Universal adversarial attacks on deep neural networks for medical image classification

Hokuto Hirano, Akinori Minagi and Kazuhiro Takemoto\*[](http://orcid.org/0000-0002-6355-1366)

**Abstract**

**Background:** Deep neural networks (DNNs) are widely investigated in medical image classification to achieve automated support for clinical diagnosis. It is necessary to evaluate the robustness of medical DNN tasks against adversarial attacks, as high-stake decision-making will be made based on the diagnosis. Several previous studies have considered simple adversarial attacks. However, the vulnerability of DNNs to more realistic and higher risk attacks, such as universal adversarial perturbation (UAP), which is a single perturbation that can induce DNN failure in most classification tasks has not been evaluated yet.

**Methods:** We focus on three representative DNN-based medical image classification tasks (i.e., skin cancer, referable diabetic retinopathy, and pneumonia classifications) and investigate their vulnerability to the seven model architec- tures of UAPs.

**Results:** We demonstrate that DNNs are vulnerable to both nontargeted UAPs, which cause a task failure resulting in an input being assigned an incorrect class, and to targeted UAPs, which cause the DNN to classify an input into a specific class. The almost imperceptible UAPs achieved > 80% success rates for nontargeted and targeted attacks. The

vulnerability to UAPs depended very little on the model architecture. Moreover, we discovered that adversarial retrain- ing, which is known to be an effective method for adversarial defenses, increased DNNs’ robustness against UAPs in only very few cases.

**Conclusion:** Unlike previous assumptions, the results indicate that DNN-based clinical diagnosis is easier to deceive because of adversarial attacks. Adversaries can cause failed diagnoses at lower costs (e.g., without consideration of data distribution); moreover, they can affect the diagnosis. The effects of adversarial defenses may not be limited. Our findings emphasize that more careful consideration is required in developing DNNs for medical imaging and their practical applications.

**Keywords:** Deep neural networks, Medical imaging, Adversarial attacks, Security and privacy

# Background

Deep neural networks (DNNs) are effective for image classification and are beginning to be applied to medi- cal image diagnosis to empower physicians and accel- erate decision making in clinical environments [[1](#_bookmark9)]. For example, DNNs have been used to classify skin cancer based on photographic images [[2](#_bookmark10)], referable diabetic

\*Correspondence: [takemoto@bio.kyutech.ac.jp](mailto:takemoto@bio.kyutech.ac.jp)

Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502, Japan

retinopathy based on optical coherence tomography (OCT) images of the retina [[3](#_bookmark11)], and pneumonia based on chest X-ray images [[3](#_bookmark11)]. They have demonstrated high diagnostic performances. A meta-analysis [[4](#_bookmark12)] has indi- cated that the diagnostic performance of DNNs is equiv- alent to that of healthcare professionals.

Despite DNNs’ high performance, their practical appli- cation in disease diagnosis is still debatable. High-stake decision making will be based on disease diagnosis. Complex classifiers, including DNNs, can potentially cause catastrophic harm to society because they are often

difficult to interpret [[5](#_bookmark13)]. More importantly, DNNs present a number of security concerns [[6](#_bookmark14)]. Specifically, DNNs are known to be vulnerable to adversarial examples [[7](#_bookmark15), [8](#_bookmark16)], which are input images that cause misclassifications by DNNs and are typically generated by adding specific, imperceptible perturbations to original input images that have been correctly classified using DNNs. The existence of adversarial examples raises questions about DNNs’ generalization ability, reduces model interpretability, and limits deep learning applications in safety- and security- critical environments [[9](#_bookmark17)]. In particular, adversarial exam- ples cause not only misdiagnosis but also various social disturbances [[10](#_bookmark18)]. The vulnerability of DNNs to adversar- ial attacks has been claimed in skin cancer [[10](#_bookmark18)] and pneu- monia classifications based on chest X-ray images [[11](#_bookmark19)].

Nevertheless, more focused investigations are needed on DNNs’ vulnerability to adversarial attacks. Previous studies have only considered input-dependent adversarial attacks (i.e., an individual adversarial perturbation is used such that each input image is misclassified). Such adver- sarial attacks are difficult because they require high com- putational costs. More realistic adversarial attacks must be further considered. Notably, a single small, image agnostic perturbation, called *universal adversarial per- turbation (UAP)*, that can induce DNN failure in most image classification tasks, has been reported [[12](#_bookmark20)]. A pre- vious study [[12](#_bookmark20)] considered only UAPs for nontargeted attacks, which cause misclassification (i.e., a task failure resulting in an input image being assigned an incorrect class). However, we previously extended the UAPs gen- erating algorithm to enable targeted attacks [[13](#_bookmark21)], which caused the DNN to classify an input image into a specific class. UAPs are difficult to detect because such perturba- tions are extremely small and, hence, do not significantly affect data distributions. UAP-based adversarial attacks can be more straightforward to implement by adversar- ies in real-world environments. UAPs are vulnerable to security threats in medical image diagnosis; however, UAP vulnerability is still poorly evaluated in DNN- based medical image diagnosis to date. Specifically, many researchers and engineers have simply developed DNNs using transfer learning (by fine-tuning pretrained DNN models with medical images), inspired by famous studies on medical image classification based on DNNs [[2](#_bookmark10), [3](#_bookmark11)] and have applied DNNs to medical image classification with- out consideration for their vulnerability to UAPs. Addi- tionally, defense strategies against UAPs in DNN-based medical image classification are still poorly investigated, although the vulnerability of DNNs to adversarial attacks indicates the need for strategies to address security con- cerns (i.e., adversarial defense [[8](#_bookmark16)]). Specifically, adversar- ial retraining is one of the few approaches that could not be defeated thus far [[14](#_bookmark22)].

This study aims to evaluate the vulnerability of DNNs to UAPs for medical image classification and to warn against facile applications of DNNs for medical image classification. We focused on representative medical image classifications: skin cancer classification based on photographic images [[2](#_bookmark10)], referable diabetic retinopathy classification based on OCT images [[3](#_bookmark11)], and pneumo- nia classification based on chest X-ray images [[3](#_bookmark11)]. We obtained DNN models with various architectures for medical image diagnosis inspired by previous studies [[2](#_bookmark10), [3](#_bookmark11)] and investigated their vulnerability to nontargeted and targeted attacks based on UAPs. Moreover, adversarial defense was considered; in particular, we evaluated the increased robustness of DNNs to nontargeted and tar- geted UAPs using adversarial retraining [[12](#_bookmark20), [14](#_bookmark22)–[16](#_bookmark24)], a representative method for adversarial defenses.

# Methods

**Medical image datasets**

We used three types of medical images: skin lesion images for skin cancer classification, OCT images for ref- erable diabetic retinopathy classification, and chest X-ray images for pneumonia classification.

In a previous study [[2](#_bookmark10)], skin lesion images (red– green–blue color) were obtained from the International Skin Imaging Collaboration (ISIC) 2018 dataset (chal- lenge2018.isic-archive.com/task3/training/), in which the images were classified into seven classes: melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis/Bowens disease (intraepithe- lial carcinoma; AKIEC), benign keratosis (solar lentigo/ seborrheic keratosis/lichen planus-like keratosis; BKL), dermatofibroma (DF), and vascular lesions (VASC). The dataset comprised 10,015 images. We randomly divided these images into 7,000 training images (778 MEL, 4,689

NV, 370 BCC, 229 AKIEC, 764 BKL, 76 DF, and 94 VASC

images, respectively) and 3,015 test images (335 MEL, 2016 NV, 144 BCC, 98 AKIEC, 335 BKL, 39 DF, and 48

VASC images, respectively).

