Journal of
Clinical Medicine

MDPI

Review

Deceptive Tricks in Artificial Intelligence: Adversarial Attacks in Ophthalmology

Agnieszka M. Zbrzezny [1,2,*] and Andrzej E. Grzybowski [3,*]

[1] Faculty of Mathematics and Computer Science, University of Warmia and Mazury, 10-710 Olsztyn, Poland
[2] Faculty of Design, SWPS University of Social Sciences and Humanities, Chodakowska 19/31, 03-815 Warsaw, Poland
[3] Institute for Research in Ophthalmology, Foundation for Ophthalmology Development, 60-836 Poznan, Poland
* Correspondence: agnieszka.zbrzezny@matman.uwm.edu.pl (A.M.Z.); ae.grzybowski@gmail.com (A.E.G.)

Abstract: The artificial intelligence (AI) systems used for diagnosing ophthalmic diseases have significantly progressed in recent years. The diagnosis of difficult eye conditions, such as cataracts, diabetic retinopathy, age-related macular degeneration, glaucoma, and retinopathy of prematurity, has become significantly less complicated as a result of the development of AI algorithms, which are currently on par with ophthalmologists in terms of their level of effectiveness. However, in the context of building AI systems for medical applications such as identifying eye diseases, addressing the challenges of safety and trustworthiness is paramount, including the emerging threat of adversarial attacks. Research has increasingly focused on understanding and mitigating these attacks, with numerous articles discussing this topic in recent years. As a starting point for our discussion, we used the paper by Ma et al. "Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems". A literature review was performed for this study, which included a thorough search of open-access research papers using online sources (PubMed and Google). The research provides examples of unique attack strategies for medical images. Unfortunately, unique algorithms for attacks on the various ophthalmic image types have yet to be developed. It is a task that needs to be performed. As a result, it is necessary to build algorithms that validate the computation and explain the findings of artificial intelligence models. In this article, we focus on adversarial attacks, one of the most well-known attack methods, which provide evidence (i.e., adversarial examples) of the lack of resilience of decision models that do not include provable guarantees. Adversarial attacks have the potential to provide inaccurate findings in deep learning systems and can have catastrophic effects in the healthcare industry, such as healthcare financing fraud and wrong diagnosis.

Keywords: adversarial attacks; ophthalmology; artificial intelligence

check for updates

Citation: Zbrzezny, A.M.; Grzybowski, A.E. Deceptive Tricks in Artificial Intelligence: Adversarial Attacks in Ophthalmology. *J. Clin. Med.* **2023**, *12*, 3266. https://doi.org/10.3390/jcm12093266

Academic Editors: Vito Romano, Yalin Zheng and Mariantonia Ferrara

Received: 16 March 2023
Revised: 20 April 2023
Accepted: 26 April 2023
Published: 4 May 2023

1. Introduction

In many areas of artificial intelligence, machine learning algorithms based on big data (Wang et al. [1] and Ching-Yu et al. [2]) and deep learning have facilitated extraordinary progress, with many applications in ophthalmology (Keenan et al. [3], Papadopoulos et al. [4], Rampasek et al. [5], Ishii et al. [6], Kermany et al. [7], Liu Y. et al. [8], Liu T. et al. [9], Burlina et al. [10], Cen et al. [11], Zheng et al. [12], Shekar et al. [13], and by Zhao et al. [14]).

Advances in artificial intelligence (AI) algorithms and easy access to large public datasets have made AI models ubiquitous, from funny filters on Instagram to automatic translators that help us perform our work to specialised models that solve complicated problems such as analysing and classifying medical photos.

With the help of artificial intelligence algorithms, we can solve complex image classification problems in different areas of life. The development of big data has given us access to millions of correctly classified images. This allows the algorithms to learn better.

The latest models can achieve up to 98% accuracy, which can be misleading. Suppose we consider even minor perturbations to the image, such as the change in colour of just one pixel. Then, such models are uncertain for small perturbations. A disturbance of the image, which does not change in termsof human perception, makes it a completely different image for the AI model (Figure 1).

