algorithms

MDPI

Article

Simple Black-Box Universal Adversarial Attacks on Deep Neural Networks for Medical Image Classification

Kazuki Koga and Kazuhiro Takemoto *

Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Fukuoka 820-8502, Japan; koga.kazuki835@mail.kyutech.jp
* Correspondence: takemoto@bio.kyutech.ac.jp; Tel.: +81-948-29-7822

Abstract: Universal adversarial attacks, which hinder most deep neural network (DNN) tasks using only a single perturbation called universal adversarial perturbation (UAP), are a realistic security threat to the practical application of a DNN for medical imaging. Given that computer-based systems are generally operated under a black-box condition in which only input queries are allowed and outputs are accessible, the impact of UAPs seems to be limited because well-used algorithms for generating UAPs are limited to white-box conditions in which adversaries can access model parameters. Nevertheless, we propose a method for generating UAPs using a simple hill-climbing search based only on DNN outputs to demonstrate that UAPs are easily generatable using a relatively small dataset under black-box conditions with representative DNN-based medical image classifications. Black-box UAPs can be used to conduct both nontargeted and targeted attacks. Overall, the black-box UAPs showed high attack success rates (40–90%). The vulnerability of the black-box UAPs was observed in several model architectures. The results indicate that adversaries can also generate UAPs through a simple procedure under the black-box condition to foil or control diagnostic medical imaging systems based on DNNs, and that UAPs are a more serious security threat.

Keywords: black-box algorithm; deep neural networks; adversarial attacks; medical imaging

Citation: Koga, K.; Takemoto, K. Simple Black-Box Universal Adversarial Attacks on Deep Neural Networks for Medical Image Classification. *Algorithms* **2022**, *15*, 144. https://doi.org/10.3390/a15050144

Academic Editor: Huan Zhang

Received: 23 March 2022
Accepted: 21 April 2022
Published: 22 April 2022

1. Introduction

Adversarial examples [1–4] are inputs (e.g., images) perturbated with specific extremely small patterns, leading to the misclassification of deep neural networks (DNNs), thus questioning the generalization ability of a DNN, limiting its practical application in safety- and security-critical environments, and reducing model interpretability [4–7]. In particular, because the diagnostic performance of DNN-based systems is equivalent to or higher than that of healthcare professionals, such systems are beginning to be used as diagnostic medical imaging systems [8,9]. The existence of adversarial examples can cause serious security [10] and social problems [11] because disease diagnosis is high-stake decision making. To avoid these problems (e.g., to evaluate the vulnerability of DNNs to adversarial attacks), the development of methods for generating adversarial perturbations is required.

Adversarial attacks first proposed in the literature were input-dependent [1,2]. However, such input-dependent attacks are difficult to perform because they need to compute a different adversarial perturbation for each input (i.e., they require high computational costs); thus, universal adversarial perturbations (UAPs) [12,13] are more realistic adversarial attacks. A UAP is a single input-agnostic perturbation. Adversaries can foil most DNN tasks using only a single UAP (i.e., at lower costs); specifically, they do not need to pay attention to input queries. Moreover, UAPs can be used for both nontargeted [12] and targeted attacks [13]. Adversarial attacks based on UAPs are more practical for adversaries because they can be easily performed in more realistic environments [12–14].

However, widely used UAP algorithms [12,13] are limited to white-box conditions in which adversaries can use model parameters (e.g., the gradient of the loss function)

and the training data. Given that DNN-based systems are generally operated under black-box conditions (i.e., closed-source software and closed application programming interfaces in which only input queries are allowed and outputs are accessible) in terms of security, white-box attacks may be non-threatening. However, adversarial perturbations can also be generated under black-box conditions; specifically, they are generally obtained only using model outputs (e.g., confidence scores). The zeroth-order optimization (ZOO) method [15] generates adversarial examples by estimating the loss gradients of a targeted DNN from their model outputs. Black-box adversarial perturbations can be generated in the context of optimization problems (i.e., minimizing the confidence score for the correct label for nontargeted attacks and maximizing the score for the target class for targeted attacks). As an example, a genetic algorithm-based method has been proposed [16]. Black-box adversarial perturbations are generated with a low query cost using low-frequency noise. [17]. Moreover, the simple black-box adversarial attack (SimBA) method considers a simple iterative search algorithm with random hill climbing to generate adversarial perturbation under the black-box condition [18]. However, these black-box attack methods are limited to input-dependent adversarial attacks. In addition, generative network models can be used to generate UAPs under black-box conditions [19]; however, they require a high computational cost (it is costly to train the network models). The Fourier basis directions can be used as black-box UAPs [20]. However, this method is limited to convolutional neural networks; moreover, it does not allow targeted attacks. A simpler and more effective algorithm for generating UAPs under black-box conditions is required.

Thus, herein, we propose a simple method for searching for black-box UPAs inspired by the SimBA method [18]; specifically, we extended the SimBA method to allow us to generate UAPs for both nontargeted and targeted attacks. To demonstrate the validity of the proposed method, according to our previous studies [14,21], three representative DNN-based medical image classifications were considered because of the importance of adversarial attacks on medical machine learning [11], that is, skin cancer classification using skin lesion images [22], diabetic macular edema classification using optical coherence tomography (OCT) images [23], and pneumonia classification using chest X-ray images [23]. Instead of the white-box conditions in [14,21], we evaluated how much DNN models with several architectures are vulnerable to black-box UAPs for both nontargeted and targeted attacks.