The OCT and chest X-ray images (grayscale) were obtained from a previous study [[3](#_bookmark11)] (data.mendeley. com/datasets/rscbjbr9sj/3). The OCT images were clas- sified into four classes: choroidal neovascularization with neovascular membrane and associated subreti- nal fluid (CNV), diabetic macular edema with retinal- thickening-associated intraretinal fluid (DME), multiple drusen present in early age-related macular degeneration (DRUSEN), and normal retina with preserved foveal con- tour and absence of any retinal fluid/edema (NM). The original dataset comprised 37,455 CNV, 11,598 DME, 8866 DRUSEN, and 51,390 NM images, respectively. We constructed a class-balanced training image set and test image set by randomly selecting 1960 and 840 images

per class, without duplicates, respectively. We finally obtained 7840 training and 3360 test images.

The chest X-ray images were classified into binary classes: no pneumonia (NORMAL) or viral or bacterial pneumonia (PNEUMONIA). The original dataset com- prised 1583 NORMAL and 4273 PNEUMONIA images. We constructed a class-balanced training image set and test image set by randomly selecting 900 and 270 images per class, without duplicates, respectively. We finally obtained 1800 training and 540 test images.

**Transfer learning methods**

Following previous studies [[2](#_bookmark10), [3](#_bookmark11)], we obtained the DNN models using transfer learning; in particular, we fine- tuned DNN models pretrained using the ImageNet dataset [[17](#_bookmark25)] with a medical image dataset. We mainly considered the Inception V3 architecture [[18](#_bookmark26)], following previous studies. To investigate the effect of model archi- tecture on adversarial robustness, we considered different model architectures: VGG16 [[19](#_bookmark27)], VGG19 [[19](#_bookmark27)], ResNet50 [[20](#_bookmark28)], Inception ResNet V2 [[21](#_bookmark29)], DenseNet 121 [[22](#_bookmark30)], and DenseNet 169 [[22](#_bookmark30)]. For each model architecture, we replaced the original last fully connected (FC) layer with a new FC layer in which the output size is the number of classes. The images were resized to 299 *×* 299 pixels. All layer parameters were fine-tuned using the training images in a medical image dataset. We used the stochas- tic gradient descent optimizer with a momentum of 0.9. The batch size and number of epochs were set to 32 and 50, respectively. The learning rates were scheduled based on the number of epochs: 0.001 for *≤* 40 epochs, 1e*−*4 for 41–45 epochs, and 1e*−*5 for > 45 epochs. To avoid overfitting, data augmentation was considered: random image rotations with angles ranging between *−* 5° and 5° and random 5% height and width image shifts. For the skin cancer classification, we adopted oversampling to account for imbalances in the dataset. The transfer learn- ing procedures were performed using Keras (version 2.2.4; Keras.io).

**Universal adversarial perturbations**

Simple iterative algorithms [[12](#_bookmark20), [13](#_bookmark21)] were used to gen- erate UAPs for nontargeted and targeted attacks. The algorithms’ details are described in [[12](#_bookmark20), [13](#_bookmark21)]. We used the nontargeted UAP algorithm available in the Adver- sarial Robustness 360 Toolbox (ART) [[23](#_bookmark31)] (version 1.0;

github.com/Trusted-AI/adversarial-robustness-tool-

under the constraint that the *Lp* norm of the perturba- tion is equal to or less than a small ξ value (i.e., ¨ρ*p*¨ *≤* ξ). The algorithms start with ρ = **0** (no perturbation) and iteratively update ρ by additively obtaining an adversar-

ial perturbation for an input image , which is randomly selected from **X** without replacement. These iterative updates continue until the number of iterations reaches a maximum imax.

The fast gradient sign method (FGSM) [[7](#_bookmark15)] is used to

obtain an adversarial perturbation for the input image. Meanwhile, the original UAP algorithm [[12](#_bookmark20)] uses the DeepFool method [[24](#_bookmark32)]. This is because the FGSM is used for both nontargeted and targeted attacks, and DeepFool requires a higher computational cost than the FGSM and only generates a nontargeted adversarial example for the input image. The FGSM generates the adversarial pertur- bation for based on the loss gradient [[7](#_bookmark15)] with the attack strength parameter .

Nontargeted and targeted UAPs were generated using the training images in the dataset. The parameter was set to 0.0024; cases where *p* = 2 and were consid-

ered. The parameter ξ was determined based on the ratio

ζ of the *Lp* norm of the UAP to the average *Lp* norm of

an image in the dataset. For the ISIC 2018 (skin lesion image) dataset, the average L∞ and L2 norms were 237 and 85,662, respectively. For the OCT image dataset, the average L∞ and L2 norms were 253 and 15,077, respec- tively. For the chest X-ray image dataset, the average L∞ and L2 norms were 253 and 40,738, respectively. The parameter imax was set to 15.

To compare the performances of the generated UAPs

with those of the random controls, we generated ran- dom vectors (random UAPs) sampled uniformly from the sphere of a specified radius [[12](#_bookmark20)].

**Vulnerability evaluation**

The fooling rate Rf and targeted attack success rate Rs were computed to evaluate the vulnerability of the DNN models to a nontargeted UAP (ρnt) and targeted UAP (ρt), respectively. Further, Rf for an image set **X** is defined as adversarial images for which predicted labels are inconsistent with the labels predicted from their associ- ated clean images to all images in the set (i.e., the proba- bility that the labels predicted from clean images are

inconsistent with the labels predicted from their adver- sarial images). Let C(**x**) be an output (class or label) of a

box). The targeted UAP algorithm was implemented

classifier−1 Σ(DN N) for an input image ,

by modifying the nontargeted UAP algorithm from our

Rf = |**X** |

**x**∈**X**

I C(**x**) /= C

**x** + ρnt

, where the func-

previous study in ART [[13](#_bookmark21)] (github.com/hkthirano/ targeted\_UAP\_CIFAR10).

The algorithms consider a classifier and generate non-

tion I(A) returns 1 if the condition A is true and 0 other- wise. Rs for an image set is the proportion of adversarial

images classified inΣto th e target c lass *y* to all images in

targeted (targeted) UPAs ρ from an input image set **X** ,

the set *Rs* = |**X** |−1

I *C*

**x**∈**X**

**x** + ρt

= *y* . It is noteworthy

that Rs has a baseline, defined as Rs, observed without UAPs. Class (label) composition of image data and pre- diction performance of DNNs both affect the baseline. In this study, for the OCT and chest X-ray image datasets, the Rs baselines of UAPs targeted to a specified class were ~ 25% and ~ 50%, respectively. For the skin lesion dataset, the Rs baselines of UAPs targeted to MEL and NV were ~ 10% and ~ 65%, respectively.

