithmsGE Reports. Accessed: 2020, June 10.

Gee, J., Button, M., 2015. The financial cost of healthcare fraud 2015: What data from around the world shows.

Ghafur, S., Grass, E., Jennings, N.R., Darzi, A., 2019. The challenges of cybersecurity in health care: the uk national health service as a case study. The Lancet Digital Health 1 (1), e10–e12.

González-Gonzalo, C., Sánchez-Gutiérrez, V., Hernández-Martínez, P., Contreras, I., Lechanteur, Y.T., Domanian, A., van Ginneken, B., Sánchez, C.I., 2020. Evaluation of a deep learning system for the joint automated detection of diabetic retinopathy and age-related macular degeneration. Acta Ophthalmol (Copenh) 98 (4), 368–377.

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

Guendel, S., Grbic, S., Georgescu, B., Liu, S., Maier, A., Comaniciu, D., 2018. Learning to recognize abnormalities in chest x-rays with location-aware dense networks. In: Iberoamerican Congress on Pattern Recognition. Springer, pp. 757–765.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 316 (22), 2402–2410.

Healthcare Innovation 2018. IoT report: Imaging systems present biggest security risk in healthcare. https://www.hcinnovationgroup.com/cybersecurity/news/13029895/iot-report-imaging-systems-present-biggest-security-risk-in-healthcare,Accessed: 2021, February 2.

Hendrycks, D., Lee, K., Mazeika, M., 2019. Using pre-training can improve model robustness and uncertainty. arXiv preprint arXiv:1901.09960.

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

IBM 2020. Cost of a data breach report 2020. https://www.ibm.com/security/data-breach, IBM. Accessed: 2021, February 16.

Kaggle 2015. Diabetic retinopathy detection competition. Onlinehttps://www.kaggle.com/c/diabetic-retinopathy-detection/.

Kalb, P.E., 1999. Health care fraud and abuse. JAMA 282 (12), 1163–1168.

Kurakin, A., Goodfellow, I., Bengio, S., 2016. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533.

Li, Y., Gal, Y., 2017. Dropout inference in bayesian neural networks with alpha-divergences. In: International conference on machine learning. PMLR, pp. 2052–2061.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med Image Anal 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.

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., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., Lu, F., 2021. Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognit 110, 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.

Martin, G., Martin, P., Hankin, C., Darzi, A., Kinross, J., 2017. Cybersecurity and healthcare: how safe are we? BMJ 358.

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.
Murphy, K., Habib, S.S., Zaidi, S.M.A., Khowaja, S., Khan, A., Melendez, J., Scholten, E.T., Amad, F., Schalekamp, S., Verhagen, M., et al., 2020. Computer aided detection of tuberculosis on chest radiographs: an evaluation of the cad4tb v6 system. Sci Rep 10 (1), 1–11.
Ozbulak, U., Van Messem, A., De Neve, W., 2019. Impact of adversarial examples on deep learning models for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 300–308.
Papernot, N., McDaniel, P., Goodfellow, I., 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277.
Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A., 2017. Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia conference on computer and communications security, pp. 506–519.
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 (EuroS&P). IEEE, pp. 372–387.
Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A., 2016. Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE Symposium on Security and Privacy (SP). IEEE, pp. 582–597.
Paschali, M., Conjeti, S., Navarro, F., Navab, N., 2018. Generalizability vs. robustness: adversarial examples for medical imaging. arXiv preprint arXiv:1804.00504.
Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al., 2017. Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225.
Rudman, W.J., Eberhardt, J.S., Pierce, W., Hart-Hester, S., 2009. Healthcare fraud and abuse. Perspectives in Health Information Management/AHIMA, American Health Information Management Association 6 (Fall).
Smith, L., Gal, Y., 2018. Understanding measures of uncertainty for adversarial example detection. arXiv preprint arXiv:1803.08533.
Song, Y., Kim, T., Nowozin, S., Ermon, S., Kushman, N., 2017. Pixeldefend: leveraging generative models to understand and defend against adversarial examples. arXiv preprint arXiv:1710.10766.
Stites, M., Pianykh, O.S., 2016. How secure is your radiology department? mapping digital radiology adoption and security worldwide. American Journal of Roentgenology 206 (4), 797–804.
Su, D., Zhang, H., Chen, H., Yi, J., Chen, P.-Y., Gao, Y., 2018. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 631–648.
Su, J., Vargas, D.V., Sakurai, K., 2019. One pixel attack for fooling deep neural networks. IEEE Trans. Evol. Comput. 23 (5), 828–841.
Sun, L., Wang, J., Huang, Y., Ding, X., Greenspan, H., Paisley, J., 2020. An adversarial learning approach to medical image synthesis for lesion detection. IEEE J Biomed Health Inform 24 (8), 2303–2314.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. 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., Fergus, R., 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. Springer, pp. 87–94.
Tao, G., Ma, S., Liu, Y., Zhang, X., 2018. Attacks meet interpretability: attribute-steered detection of adversarial samples. arXiv preprint arXiv:1810.11580.
Ting, D.S.W., Cheung, C.Y.-L., Lim, G., Tan, G.S.W., Quang, N.D., Gan, A., Hamzah, H., Garcia-Franco, R., San Yeo, I.Y., Lee, S.Y., et al., 2017. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA 318 (22), 2211–2223.
Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P., 2017. Ensemble adversarial training: attacks and defenses. arXiv preprint arXiv:1705.07204.
Tramèr, F., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P., 2017. The space of transferable adversarial examples. arXiv preprint arXiv:1704.03453.
Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., Madry, A., 2018. Robustness may be at odds with accuracy. arXiv preprint arXiv:1805.12152.
Uesato, J., O'Donoghue, B., Oord, A.v.d., Kohli, P., 2018. Adversarial risk and the dangers of evaluating against weak attacks. arXiv preprint arXiv:1802.05666.
Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M., 2018. Rotation equivariant CNNs for digital pathology. In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 210–218.
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.
Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.
Wetstein, S.C., Onken, A.M., Luffman, C., Baker, G.M., Pyle, M.E., Kensler, K.H., Liu, Y., Bakker, B., Vlutters, R., van Leeuwen, M.B., et al., 2020. Deep learning assessment of breast terminal duct lobular unit involution: towards automated prediction of breast cancer risk. PLoS ONE 15 (4), e0231653.
Wu, D., Wang, Y., Xia, S.-T., Bailey, J., Ma, X., 2020. Skip connections matter: on the transferability of adversarial examples generated with resnets. arXiv preprint arXiv:2002.05990.
Xia, T., Chartsias, A., Tsaftaris, S.A., 2020. Pseudo-healthy synthesis with pathology disentanglement and adversarial learning. Med Image Anal 64, 101719.
Yuan, X., He, P., Zhu, Q., Li, X., 2019. Adversarial examples: attacks and defenses for deep learning. IEEE Trans Neural Netw Learn Syst 30 (9), 2805–2824.
Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M., 2019. Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning. PMLR, pp. 7472–7482.
Zhang, X., Wang, N., Shen, H., Ji, S., Luo, X., Wang, T., 2020. Interpretable deep learning under fire. 29th {USENIX} Security Symposium ({USENIX} Security 20).