2. Materials and Methods

2.1. Simple Black-Box Algorithm for UAPs

Our black-box algorithm (Algorithm 1) is an extension of the SimBA method [18], which generates input-dependent perturbations to generate UAPs; in particular, we combined the SimBA method with simple iterative methods for generating UAPs under white-box conditions [12,13]. Similar to the SimBA method, our algorithm considers a black-box classifier C that returns the confidence score output probability $p_C(y|\boldsymbol{x})$ of class y given input image $\boldsymbol{x}$. Here, $C(\boldsymbol{x})$ indicates the class or label (with the highest confidence score) for $\boldsymbol{x}$ (i.e., $\arg\max_{y'} p_C(y'|\boldsymbol{x})$). The algorithm starts with $\boldsymbol{\delta} = \mathbf{0}$ (i.e., no perturbations) and iteratively updates the UAP $\boldsymbol{\delta}$ using a direction $\boldsymbol{q}$ randomly sampled from a set $\boldsymbol{Q}$ of search directions with attack strength ϵ under the constraint in which the L_p norm of the UAP is equal to or less than a small value ξ (i.e., $\delta_p \leq \xi$). For project (δ, p, ξ), a projection function is used to satisfy the constraint and is specifically defined as follows:

$$\text{project}(\boldsymbol{\delta}, p, \xi) = \arg\min_{\boldsymbol{\delta}^*} ||\boldsymbol{\delta} - \boldsymbol{\delta}^*||_2 \quad \text{s.t.} \quad ||\boldsymbol{\delta}^*||_p \leq \xi \tag{1}$$

For nontargeted attacks, the updated UAP vector $\boldsymbol{\delta}'$ is accepted if the update from $\boldsymbol{\delta}$ to $\boldsymbol{\delta}'$ contributes to decreasing the confidence scores for their labels $C(\boldsymbol{x})$ predicted without

the UAP of any $\boldsymbol{x}$ in an input image set $\boldsymbol{X}$. As shown in Algorithm 1, the update is accepted if the following simple condition is satisfied (true):

$$\sum_{\boldsymbol{x} \in \boldsymbol{X}} p_C(C(\boldsymbol{x})|\boldsymbol{x} + \boldsymbol{\delta}') < \sum_{\boldsymbol{x} \in \boldsymbol{X}} p_C(C(\boldsymbol{x})|\boldsymbol{x} + \boldsymbol{\delta}). \tag{2}$$

This corresponds to considering the average search directions obtained from input images to generate a UAP; thus, it is analogous to an approach [24] to generate a UAP using the norm of the loss gradients estimated from the outputs of a hidden layer with respect to inputs.

Targeted UAPs can also be implemented by modifying this condition (electronic Supplementary Material, Algorithm S1). Specifically, in this case, because the algorithm accepts an updated UAP vector if the update overall contributes to increasing the confidence score for a target class y of any $\boldsymbol{x}$ in $\boldsymbol{X}$, the update is accepted if

$$\sum_{\boldsymbol{x} \in \boldsymbol{X}} p_C(y|\boldsymbol{x} + \boldsymbol{\delta}') > \sum_{\boldsymbol{x} \in \boldsymbol{X}} p_C(y|\boldsymbol{x} + \boldsymbol{\delta}). \tag{3}$$

Several types of search directions $\boldsymbol{Q}$ were considered. As mentioned in a previous study [18], a natural choice for $\boldsymbol{Q}$ may be the standard basis $\boldsymbol{Q} = \boldsymbol{I}$, for which $\boldsymbol{q}$ corresponds to a random pixel attack. However, this may be less effective for inputs with a large space (i.e., large images). In this case, a discrete cosine basis is useful because random noise in a low-frequency space contributes to adversarial attacks [17]. The set $\boldsymbol{Q}_{\mathrm{DCT}}$ of the orthonormal frequencies was extracted using the discrete cosine transform. Assuming a two-dimensional image space $\mathbb{R}^{d \times d}$, $\boldsymbol{Q}_{\mathrm{DCT}}$ has $d \times d$ frequencies; however, a fraction f_d of the lowest frequency directions is only used to generate the UAP more effectively (faster).

This update procedure terminates when the attack success rate for $\boldsymbol{X}$ is 100%, or the number of iterations reaches the maximum $i_{\max}$. When generating a nontargeted UAP $\boldsymbol{\delta}_{\mathrm{nt}}$ the attack success rate corresponds to the fooling rate R_f; that is, the fraction of input images for which their predicted labels are altered due to adversarial attacks to all input images in set $\boldsymbol{X}$, $R_f = |\boldsymbol{X}|^{-1} \sum_{\boldsymbol{x} \in \boldsymbol{X}} \mathbb{I}(C(\boldsymbol{x}) \neq C(\boldsymbol{x} + \boldsymbol{\delta}_{\mathrm{nt}}))$, where the function $\mathbb{I}(A) = 1$ if condition A is true, and $\mathbb{I}(A) = 0$ otherwise. When generating a targeted UAP $\boldsymbol{\delta}_{\mathrm{t}}$, the attack success rate corresponds to the targeted attack success rate $R_s = |\boldsymbol{X}|^{-1} \sum_{\boldsymbol{x} \in \boldsymbol{X}} \mathbb{I}(C(\boldsymbol{x} + \boldsymbol{\delta}_{\mathrm{t}}) = y)$; that is, the fraction of input images predicted as target class y due to adversarial attacks to all input images in set $\boldsymbol{X}$.