Additionally, we obtained the confusion matrices, to evaluate the change in prediction owing to the UAPs for each class. The rows and columns in the matrices rep- resent true and predicted classes, respectively. The con- fusion matrices were row-normalized to account for an imbalanced dataset (ISIC 2018 dataset); in particular, each cell value was normalized by the number of obser- vations with the same true class (label).

**Adversarial retraining**

Adversarial retraining was performed to increase the robustness of the DNN models to UAPs [[12](#_bookmark20), [15](#_bookmark23)]; in par- ticular, the models were fine-tuned with adversarial images. The procedure was described in a previous study [[12](#_bookmark20)]. A schematic diagram of the adversarial retraining procedure is shown in Additional file [1](#_bookmark8): Fig. S1. A brief description is provided here: (1) 10 UAPs against a DNN model were generated with the (clean) training image set;

(2) a modified training image set was obtained by ran- domly selecting half of the training images and combin- ing them with the remaining images in which each image was perturbed by a UAP randomly selected from the 10 UAPs; (3) the model was fine-tuned by performing five additional epochs of training on the modified training image set; (4) a new UAP was generated against the fine- tuned model using the algorithm with the training image

set; (5) the UAP Rf and Rs values for the test images were computed; and steps (1)–(5) were repeated five times.

# Results

**Medical images classification performance**

We evaluated the prediction performance of seven DNN models for three medical image datasets. The test and training accuracies of the models for the datasets are summarized in Additional file [1](#_bookmark8): Table S1. The DNN models achieved good accuracy. For the skin lesion, OCT, and chest X-ray image datasets, the average test accuracies across the seven models were 87.3%, 95.8%, and 98.4%, respectively. Specifically, the test accuracies of Inception V3 models, which were frequently used in previous studies on medical image diagnosis (e.g., [[2](#_bookmark10), [3](#_bookmark11)]), were 87.7%, 95.5%, and 97.6%, respectively. The normal- ized confusion matrices for the Inception V3 models on the test images are shown in Additional file [1](#_bookmark8): Fig. S2.

**Nontargeted universal adversarial attacks**

We evaluated the vulnerability of the DNN models to nontargeted UAPs. We first considered Inception V3 models because well-known previous studies on DNN- based medical image classification used the Inception V3 architecture [[2](#_bookmark10), [3](#_bookmark11)]. Figure [1](#_bookmark0) shows the case of non-

targeted UAPs *p* = 2 against the Inception V3 models.

The fooling rates Rf for both the training and test images increased rapidly with the perturbation magnitude ζ and reached a high Rf , despite a low ζ. The UAPs with ζ = 4% achieved Rf > 80% for the skin lesion (Fig. [1](#_bookmark0)a) and chest X-ray image datasets (Fig. [1](#_bookmark0)c), whereas slightly larger UAPs (with ζ = 6%) were required to achieve Rf ~ 70% for the OCT image dataset (Fig. [1](#_bookmark0)b). Further, Rf of the nontargeted UAPs was significantly higher than that of

random UAPs. The confusion matrices on test images show that the models classified most images into several specific classes (i.e., dominant classes) due to the UAPs for the skin lesion and OCT image datasets. Specifically, most skin lesion images tended to be classified as AKIEC or DF (Fig. [1](#_bookmark0)d); moreover, most OCT images were classi- fied as CNV (Fig. [1](#_bookmark0)e). For the chest X-ray image dataset, the model incorrectly predicted the true labels (Fig. [1](#_bookmark0)f ).

A high Rf at low ζ and dominant labels was observed in the case of UAP with *p* = ∞ against the Inception V3

models for all medical image datasets (Additional file [1](#_bookmark8): Fig. S3). However, the skin lesion images tended to be classified into broader classes: BCC, AKIEC, BKL, or DF (Additional file [1](#_bookmark8): Fig. S3D).

We also considered other models to evaluate whether the vulnerability to nontargeted UAPs depends on model architectures. Table [1](#_bookmark1) shows Rf of the UAPs against the

DNN models for the test images in the medical image

datasets. Overall, a vulnerability to nontargeted UAPs was observed independent of model architectures; in particular, the small UAPs (ζ = 4% for the skin lesions

and chest X-ray image datasets, and ζ = 6% for the OCT

image dataset) achieved a high Rf (70–90%). The UAPs’ Rf were significantly higher than those of the random UAPs. However, Rf partially depends on model archi- tectures; specifically, Rf of the UAPs against the VGG16 and VGG19 models were ~ 50% for the chest X-ray

image dataset, whereas those of the UAPs against the other models were between 70 and 80%. This indicates that the models classified images into either NORMAL or PNEUMONIA. In the case of UAPs with *p* = 2, the VGG16 and VGG19 models classified most test images

into PNEUMONIA and NORMAL, respectively (Addi- tional file [1](#_bookmark8): Fig. S4). In the case of UAPs with *p* = ∞, both the VGG16 and VGG19 models predicted most of

the test images as NORMAL. This indicates that the con- fusion matrix patterns (dominant classes) might change according to the model architecture and *p*. Additionally,

**T****able 1 Fooling rates *Rf* (%) of nontargeted UAPs against various DNN models for test images of skin lesions, OCT, and chest X-ray image datasets**



0.00 0.00 0.00 0.30 0.04 0.67 0.00

0.00

0.96 0.00

0.03 0.00 0.01 0.39 0.16 0.41 0.00

0.02 0.00 0.02 0.44 0.15 0.37 0.00

0.90 0.10 0.00 0.00

0.02 0.00 0.01 0.27 0.11 0.59 0.00

0.80

0.20

0.01 0.04 0.01 0.29 0.06 0.59 0.00

1.00 0.00 0.00 0.00

0.14 0.00 0.03 0.41 0.11 0.31 0.00

**f**

0.84

0.16

0.94 0.00 0.00 0.06

0.00 0.02 0.02 0.44 0.00 0.50 0.02

MEL NV BCC AKIEC BKL DF

Predicted class

VASC

CNV

DME DRUSEN NM

NORMAL

Predicted class

PNEUMONIA

Predicted class

**Fig. 1** Vulnerability to nontargeted UAPs with *p* = 2. Line plots of the fooling rate *Rf* against Inception V3 model versus perturbation magnitude

ζ for skin lesions (**a**), OCT (**b**), and chest X-ray (**c**) image datasets. Legend label indicates image set used for computing *Rf* . Additional argument “(random)” indicates that random UAPs were used instead of UAPs. Normalized confusion matrices for Inception V3 models attacked using UAPs on

test images of skin lesions (**d**), OCT (**e**), and chest X-ray (**f**) image datasets are also shown. ζ = 4% in **d** and **f**. ζ = 6% in **e**

0.6

**a**

Skin lesion

**b**

OCT

**c**

Chest X-ray

1.0 1.0 1.0

0.8

0.8

0.8

0.6

0.6

**e**

0.4

0.4

0.4

0.2

0.2

0.2

0.0

0.0

0.0

0.00 0.02 0.04 0.06 0.08 0.00 0.02 0.04 0.06 0.08 0.00 0.02 0.04 0.06 0.08

**d**

0.04

training test

training (random) test (random)

VASC DF

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

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

a change in confusion matrix patterns (on test images) was observed in the skin lesions and OCT image data- sets. For example, the VGG16 model classified most skin lesion images into BKL owing to the UAP with ζ = 4% and *p* = 2 (Additional file [1](#_bookmark8): Figure S5A), whereas the Inception V3 models classified them into AKIEC or DF

True class

BKL AKIEC BCC

NV MEL

True class

NM

DRUSEN

DME

CNV

True class

PNEUMONIA

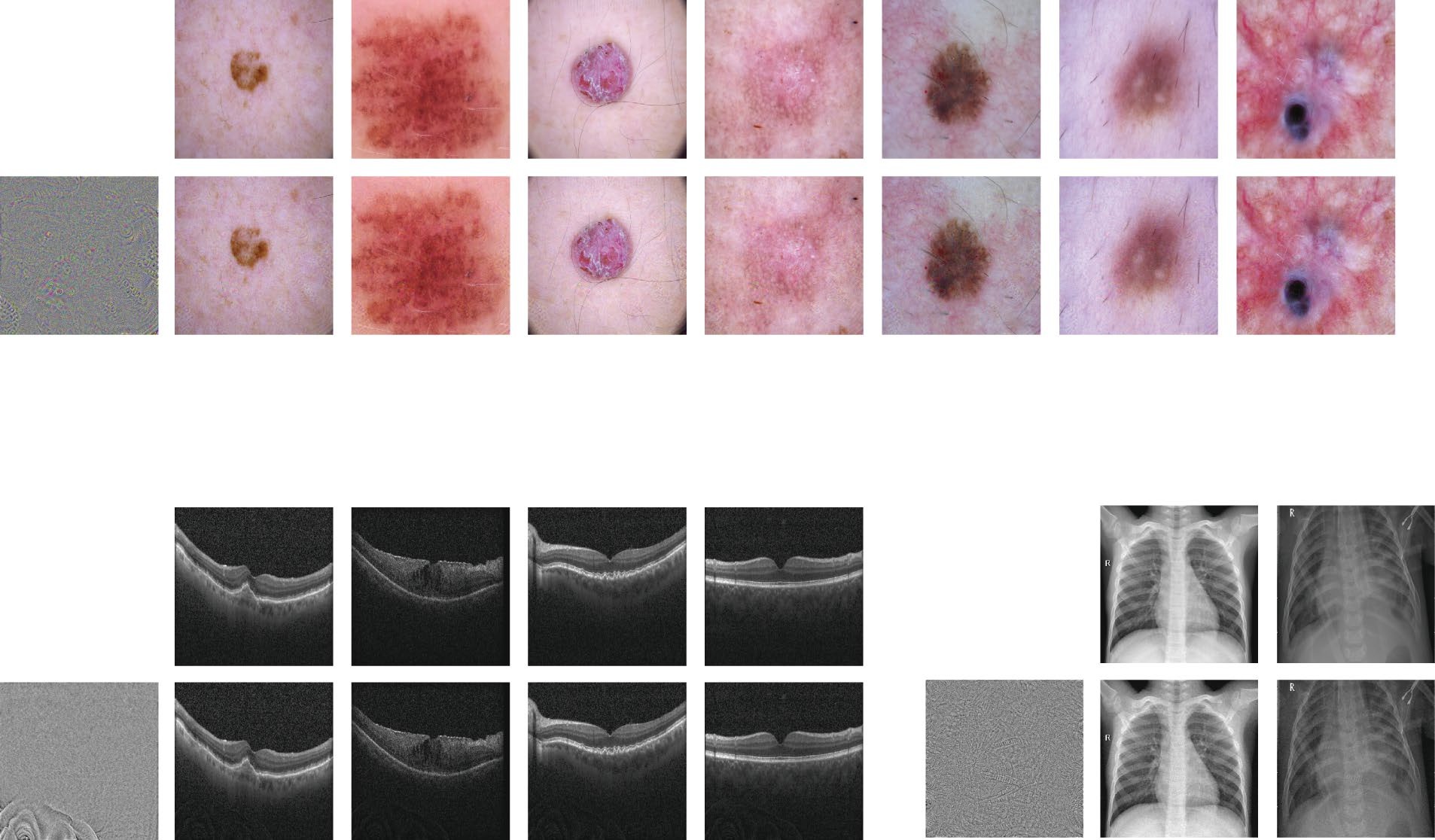
NORMAL

(Fig. [1](#_bookmark0)d). The ResNet 50 model classified most OCT images into DME owing to the UAP with ζ = 6% and *p* = 2 (Additional file [1](#_bookmark8): Fig. S5B), whereas Inception V3 models classified them into CNV (Fig. [1](#_bookmark0)e).

We investigated whether the nontargeted UAPs were perceptible. As a representative example, the nontargeted

UAPs with *p* = 2 against the Inception V3 models and examples of adversarial images for the medical image datasets are shown in Fig. [2](#_bookmark2). The UAPs with ζ = 4% for the skin lesions and chest X-ray image datasets and with ζ = 6% for the OCT image dataset were almost imper- ceptible. The models predicted the original images as their actual classes; however, they classified the adversar- ial images into incorrect classes owing to the nontargeted UAPs. The UAPs with *p* = ∞ and those against the other DNN models were also almost imperceptible for the skin lesion (Additional file [1](#_bookmark8): Fig. S6), OCT (Additional file [1](#_bookmark8): Fig. S7), and chest X-ray image datasets (Additional file [1](#_bookmark8): Fig. S8).

Moreover, we found that different UAP patterns were observed in the different model architectures for each medical image dataset (Additional file [1](#_bookmark8): Figs. S6–S8). We hypothesized that the UAPs have no transferability, which indicates that UAPs generated based on DNNs with one model architecture can be used to deceive DNNs with another model architecture and to evaluate the trans- ferability of UAPs. As expected, transferability was not



**Fig. 2** Nontargeted UAPs with *p* = 2 against Inception V3 models and their adversarial images for skin lesions (**a**), OCT (**b**), and chest X-ray image datasets (**c**). Further, ζ = 4% in **a** and **c**. ζ = 6% in **b**. Labels in brackets beside the images are the predicted classes. The original (clean) images are correctly classified into their actual labels. UAPs are emphatically displayed for clarity; in particular, each UAP is scaled by a maximum of 1 and minimum of 0

(PNEUMONIA) (NORMAL)

(CNV)

(CNV)

(CNV)

(CNV)

UAP

UAP

**c** Chest X-ray

(NORMAL) (PNEUMONIA)

(NM)

**b** OCT

(DME) (DRUSEN)

(CNV)

(AKIEC)

(AKIEC)

(AKIEC)

(AKIEC)

(AKIEC)

(AKIEC)

(AKIEC)

UAP

(VASC)

(DF)

(BKL)

**a** Skin lesion

(BCC) (AKIEC)

(NV)

(MEL)

Original

confirmed for the OCT (Additional file [1](#_bookmark8): Table S3) and chest X-ray image datasets (Additional file [1](#_bookmark8): Table S4); however, a weak transferability of UAPs was observed in the skin lesions image dataset (Additional file [1](#_bookmark8): Table S5). Specifically, the nontargeted UAPs with *p* = 2

generated based on the Inception V3 models achieved Rf of approximately 45%, ~ 2%, and ~ 10% on average against the DNNs with another model architecture for the skin lesions, OCT, and chest X-ray image datasets, respectively.