Our algorithm (please see Algorithm 1) was implemented using the adversarial robustness toolbox (ART, version 1.7.0; github.com/Trusted-AI/adversarial-robustness-toolbox, accessed on 1 February 2021).

Algorithm 1 Computation of a nontargeted UAP

Input: Set X of input images, classifier C, set Q of search directions, attack strength ϵ , cap ξ on L_p norm of the perturbation, norm type p (1, 2, or ∞), maximum number $i_{\max}$ of iterations.

Output: nontargeted UAP vector δ

$\delta \leftarrow \mathbf{0}$, $r \leftarrow 0$, $i \leftarrow 0$
while $r < 1$ and $i < i_{\max}$ **do**
  Pick a direction randomly: $\boldsymbol{q} \in \boldsymbol{Q}$
  for $\alpha \in \{-\epsilon, \epsilon\}$ **do**
    $\boldsymbol{\delta}' \leftarrow \mathrm{project}(\boldsymbol{\delta} + \alpha \boldsymbol{q}, p, \xi)$
    if $\sum_{\boldsymbol{x} \in \boldsymbol{X}} p_C(C(\boldsymbol{x})|\boldsymbol{x} + \boldsymbol{\delta}') < \sum_{\boldsymbol{x} \in \boldsymbol{X}} p_C(C(\boldsymbol{x})|\boldsymbol{x} + \boldsymbol{\delta})$ **then**
      $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta}'$
      break
    end if
  end for
  $r \leftarrow |\boldsymbol{X}|^{-1} \sum_{\boldsymbol{x} \in \boldsymbol{X}} \mathbb{I}(C(\boldsymbol{x}) \neq C(\boldsymbol{x} + \boldsymbol{\delta}))$
  $i \leftarrow i + 1$
end while

2.2. Medical Images and DNN Models

To evaluate the performance of black-box UAPs generated using our algorithm, we used the medical image datasets and DNN models in our previous studies on DNN-based medical image classifications [14,21]. The datasets contain the skin lesion images (in seven classes) for skin cancer classification, OCT images (in four classes) for referable diabetic retinopathy classification, and chest X-ray images (in binary classes) for pneumonia classification (also see Table S1 for the details of the class labels). All images had a pixel resolution of 299 × 299 pixels. The skin lesion images are red–green–blue, whereas the OCT and chest X-ray images are in grayscale. We used the test images to generate UAPs and evaluate the UAP performance because the black-box condition assumes that adversaries cannot access the training data. The skin lesion, OCT, and chest X-ray image datasets contained 3015, 3360, and 540 test images, respectively. Note that the skin lesion image dataset was only class-imbalanced (also see Table S1).

To investigate the relationship between the model architecture and UAP performance, following [14] and [21], we used the Inception V3 architecture [25], Visual Geometry Group-16 (VGG16) [26], and Residual Network with 50 layers (ResNet50) [27] architectures. The DNN models were obtained using transfer learning from the ImageNet dataset [28] (see [14] for the test accuracies of the DNN models).

2.3. Generating UAPs

A portion of the test images in the dataset was used as input images to generate UAPs. For the OCT image dataset, 800 test images (200 randomly selected images per class) were used. For the chest X-ray image dataset, 200 test images (100 images randomly selected per class) were used. For the skin lesion image dataset, 1000 randomly selected test images were used. We considered $\epsilon = 0.5$ and $i_{\max} = 5000$. As the set of search directions, $\boldsymbol{Q}_{\mathrm{DCT}}$ was considered, where f_d was set to ~9.4% (28/299). The parameters ϵ and f_d were selected using a grid search to maximize the performance of the UAPs for the input images. The ratio ζ of the L_p norm of the UAP to the average L_p norm of an image in the dataset (see [14] for details) was used to determine the cap parameter ξ. Following [14], $\zeta = 4\%$ for the skin lesion and chest X-ray image dataset, and $\zeta = 8\%$ for the OCT image dataset.

2.4. Evaluating the Performance of UAPs

Following [14] and [21], R_f and R_s were used to evaluate the performances of a nontargeted UAP and a targeted UAP, respectively. In addition, R_f and R_s of a UAP were computed using the validation dataset, which consists of the rest of the test images (i.e., the test images excluded the images used to generate the UAP) in each dataset. However, when evaluating the effect of the number of images used to generate a UAP on the performance of the UAP (Section 3.3), a validation dataset of the same size was used to evaluate the performance, and a fixed number of test images were randomly selected from the remaining images (validation dataset) for each medical dataset. Random UAPs, vectors randomly sampled from a sphere with a specific radius [12], were used as random controls. As mentioned in [14], note that R_s has a baseline (i.e., nonzero R_s observed without UAPs).

To evaluate how the predicted labels changed for each class due to the UAPs, the confusion matrices on the validation dataset (i.e., the rest of the test images used to generate a UAP) were also obtained. Normalized confusion matrices were computed to account for the imbalanced datasets.