**Targeted universal adversarial attacks**

We have developed targeted UAPs in our previous study [[13](#_bookmark21)]. We evaluated whether the DNNs are vulnerable not only to nontargeted UAPs but also to targeted UAPs (i.e., whether UAPs can control DNN outputs). Table [2](#_bookmark3) shows the targeted attack success rates Rs of the UAPs

with *p* = 2 against the DNN models for the test images in

the medical image datasets. As representative examples, we considered targeted attacks to be the most significant case and the control in each medical image dataset. For

Original

Adversarial

Adversarial

Adversarial

Original

**Table 2 Targeted attack success rates** *Rs* **(%) of targeted UAPs with** *p* = 2 **against various DNN models to each target class**

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

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

skin lesion image datasets, targeted attacks on MEL and NV were considered. For the OCT image dataset, tar- geted attacks on CNV and NM were considered. For the chest X-ray image dataset, targeted attacks on PNEU- MONIA and NORMAL were considered. Overall, a high (> 85%) Rs was observed regardless of the model architec- ture, despite small UAPs (with ζ = 2% for the skin lesions and chest X-ray image datasets, and ζ = 6% for the OCT image dataset). Furthermore, the confusion matrices (Fig. [3](#_bookmark4)) indicate that the UAP-based targeted attacks were successful: most (Rs% of) test images were classi- fied into the targeted class owing to the UAPs (Table [2](#_bookmark3)).

However, a smaller Rs was partially observed according to the model architectures and datasets. For the skin lesions image dataset, Rs of the UAPs against VGG16 (~ 40%) and VGG19 (~ 65%) models were lower than those (~ 90%) of the UAPs against the other models. For the targeted attacks on NM in the OCT image dataset, Rs (30–40%) of the UAPs against the VGG and DensNet models were lower than those (~ 85%) of the UAPs against the other models. Further, Rs of random UAPs was almost equiva- lent to those of the baselines. The Rs values of the UAPs were significantly higher than those of the random UAPs. Furthermore, a high Rs for a small ζ was observed for the targeted UAPs with *p* = ∞ (Additional file [1](#_bookmark8): Table S2). However, Rs for targeted attacks on MEL was lower over- all than Rs of the UAPs with *p* = 2. For example, Rs of the UAPs with *p* = 2 and *p* = ∞ against the Inception V3 model were ~ 95% and ~ 75%, respectively.

We investigated whether the targeted UAPs were perceptible. As a representative example, the targeted UAPs with *p* = 2 against the Inception V3 models and

examples of adversarial images for the medical image

datasets are shown in Fig. [4](#_bookmark5). The targeted UAPs with ζ = 2% for the skin lesions and chest X-ray image data- sets and ζ = 6% for the OCT image dataset were also almost imperceptible. The models predicted the original images as their actual classes; however, they classified the

adversarial images into the targeted class owing to the UAPs. The UAPs with *p* = ∞ and those against the other DNN models were also almost imperceptible. For the skin lesion image dataset, Additional file [1](#_bookmark8): Figures S9 and S10 show the targeted attacks on NV and MEL, respec- tively. For the OCT image dataset, Additional file [1](#_bookmark8): Fig- ures S11 and S12 show the targeted attacks on NM and CNV, respectively. For the chest X-ray image dataset,

Additional file [1](#_bookmark8): Figures S13 and S14 show the targeted attacks on NORMAL and PNEUMONIA, respectively.

We also evaluated whether UAP patterns depend on model architectures and found that they did so for each medical image dataset (Additional file [1](#_bookmark8): Figs. S9–S14). The non-transferability of UAPs was also confirmed for the skin lesions (Additional file [1](#_bookmark8): Table S6), OCT (Addi- tional file [1](#_bookmark8): Table S7), and chest X-ray image datasets (Additional file [1](#_bookmark8): Table S8); specifically, Rs observed

when the targeted UAPs with *p* = 2 generated based

on the Inception V3 model that attacked the DNN models with another architecture were almost equiva- lent to their baselines of Rs ~ 10%, ~ 25%, and ~ 50% for the skin lesions, OCT, and chest X-ray image datasets, respectively.

**Adversarial retraining**

We analyzed the usefulness of adversarial retraining against universal adversarial attacks (both nontargeted and targeted UAPs). We considered Inception V3 mod- els because well-known previous studies on DNN-based medical image classification used the Inception V3 archi- tecture [[2](#_bookmark10), [3](#_bookmark11)].

Figure [5](#_bookmark6) shows the effect of adversarial retraining on Rf of nontargeted UAPs with *p* = 2 against Inception V3 models for the skin lesions, OCT, and chest X-ray image datasets, ζ = 4% for the skin lesions and chest X-ray

image datasets, and ζ = 6% for the OCT image dataset.

Adversarial retraining did not affect test accuracy. For the OCT image dataset, Rf decreased with the adversarial

retraining iterations; specifically, Rf decreased from 70.2 to 13.1% after five iterations (Fig. [5](#_bookmark6)b); however, ~ 40% of the NM images were still classified into an incorrect class (DME, Fig. [5](#_bookmark6)e). The adversarial retraining effect on Rf was limited for the skin lesions (Fig. [5](#_bookmark6)a) and chest X-ray

Predicted class

**Fig. 3** Normalized confusion matrices for Inception V3 models attacked with targeted UPAs with *p* = 2 on test images in skin lesions (left panels), OCT (middle panels), and chest X-ray image datasets (right panels). Further, ζ = 2% for skin lesions and chest X-ray image datasets, and ζ = 6% for OCT image dataset

PNEUMONIA

Predicted class

Targeted to NV

0.21 0.77 0.00 0.00 0.01 0.00 0.00

Targeted to NM

Targeted to NORMAL

0.69

0.00 1.00 0.00 0.00 0.00 0.00 0.00

1.00

0.00

0.01 0.86 0.09 0.01 0.04 0.00 0.00

0.80

0.03 0.67 0.00 0.20 0.09 0.01 0.00

0.00 0.85 0.00 0.00 0.14 0.01 0.00

0.86

0.00 0.81 0.00 0.00 0.00 0.19 0.00

0.92

0.08

1.00

0.00 0.82 0.00 0.00 0.00 0.00 0.18

MEL NV BCC AKIEC BKL DF VASC

Predicted class

CNV

DME DRUSEN NM

NORMAL

Predicted class

PNEUMONIA

Predicted class

0.95 0.00 0.00 0.00 0.05 0.00 0.00

Skin lesion

Targeted to MEL

1.00 0.00 0.00 0.00 0.00 0.00 0.00

OCT

Targeted to CNV

Chest X-ray

Targeted to PNEUMONIA

1.00 0.00 0.00 0.00

0.94 0.06 0.00 0.00 0.00 0.00 0.00

0.13

0.87

0.94 0.01 0.03 0.00 0.02 0.00 0.00

0.91 0.09 0.00 0.00

0.92 0.00 0.00

0.00

NORMAL

0.96 0.00

0.00

0.93 0.00 0.00 0.00 0.04 0.04 0.00

0.00

1.00

0.96 0.00 0.00 0.04

0.92 0.08 0.00 0.00 0.00 0.00 0.00

MEL NV BCC AKIEC BKL DF

Predicted class

VASC

CNV

DME DRUSEN NM

0.30 0.00 0.00

0.00 0.20 0.00

0.00 0.00 0.13

0.00 0.00 0.00

0.04

0.04 0.03 0.01

True class

BKL AKIEC BCC

True class

BKL AKIEC BCC

VASC DF

NV MEL

VASC DF

NV MEL

True class

True class

NM

DRUSEN

DME

CNV

NM

DRUSEN

DME

CNV

True class

True class

PNEUMONIA

NORMAL

PNEUMONIA

NORMAL

image datasets (Fig. [5](#_bookmark6)b). For the chest X-ray image data- set, Rf decreased from 81.7 to 46.7%. A Rf of ~ 50% indi- cates that the model classified most images into either