3. Results

3.1. Nontargeted Attacks Using Black-Box UAPs

The performance of non-targeted UAPs was evaluated (Table 1). Overall, the fooling rate R_f of the UAPs was significantly higher than that of the random UAPs (random controls) and was hardly influenced by the randomness of the algorithm (i.e., random seed setting; see Table S2). Moreover, the difference between the clean images and their adversarial versions through the use of UAPs was almost imperceptible (Figure 1). The results

indicate that small UAPs are generatable under black-box conditions. For each medical image dataset, however, the performance (R_f) might depend on the model architectures and norm type p of the UAPs. For the skin lesion image dataset, R_f achieved a score of >70% when $p = 2$ regardless of the model architecture; however, it was relatively low (~35%) for the UAPs with $p = \infty$ against the ResNet50 and VGG16 models. A similar R_f (~65%) was also observed for the Inception V3 model when $p = \infty$. For the OCT image dataset, R_f of the UAP with $p = \infty$ was higher than that of the UAP with $p = 2$; in particular, the UAP with $p = 2$ against the ResNet50 model was less effective in causing a misclassification of the DNN model, although it still showed a slightly higher R_f in comparison to random UAPs. For the chest X-ray image dataset, R_f (~40%) of the UAPs against the Inception V3 model was slightly lower than that (~50%) of the UAPs against the ResNet50 and VGG16 models, regardless of the norm type.

Table 1. Fooling rates R_f (%) of UAPs for non-targeted attacks against DNN models for medical image datasets. Values in brackets indicate random controls (R_f of random UAPs).

Dataset/Architecture	Skin Lesion		OCT		Chest X-ray	
	$p = 2$	$p = \infty$	$p = 2$	$p = \infty$	$p = 2$	$p = \infty$
Inception V3	78.8 (13.6)	65.6 (10.2)	31.7 (1.6)	44.9 (3.3)	41.8 (2.1)	44.1 (2.6)
ResNet50	71.9 (11.1)	33.9 (8.6)	5.5 (1.3)	69.3 (4.3)	51.5 (5.9)	50.9 (6.2)
VGG16	76.6 (5.3)	38.9 (3.6)	40.9 (0.7)	75.1 (2.0)	50.0 (1.8)	50.0 (2.4)

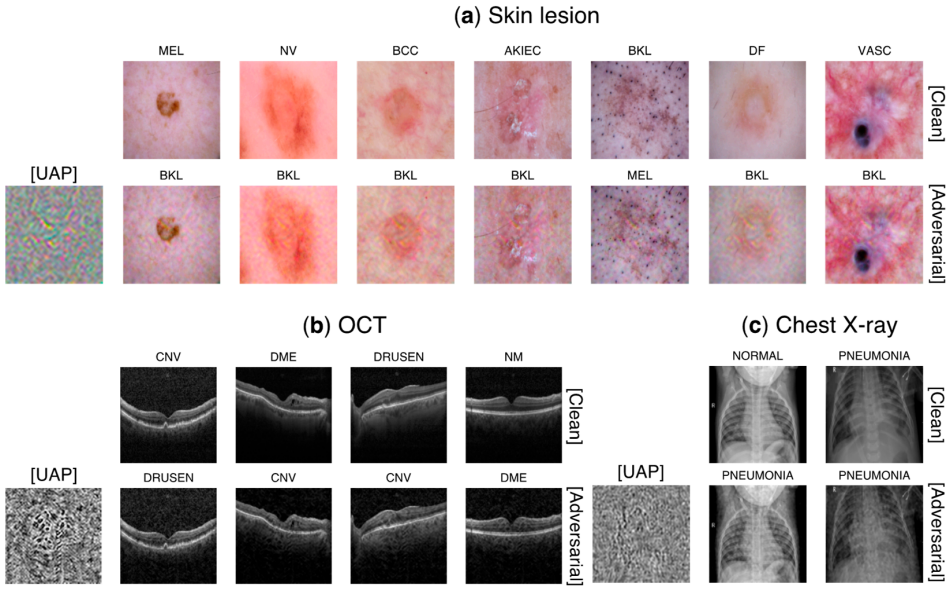

Figure 1. Clean images and their adversarial images generated using nontargeted UAPs against Inception V3 model, which has been widely used in previous studies on DNN-based medical imaging (e.g., [22,23]) for (**a**) skin lesion, (**b**) OCT, and (**c**) chest X-ray image classifications. Here, $p = 2$ in (**a**,**c**) $p = \infty$ in (**b**) in terms of the UAP performance (Table 1). Labels (without square brackets) next to the images are the predicted classes (see Table S1 for details). Each UAP is scaled by a maximum of 1 and a minimum of zero to visually emphasize UAPs.

The confusion matrices (Figure 2) indicate that the adversarial images were classified into a few specific classes (e.g., dominant classes); however, the dominant classes might be different between the model architectures.

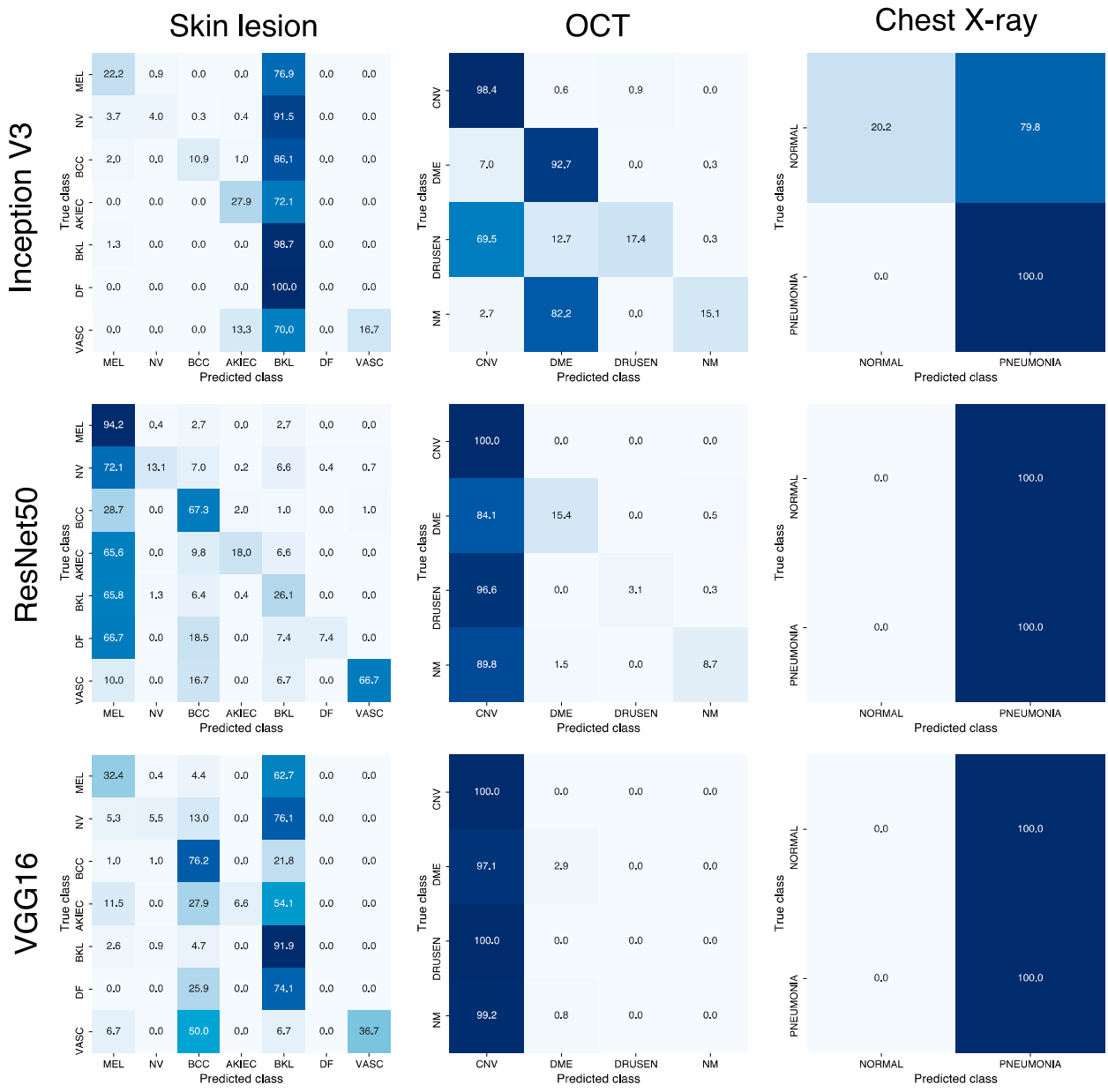

Figure 2. Normalized confusion matrices for Inception V3, ResNet50, and VGG16 models attacked using nontargeted UAPs for skin lesions, OCT, and chest X-ray image datasets. Here, $p = 2$ for skin lesion and chest X-ray image datasets, and $p = \infty$ for the OCT image dataset. See Table S1 for the abbreviations of the class labels.

3.2. Targeted Attacks Using UAPs

The performance of targeted UAPs was evaluated (Table 2), where $p = 2$ for the skin lesion and chest X-ray image datasets, and $p = \infty$ for the OCT image dataset, when considering the performance of nontargeted UAPs (Table 1). Following our previous study [14], we selected the target classes for each medical image dataset (also see Table S1): the most significant cases (hereinafter called 'cases' for simplicity) and controls.

Table 2. Target attack success rates R_s (%) of UAPs for targeted attacks against the DNN models for medical image datasets. Values in brackets indicate random controls (R_s of random UAP).

Dataset Target Class/ Architecture	Skin Lesion		OCT		Chest X-ray	
	Control	Case	Control	Case	Control	Case
Inception V3	63.8 (64.8)	60.9 (10.4)	27.3 (27.2)	92.2 (25.2)	67.1 (54.4)	91.8 (45.9)
ResNet50	76.0 (66.6)	81.2 (10.6)	28.6 (28.3)	93.5 (24.8)	53.8 (57.6)	97.6 (42.4)
VGG16	80.0 (72.4)	79.0 (7.7)	25.1 (26.5)	97.9 (24.6)	97.4 (51.5)	97.1 (48.5)

Overall, the targeted attacks on the cases were successful for all types of medical image classification. The UAP performance ($R_s = 60\%$ to 95%) of the UAPs was significantly high compared to the random controls. For the OCT and chest X-ray image datasets, R_s was independent of the model architecture. However, a weak model architecture dependency of R_s was observed for the skin lesion image dataset. The value of R_s (~60%) of the UAP against the Inception V3 model was slightly lower than that (~80%) of the UAP against the ResNet50 and VGG16 models.

In contrast, attacks targeting the controls were only partly successful. For the skin lesion image dataset, R_s of the UAP were 75% to 80% for the ResNet50 and VGG16 models, which is significantly higher than that of the random controls; however, it is similar to that of the random control for Inception V3, indicating that the targeted attacks failed. For the OCT image dataset, the values of R_s of the UAP were also almost equivalent to those of random UAPs, regardless of the model architecture. For the chest X-ray image dataset, the performance of UAPss was remarkably high for the VGG16 model (R_s ~95%); however, R_s was relatively low (~65%) for the Inception V3 model compared to the UAP against the VGG16 model, although R_s was higher than that of the random control. For the ResNet50 model, R_s was similar to that of the random control.