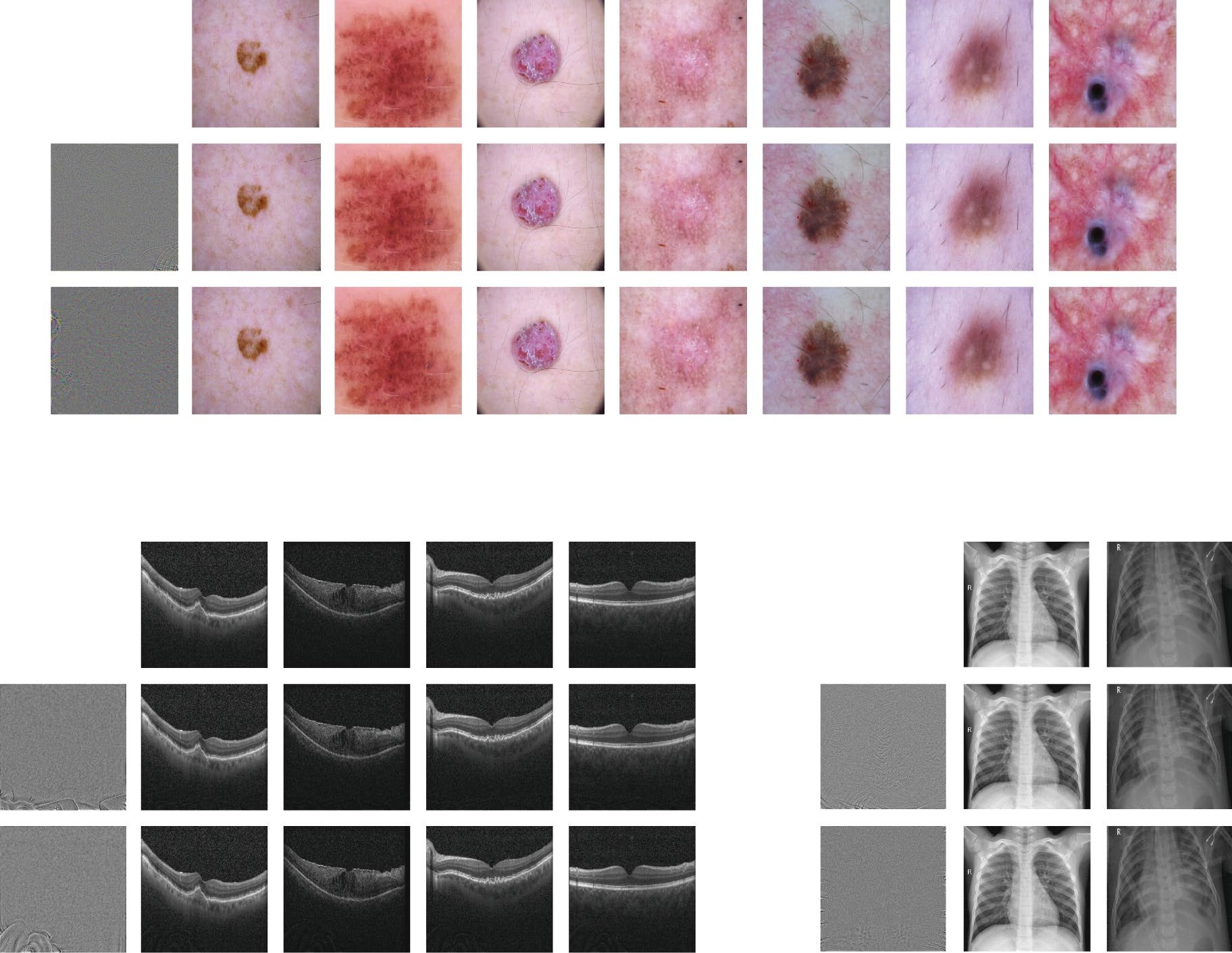
one of two classes; specifically, most images were classi- fied into NORMAL at the fifth iteration (Fig. [5](#_bookmark6)f ). For the skin lesions image dataset, no remarkable decrease in Rf due to adversarial retraining was confirmed; specifically, Rf decreased from 92.2 to 82.1% (Fig. [5](#_bookmark6)a). Most images were classified into MEL at the fifth iteration (Fig. [5](#_bookmark6)c). However, the dominant classes changed for each itera- tion. For example, the dominant classes were AKIEC and

BKL at the third and fourth iterations, respectively (Fig. S15 in Additional file [1](#_bookmark8)).

Figure [6](#_bookmark7) shows the effect of adversarial retraining on the Rs of targeted UAPs with *p* = 2 against the Inception V3 models for the skin lesions, OCT, and chest X-ray image datasets. As representative examples, we considered tar- geted attacks on the most significant cases, namely, MEL, CNV, and PNEUMONIA for the skin lesions, OCT, and chest X-ray image datasets, respectively; ζ = 2% for the skin lesions and chest X-ray image datasets; and ζ = 6% for the OCT image dataset. Adversarial retraining did not affect the test accuracy and reduced Rs for all medical image datasets (Fig. [6](#_bookmark7)a–c). For the OCT and chest X-ray image dataset, Rs decreased from ~ 95% to the baseline Rs (~ 25% and ~ 50%, respectively) after five iterations. For the skin lesions image dataset, Rs decreased from ~ 95 to ~ 30%; however, Rs at the fifth iteration was higher than the baseline (~ 10%). The confusion matrices (Fig. [6](#_bookmark7)d–f)

indicated that adversarial retraining was useful against UAP-based targeted attacks: most images were correctly

classified into the original classes despite the adversarial attacks. However, the effect of adversarial retraining was partially limited for the skin lesions image dataset. For example, 30% of the NV images were still classified into the target class (MEL) despite five iterations of adver- sarial retraining (Fig. [6](#_bookmark7)c). Furthermore, ~ 20% of the BKL and VASC images were still classified into the target class.