3.3. Effect of the Input Dataset Size on the UAP Performance

Finally, we investigated the effect of the size of the input dataset used to generate a UAP on UAP performance. As a representative example, in terms of such performance, we focused on the skin lesion image dataset and evaluated the performance of nontargeted UAPs with $p = 2$. The UAPs were generated using an input dataset consisting of N images randomly selected from the test images in the medical image dataset. Here, we considered $N = 100$, $N = 500$, and $N = 1000$. The value of R_f was computed using the input and validation datasets. The validation dataset consisted of 2015 images randomly selected from the test images, excluding the images in the input dataset.

As shown in Figure 3, R_f increases with N. In addition, R_f for the validation dataset was lower than that for the input dataset; however, the difference in R_f between the validation dataset and input dataset decreased with an increase in N. These tendencies were observed regardless of the model architecture. The results indicate that UAPs achieve high performance and generalization, resulting in a misclassification of the DNN model with an increasing in N. However, the UAPs showed a relatively high performance despite a small N. When $N = 100$, R_f for the validation dataset was >50% for the Inception V3 and VGG models, although it was ~30% for the ResNet50 model. Moreover, R_f for the validation dataset was >60%, regardless of the model architecture, when $N = 500$.

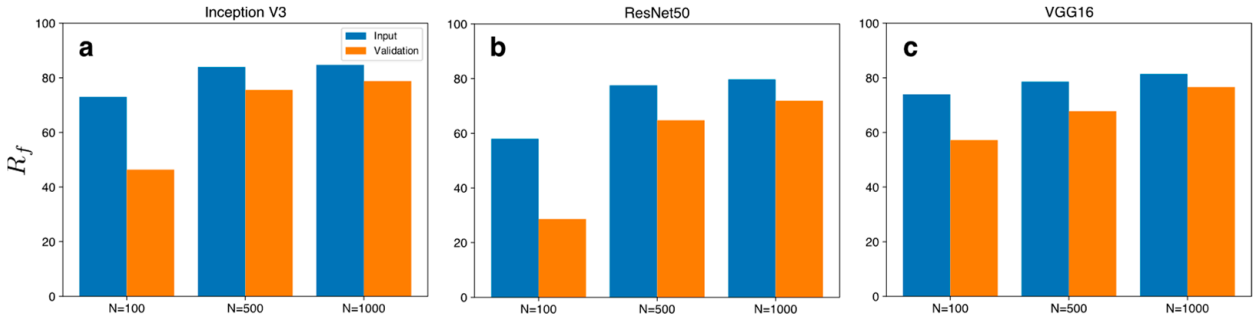

Figure 3. Effect of input dataset size on the fooling rates R_f (%) of UAPs against Inception V3 (**a**), ResNet50 (**b**), and VGG16 models (**c**) for skin lesion image dataset. The values of R_f for both the input and validation datasets are shown.

4. Discussion

We proposed a simple method for generating UAPs under the black-box condition inspired by the SimBA method [18] and applied this method to DNN-based medical image classification. Overall, the results showed that small (almost imperceptible) UAPs are generatable under the black-box condition using only a simple hill-climbing search based on the model outputs (i.e., confidence scores) to perform both nontargeted and targeted attacks. The vulnerability of black-box UAPs was observed in several model architectures, indicating the versatility of our method; thus, the vulnerability may be a general property of a DNN. The results indicate that adversaries can foil or control DNN tasks even if they never use model parameters (e.g., loss gradients) and training data because target systems are operated under black-box conditions in terms of security. DNN-based medical image diagnoses may be easy to conduct in a realistic environment.

It can be argued that the performance of black-box UAPs generated using our method compares favorably with that of white-box UAPs [12,13]. In addition to the model parameters, the white-box UAPs were generated using a larger dataset (7–10 times) than that of the black-box UAPs. For instance, for the nontargeted UAPs with $p = 2$ and $\zeta = 4\%$ against the Inception V3 model for the skin lesion image dataset, R_f of the black-box UAP was ~80% (Table 1), whereas that of the white-box UAP was ~90% (see Table 1 in [14]), where 1000 test images and 7000 training images were used to generate the black-box and white-box UAPs, respectively. For nontargeted attacks, overall, the values of R_f of the black-box UAPs were 40% to 80% (Table 1, except for the nonsignificant cases in which R_f is almost similar to the random control), whereas R_f of the white-box UAPs was 70% to 90% [14]. For targeted attacks, the values of R_s of the black-box UAPs were 60% to 90% (Table 2, except for the controls), whereas those of the white-box UAPs were >90% (see Table 2 in [14]).

Given that the DNNs for medical image classification are vulnerable to white-box (i.e., loss gradient-based) UAPs [14], it is expected that black-box UAPs can also be generated by estimating loss gradients from model outputs. The white-box methods for generating UAPs [12,13], used in [14], iteratively update a UAP vector using loss-gradient-based methods (e.g., the fast gradient sign method [2] and DeepFool [29]). Thus, instead of such white-box methods, black-box methods that craft an input-dependent adversarial perturbation by estimating the loss gradient from model outputs (e.g., the ZOO method [15]) can be used in the iterative algorithms to generate black-box UAPs. However, this approach may be unrealistic in terms of computational time because crafting an adversarial perturbation while estimating loss gradients is time-consuming in black-box methods. Therefore, we considered an extension of SimBA (i.e., a simple hill-climbing approach) as an alternative approach.

In general, the number of queries per fixed time is limited in terms of software and application programming interfaces in terms of security; thus, query-efficient methods are needed for input-dependent adversarial attacks. Our proposed method generated UAPs with a relatively high performance despite a small number of queries because it is based

on the SimBA method [18], which is a query-efficient approach. The upper of the overall number of queries is the same for the SimBA method when considering adversarial attacks on classification tasks for an image set; specifically, it is the product of the maximum number of iterations ($i_{\max}$) and the number of input images (N). Because the SimBA method needs to generate an adversarial perturbation per image, the upper part of the overall number of queries is the product of the maximum number of queries per image, which corresponds to $i_{\max}$, and the size of the image set. However, our method may be more efficient than the SimBA method for adversarial attacks on classification tasks for an image set. Our method can generate black-box UAPs using a relatively small number of input images and is effective for adversarial attacks on classification tasks for many other images (Figure 3) because UAPs are input-agnostic. Assuming that the number of input images is lower than the size of the image set, the SimBA method may require a larger number of queries because it generates adversarial perturbation per image in the set. In addition, given that UAPs are input-agnostic, the number of queries required for generating UAPs poses a few problems because UAPs are also useful for adversarial attacks to classification tasks for many other images, even if they require generating many queries. Black-box UAPs have an advantage in that adversaries do not need to be concerned with the query efficiency in black-box attack methods, in addition to the advantages mentioned in Section 1.

Black-box UAPs are likely useful for avoiding adversarial defenses. For example, a discontinuous activation function is often used to make it difficult for adversaries to estimate the loss gradients [30]; however, black-box attack methods have led to many defense methods [31], including the use of a discontinuous activation function. Black-box attacks based on UAPs may provide insight into the development of more efficient defense methods.

However, the performance of black-box UAPs is limited in some cases. For nontargeted attacks on the models for the chest X-ray image dataset, the values R_f of the black-box UAPs were ~50% at most (Table 1 and Figure 2), whereas those of the white-box UAPs were ~80% at most (the models incorrectly predicted the true labels; Table 1 in [14]). In addition, the success rates of the targeted attacks on the controls were mostly equivalent to the random controls (Table 2), although the white-box UAPs achieved a high performance (Table 2 in [14]). This is because the black-box attack method only utilizes limited information to generate UAPs compared with white-box methods. Moreover, this may be due to an unbalanced dataset. The images in the control class were abundant in the skin lesion image dataset (also see Table S1). The proposed algorithm considers maximizing the confidence score for a targeted class; thus, the algorithm may stop to search for UAPs for targeted attacks to an abundant label in a dataset because a large R_s will have already been achieved. Simple solutions for improving the performance include the use of more input images (e.g., as shown in Figure 3 and using data augmentation) and considering a larger number of queries. Another solution is to consider different types of search directions. For example, the Fourier basis [20], texture bias [32], and Turing patterns [33] may be useful for efficiently searching for UAPs because they are also useful for universal adversarial attacks.

Our proposed method is limited to a black-box condition in which the confidence scores for all labels are available. A harder black-box condition can also be considered, that is, a case in which the classifiers return the predicted label only. For example, the boundary attack method [34] generates input-dependent adversarial perturbations based on a decision-based attack that starts from a large adversarial perturbation and then seeks to reduce the perturbations while remaining adversarial. This method requires a relatively large number of model queries; however, the HopSkipJumpAttack method [35] can conduct decision-based adversarial attacks with significantly fewer model queries by using binary information at the decision boundary. Nevertheless, this study considered confidence-score-based black-box attacks because the use of confidence scores is important in deciding whether to trust a classifier's decision in terms of machine learning trust and safety [36] (for healthcare in particular [37,38]). However, it is important to evaluate whether adversarial attacks are possible under hard black-box conditions in terms of the reliability and safety of

a DNN. Although further investigations are needed, our algorithm may provide insight into the development of decision-based universal adversarial attacks.

5. Conclusions

We proposed a simple method for generating UAPs under black-box conditions and demonstrated that black-box UAPs could be generated easily using a relatively small dataset. Our finding that adversaries can generate UAPs under the black-box condition using a simple procedure provides insight into increasing the reliability and safety of a DNN and designing its operational strategy (for medical imaging in particular [10]).

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/a15050144/s1, Algorithm S1: Computation of a targeted UAP; Table S1; List of the abbreviations of the image labels and label composition in each medical image dataset. Table S2: Fooling rates R_f (%) of nontargeted UAPs with $p = 2$ against Inception V3 model for skin lesion images dataset under different random seed settings.