(MEL)

(NV)

(BCC)

**a** Skin lesion

(AKIEC) (BKL)

(DF)

(VASC)

UAP

(CNV)

**b** OCT

(DME) (DRUSEN)

(NM)

**c** Chest X-ray

(NORMAL) (PNEUMONIA)

UAP

UAP

**Fig. 4** Targeted UAPs with *p* = 2 against Inception V3 models and their adversarial images for skin lesions (**a**), OCT (**b**), and chest X-ray image datasets. Further, ζ = 2% in **a** and **c**. ζ = 6% in **b**. Labels in brackets beside the images are predicted classes. Original (clean) images were correctly classified into their actual labels. Adversarial images were classified into the target classes. UAPs are emphatically displayed for clarity; in particular, each UAP is scaled by a maximum of 1 and minimum of 0

Original

Adversarial

Original

Adversarial

Original

Adversarial

Targeted to CVN

Targeted to NM

Targeted to MEL

Targeted to NV

Targeted to PNEUMONIA

Targeted to NORMAL

# Discussion

We showed the vulnerability of the DNN models for medical image classification to small UAPs. Previous studies [[10](#_bookmark18), [11](#_bookmark19)] have indicated the vulnerability to adver- sarial attacks toward medical DNNs; however, they were limited to input image-dependent perturbations. In this

study, we demonstrated that almost imperceptible UAPs caused DNN misclassifications. Unlike previous assump- tions, these results indicate that a DNN-based medi- cal image diagnosis is easier to deceive. Adversaries can result in failed DNN-based medical image diagnoses at lower costs (i.e., using a single perturbation). Specifically, they do not need to consider the distribution and diver- sity of input images when attacking DNNs using UAPs, as UPAs are image agnostic.

We demonstrated that nontargeted attacks based on UAPs were possible (Figs. [1](#_bookmark0) and [2](#_bookmark2), Table [1](#_bookmark1)). Most images were classified into a few specific classes for the skin lesions and OCT image (multiclass) datasets. This result is consistent with the existence of domi- nant classes in UAP-based nontargeted attacks [[12](#_bookmark20)].

For the skin lesions image dataset, the AKIEC and DF dominant classes observed in this study may be owing to the imbalanced dataset. The number of AKIEC and DF images is relatively lower than that of the other class



0.00 0.01 0.01 0.00

0.84 0.01

0.00

1.00

0.91 0.09 0.00 0.00 0.00 0.00 0.00

0.97 0.02 0.01 0.00

1.00 0.00 0.00 0.00 0.00 0.00 0.00

**f**

**e**

**d**

5

1 2 3 4

# of iterations for adversarial retraining

**Fig. 5** Effect of adversarial retraining on robustness of nontargeted UAPs with *p* = 2 against Inception V3 models for skin lesions, OCT, and chest

X-ray image datasets. ζ = 4% for the skin lesions and chest X-ray image datasets. ζ = 6% for OCT image dataset. The top panels indicate the scatter plots of fooling rate *Rf* (%) of UAPs versus number of iterations for adversarial retraining. Bottom panels indicate normalized confusion matrices for

fine-tuned models obtained after five iterations of adversarial retraining. These confusion matrices are on adversarial test images

0

0.00 1.00 0.00 0.00

0.93 0.01 0.02 0.03 0.00 0.00 0.00

0.92 0.00 0.01 0.00 0.07 0.00 0.00

0.01 0.05 0.93 0.01

0.78 0.04 0.00 0.00 0.00 0.19 0.00

0.92

0.08

0.00 0.40 0.01 0.58

0.87 0.04 0.00 0.00 0.00 0.00 0.09

MEL NV BCC AKIEC BKL DF VASC

Predicted class

CNV

DME DRUSEN NM

NORMAL

PNEUMONIA

Predicted class

Predicted class

0.4

**a**

1.0

Skin lesion

**b**

OCT

**c**

1.0

Chest X-ray

1.0

0.8

0.8

0.8

0.6

0.6

0.6

5

0.4

0.4

0.2

0.2

0.2

0.0

0.0

0.0

0

1 2 3 4

# of iterations for adversarial retraining

5

0

1 2 3 4

# of iterations for adversarial retraining

0.14

Test accuracy

True class

VASC DF

BKL AKIEC BCC

NV MEL

True class

NM

DRUSEN

DME

CNV

True class

PNEUMONIA

NORMAL

images. As the algorithm considers maximizing Rf , a relatively large Rf is achieved when all inputs are clas- sified into AKIEC and DF owing to UAPs. The use of

imbalanced datasets may be one of the causes of vul- nerability to UAPs. To avoid this problem, domain adaptation [[25](#_bookmark33), [26](#_bookmark34)] may be useful. For the OCT image (binary-class) dataset, the DNN models wrongly pre- dicted the actual labels because of Rf maximization; however, the existence of dominant classes was par- tially confirmed according to the model architecture.

These misclassifications result in false positives and false negatives in medical diagnosis. False positives may cause unwanted mental stress to patients, whereas false negatives may result in significant misdiagnoses involving human lives; specifically, they fail to perform early detection and render therapeutic strategies dif- ficult. Moreover, they can cause the social credibility

of medical doctors and medical organizations to be undermined.

The transferability of nontargeted UAPs across model architectures was limited (Additional file [1](#_bookmark8): Tables S3– S5). This indicates that UAPs are architecture-specific, which is inconsistent with a previous study [[12](#_bookmark20)]. This discrepancy might be due to differences in the image datasets. Specifically, the number of classes (2–7) in the medical image datasets was lower than that (1000) of the dataset used in the previous study. This study partly considered grayscale images, whereas the previ- ous study used colored images only. Transferability may be observed in datasets comprising colored images with more classes. In fact, a weak transferability was observed for the skin lesions image dataset (Additional file [1](#_bookmark8): Table S5).

Furthermore, we showed that targeted attacks based on UAPs were possible in medical image diagnosis (Figs. [3](#_bookmark4) and [4](#_bookmark5), Table [2](#_bookmark3)), although the UAPs were not trans- ferable across model architectures (Additional file [1](#_bookmark8): Tables S6–S8). The results imply that adversaries can

control DNN-based medical image diagnoses. As tar- geted attacks are more realistic, they may result in more significant security concerns compared with nontargeted attacks. In particular, adversaries can obtain any diag- nosis; specifically, they can intentionally cause not only problems resulting from misdiagnosis, but also various social disturbances. As mentioned in a previous study [[10](#_bookmark18)], adversarial attacks can be used for insurance fraud, as well as drug and device approval adjustments, thereby fraudulently providing and obtaining high-quality care when DNNs are used for decision making.



5

0

1 2 3 4

# of iterations for adversarial retraining

5

0

1 2 3 4

# of iterations for adversarial retraining

5

**d**

**e**

**f**

0.96 0.02 0.01 0.00 0.00 0.00 0.00

0.30 0.67 0.02 0.00 0.00 0.00 0.00

0.91

0.09

0.01 0.01 0.97 0.00 0.01 0.00 0.00

0.11 0.00 0.16 0.71 0.00 0.02 0.00

0.22 0.06 0.05 0.03 0.64 0.01 0.00

0.12 0.04 0.08 0.00 0.00 0.76 0.00

0.00

1.00

0.21 0.00 0.05 0.00 0.00 0.00 0.74

MEL NV BCC AKIEC BKL DF VASC

Predicted class

CNV

DME DRUSEN NM

NORMAL

PNEUMONIA

Predicted class

Predicted class

**Fig. 6** Effect of adversarial retraining on robustness of targeted UAPs with *p* = 2 against Inception V3 models for skin lesions, OCT, and chest X-ray image dataset. ζ = 2% for skin lesion and chest X-ray image datasets. ζ = 6% for OCT image dataset. Top panels indicate scatter plots of targeted attack success rate *Rs* (%) of UAPs versus number of iterations for adversarial retraining. Bottom panels indicate normalized confusion matrices for fine-tuned models obtained after five iterations of adversarial retraining. These confusion matrices are on adversarial test images