Author Contributions: Conceptualization, K.K. and K.T.; methodology, K.K. and K.T.; software, K.K. and K.T.; validation, K.K. and K.T.; formal analysis, K.K. and K.T.; investigation, K.K. and K.T.; resources, K.K.; data curation, K.K.; writing—original draft preparation, K.T.; writing—review and editing, K.K. and K.T.; visualization, K.T.; supervision, K.T.; project administration, K.T.; funding acquisition, K.T. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by JSPS KAKENHI (grant number JP21H03545).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The code and data used in this study are available from our GitHub repository: github.com/kztakemoto/U-SimBA (accessed on 26 October 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.J.; Fergus, R. Intriguing properties of neural networks. In Proceedings of the 2nd International Conference on Learning Representations, {ICLR} 2014, Banff, AB, Canada, 14–16 April 2014.
2. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
3. Yuan, X.; He, P.; Zhu, Q.; Li, X. Adversarial examples: Attacks and defenses for deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 2805–2824. [CrossRef] [PubMed]
4. Ortiz-Jimenez, G.; Modas, A.; Moosavi-Dezfooli, S.-M.; Frossard, P. Optimism in the face of adversity: Understanding and improving deep learning through adversarial robustness. *Proc. IEEE* **2021**, *109*, 635–659. [CrossRef]
5. Matyasko, A.; Chau, L.-P. Improved network robustness with adversary critic. In Proceedings of the 32nd International Conference on Neural Information Processing Systems 2018, Montreal, QC, Canada, 3–8 December 2018; pp. 10601–10610.
6. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
7. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; pp. 39–57. [CrossRef]
8. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
9. Liu, X.; Faes, L.; Kale, A.U.; Wagner, S.K.; Fu, D.J.; Bruynseels, A.; Mahendiran, T.; Moraes, G.; Shamdas, M.; Kern, C.; et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *Lancet Digit. Health* **2019**, *1*, e271–e297. [CrossRef]
10. Kaissis, G.A.; Makowski, M.R.; Rückert, D.; Braren, R.F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2020**, *2*, 305–311. [CrossRef]
11. Finlayson, S.G.; Bowers, J.D.; Ito, J.; Zittrain, J.L.; Beam, A.L.; Kohane, I.S. Adversarial attacks on medical machine learning. *Science* **2019**, *363*, 1287–1289. [CrossRef] [PubMed]
12. Moosavi-Dezfooli, S.M.; Fawzi, A.; Fawzi, O.; Frossard, P. Universal adversarial perturbations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21 July–26 July 2017; pp. 86–94. [CrossRef]
13. Hirano, H.; Takemoto, K. Simple iterative method for generating targeted universal adversarial perturbations. *Algorithms* **2020**, *13*, 268. [CrossRef]

14. Hirano, H.; Minagi, A.; Takemoto, K. Universal adversarial attacks on deep neural networks for medical image classification. *BMC Med. Imaging* **2021**, *21*, 9. [CrossRef] [PubMed]
15. Chen, P.-Y.; Zhang, H.; Sharma, Y.; Yi, J.; Hsieh, C.-J. ZOO: Zeroth Order Optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*; ACM: New York, NY, USA, 2017; pp. 15–26.
16. Chen, J.; Su, M.; Shen, S.; Xiong, H.; Zheng, H. POBA-GA: Perturbation optimized black-box adversarial attacks via genetic algorithm. *Comput. Secur.* **2019**, *85*, 89–106. [CrossRef]
17. Guo, C.; Frank, J.S.; Weinberger, K.Q. Low frequency adversarial perturbation. In Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, {UAI} 2019, Tel Aviv-Yafo, Israel, 22–25 July 2019; Volume 115, pp. 1127–1137.
18. Guo, C.; Gardner, J.R.; You, Y.; Wilson, A.G.; Weinberger, K.Q. Simple black-box adversarial attacks. In Proceedings of the 36th International Conference on Machine Learning, Beach, CA, USA, 9–15 June 2019; pp. 2484–2493.
19. Poursaeed, O.; Katsman, I.; Gao, B.; Belongie, S. Generative Adversarial Perturbations. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognitio, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4422–4431.
20. Tsuzuku, Y.; Sato, I. On the structural sensitivity of deep convolutional networks to the directions of Fourier basis functions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 51–60.
21. Minagi, A.; Hirano, H.; Takemoto, K. Natural Images Allow Universal Adversarial Attacks on Medical Image Classification Using Deep Neural Networks with Transfer Learning. *J. Imaging* **2022**, *8*, 38. [CrossRef] [PubMed]
22. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]
23. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.S.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **2018**, *172*, 1122–1131.e9. [CrossRef] [PubMed]
24. Khrulkov, V.; Oseledets, I. Art of singular vectors and universal adversarial perturbations. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8562–8570.
25. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception architecture for computer vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
26. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015—Conference Track Proceedings, San Diego, CA, USA, 7–9 May 2015.
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
28. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
29. Moosavi-Dezfooli, S.-M.; Fawzi, A.; Frossard, P. DeepFool: A simple and accurate method to fool deep neural networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.
30. Xiao, C.; Zhong, P.; Zheng, C. Enhancing adversarial defense by k-winners-take-all. In Proceedings of the 8th International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
31. Mahmood, K.; Gurevin, D.; van Dijk, M.; Nguyen, P.H. Beware the black-box: On the robustness of recent defenses to adversarial examples. *Entropy* **2021**, *23*, 1359. [CrossRef] [PubMed]
32. Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F.A.; Brendel, W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
33. Tursynbek, N.; Vilkoviskiy, I.; Sindeeva, M.; Oseledets, I. Adversarial Turing patterns from cellular automata. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; pp. 2683–2691.
34. Brendel, W.; Rauber, J.; Bethge, M. Decision-based adversarial attacks: Reliable attacks against black-Box machine learning models. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
35. Chen, J.; Jordan, M.I.; Wainwright, M.J. HopSkipJumpAttack: A query-efficient decision-based attack. In Proceedings of the 2020 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 18–21 May 2020; pp. 1277–1294.
36. Jiang, H.; Kim, B.; Guan, M.; Gupta, M. To trust or not to trust a classifier. In *Proceedings of the Advances in Neural Information Processing Systems*; Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2018; Volume 31.

37. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 310. [CrossRef] [PubMed]
38. Lauritsen, S.M.; Kristensen, M.; Olsen, M.V.; Larsen, M.S.; Lauritsen, K.M.; Jørgensen, M.J.; Lange, J.; Thiesson, B. Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nat. Commun.* **2020**, *11*, 3852. [CrossRef] [PubMed]