0.6

**a**

1.0

Skin lesion

**b**

1.0

OCT

**c**

Chest X-ray

1.0

0.8

0.8

0.8

0.6

0.6

1 2 3 4

# of iterations for adversarial retraining

0.4

0.4

0.4

0.2

Test accuracy

0.2

0.2

0.0

0.0

0.0

0

0.99 0.00 0.01 0.00

0.01 0.97 0.01 0.01

0.01 0.00 0.98 0.01

0.00 0.00 0.00 0.99

True class

BKL AKIEC BCC

VASC DF

NV MEL

True class

DRUSEN

CNV

NM

DME

True class

PNEUMONIA

NORMAL

We considered adversarial retraining, which is known to be an effective method for adversarial defenses [[14](#_bookmark22)], to reduce the vulnerability to UAPs. However, the effect of adversarial retraining was limited for nontargeted UAPs (Fig. [5](#_bookmark6)). For targeted attacks, adversarial retraining sig- nificantly reduced the vulnerability to UAPs, but did not completely avoid it (particularly for the skin lesions image dataset, Fig. [6](#_bookmark7)). Additionally, adversarial retrain- ing requires high computational costs, as it is an iterative fine-tuning method. Simpler alternative methods, such as dimensionality reduction (e.g., principle component

analysis), distributional detection (e.g., maximum mean discrepancy), and normalization detection (e.g., dropout randomization) are available; however, they are known to be easily detected as adversarial examples [[15](#_bookmark23)]. Despite the recent development in adversarial defenses, such as regularized surrogate loss optimization [[27](#_bookmark35)], the use of a discontinuous activation function [[28](#_bookmark36)], and improving the generalization of adversarial training with domain adaptation [[29](#_bookmark37)], many promising defense methods have failed [[30](#_bookmark38)]. Defending against adversarial attacks is a cat- and-mouse game [[10](#_bookmark18)]. Furthermore, properties inher- ent to image processing may cause misclassification. For instance, DNN-based image reconstructions are often performed to purify adversarial examples [[31](#_bookmark39)]; however, they cause image artifacts, resulting in misclassifications by DNNs [[32](#_bookmark40)]. It may be difficult to completely avoid security concerns caused by adversarial attacks.

The vulnerability to UAPs was confirmed in various model architectures. Vulnerability to UAPs may be a uni- versal feature in DNNs. However, VGG16 and VGG19 were relatively robust against UAPs compared to the other model architectures. This result is consistent with

the fact that shallower neural networks are more robust against adversarial attacks for the same task [[33](#_bookmark41)]. The use of these model architectures may be a simple solution for avoiding vulnerability to UAPs. However, such a solution may be unrealistic. The effect of the use of these model architectures on the decrease in Rf and Rs was limited

(Tables [1](#_bookmark1) and [2](#_bookmark3)). Simpler models may show a relatively

low prediction performance. Given the tradeoffs between prediction performance and robustness against adversar- ial attacks [[27](#_bookmark35)], it may be difficult to develop DNNs with both high prediction performance and high robustness against UAPs.

Another simple solution for avoiding adversarial attacks is to render DNNs closed source and publicly unavailable; however, this hinders the accelerated devel- opment of medical DNNs and practical applications of DNNs to automated support for clinical diagnosis. Because the amount of medical image data is limited, collaboration among multiple institutions is required to achieve high diagnostic performance [[34](#_bookmark42)]. For similar reasons, medical DNNs are often developed by fine-tun- ing existing DNNs, such as VGG, ResNet, and Inception, pretrained using the ImageNet dataset (i.e., via transfer learning), although a previous study [[34](#_bookmark42)] debated the effect of transfer learning on the improvement in pre- diction performance for medical imaging; consequently, model architectures and model weights may be impor- tant. Furthermore, DNNs are aimed at real-world usage (e.g., automated support for clinical diagnosis). The assumption that DNNs are a closed source and publicly unavailable may be unrealistic. Even if DNNs are black- box (e.g., model architectures and weights are unknown and loss gradient is not accessible), adversarial attacks on DNNs may be possible. Several methods for adversarial attacks on black-box DNNs, which estimate adversarial perturbations using only model outputs (e.g., confidence scores), have been proposed [[35](#_bookmark43)–[37](#_bookmark44)]. The development and operation of secure, privacy-preserving, and feder- ated DNNs are required in medical imaging [[6](#_bookmark14)].

# Conclusion

Our study is the first to show the vulnerability of DNN- based medical image classification to both nontargeted and targeted UAPs. Our findings emphasize that careful consideration is required in the development of DNNs for medical imaging and their practical applications. Inspired by the high prediction performance of DNNs, many studies have applied DNNs to medical image clas- sification; however, they have ignored the vulnerability of UAPs. Our study highlights such facile applications of DNNs. Our findings enhance our understanding of the vulnerabilities of DNNs to adversarial attacks and may help increase the security of DNNs. UAPs are useful for

reliability evaluation and for designing the operation strategy of medical DNNs.

# Supplementary Information

The online version contains supplementary material available at [https://doi.](https://doi.org/10.1186/s12880-020-00530-y) [org/10.1186/s12880-020-00530-y](https://doi.org/10.1186/s12880-020-00530-y).

**A****dditional file 1:** Supplementary tables and figures.

**Abbreviations**

AKIEC: Actinic keratosis/Bowens disease (intraepithelial carcinoma); BCC: Basal cell carcinoma; BKL: Benign keratosis (solar lentigo/seborrheic keratosis/lichen planus-like keratosis); CNV: Neovascular membrane and associated subretinal fluid; DF: Dermatofibroma; DME: Diabetic macular edema with retinal-thicken- ing-associated intraretinal fluid; DNN: Deep neural network; DRUSEN: Multiple drusen present in early age-related macular degeneration; DensNet: Dense convolutional network; FC: Fully connected; FGSM: Fast gradient sign method; ISIC: International Skin Imaging Collaboration; MEL: Melanoma; NM: Normal retina with preserved foveal contour and absence of retinal fluid/edema; NV: Melanocytic nevus; OCT: Optical coherence tomography; ResNet: Residual network; UAP: Universal adversarial perturbation; VASC: Vascular lesion; VGG: Visual geometry group.

**Acknowledgements**

We would like to thank Editage ([www.editage.jp](http://www.editage.jp/)) for the English language editing.

**Authors’ contributions**

KT conceived and designed the study. HH and AM prepared the data and models. HH coded and performed the experimental evaluation. HH and KT interpreted the results. HH and KT wrote the manuscript. All authors

provided final approval for publication. All authors read and approved the final manuscript.

**Funding**

No funding was received.

**Availability of data and materials**

All data generated and analyzed during this study are included in this pub- lished article and its supplementary information files. The code and data used in this study are available from our GitHub repository: github.com/hkthirano/ MedicalAI-UAP.

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare that they have no competing interests.

Received: 2 September 2020 Accepted: 26 November 2020