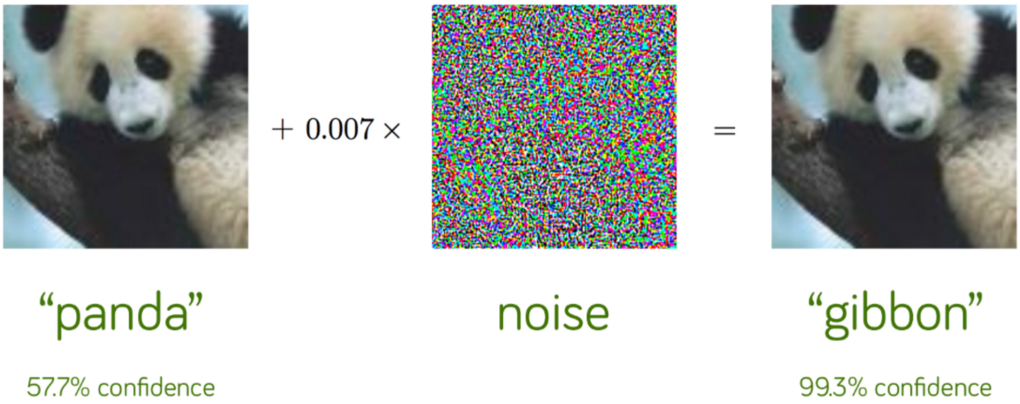


Figure 1. A classical example of an adversarial attack (Goodfellow et al. [15]).

The figure above demonstrates that the model first correctly categorised the panda image. After the perturbation was applied to this image, the model's prediction shifted to a gibbon, and the likelihood that it was a gibbon was extremely high. Humans interpret both photographs similarly. Nonetheless, the model views them as entirely distinct. This highlights the significance of adversarial attacks and that even modifications in the image that are imperceptible to humans can dramatically misrepresent the model's output, as described by Liu et al. [16].

Most of the attacks presented in this paper concern convolutional neural networks (CNNs). However, the relatively new approach in computer vision, adapted from natural language processing, vision transformers (ViTs), can achieve state-of-the-art or near-state-of-the-art performance on various image classification tasks, as demonstrated by recent advances in attention-based networks. It distinguishes transformers as a promising alternative to conventional convolutional neural networks. Mahmood et al. [17] demonstrated that white-box attacks are highly effective at producing vision transformer examples. They concluded that vision transformers are as susceptible to white-box adversaries as their CNN counterparts. Nonetheless, transferability can be utilised to achieve robustness against a black-box adversary. They developed a SAGA attack.

Hu et al. [18] also employed attacks against CNNs and ViTs. They described the IAM-UAP attack algorithm.

However, some groundbreaking studies by Naseer et al. [19] demonstrated that ViTs are more robust than CNNs against adversarial patch attacks, arguing that the dynamic, receptive field of multihead self-attention (MSA) is the reason for its superior robustness.

Wang and Ruan provided significant results [20]. The authors demonstrated that ViTs are Lipschitz continuous for vision tasks and then formally connect the local robustness of transformers to the Cauchy problem. They theoretically showed that the maximum singular value determines the local robustness of each block's Jacobian. They discovered that the initial and final layers inhibit the robustness of ViTs. In addition, contrary to existing research that suggests MSA can increase robustness, they found that the defensive power of MSA in ViT is only effective against weak adversarial attacks for large models. MSA compromises adversary robustness even against solid attacks.

Unfortunately, research using ViT to classify medical images is relatively new, and there are no presented attacks on such models for classifying medical images. It could be an interesting step in the development of specialised adversarial attacks.

In recent years, more research has been undertaken on applying artificial intelligence algorithms to solve medical problems, i.e., studies on various eye diseases. The resulting models are highly efficient. Nevertheless, we typically do not receive any information on model checking. One such difficulty may arise from the adversarial attacks presented by Goodfellow et al. [15]. They are relatively easy attacks to develop and are expected when working with images.

This study covered adversarial and specific adversarial attacks for images used in ophthalmology. In this paper, we highlight the different attacks and describe their influence on models that receive ophthalmic images as input. We also address the necessity of designing novel attack algorithms for specific sorts of pictures, such as fundus or anterior segment images.

2. Adversarial Attacks

The tendency of classifiers to overfit has led many to assume that adversarial attacks only occur in deep neural networks. All decision models have an adversarial attack as a fundamental property. Here we will explore the fundamentals of adversarial attacks to better understand attacks on ophthalmic image classification models.

Suppose d represents the number of dimensions of the input object. Consider a data point representing an image $x_0 \in \mathbb{R}^d$ of class C_i, where $0 < i \leq j$, and j is the number of classes. An adversarial attack is a malicious attempt to insert x_0 into a new data point $x \in \mathbb{R}^d$ such that the classifier misclassifies x, i.e., x belongs to a particularly hostile target class.

There are two categories of attacks: targeted and untargeted attacks. Let us consider a binary classification. In a targeted attack, a certain class $\mathcal{C}_1$ is given, and with this attack, we want a given model $\mathcal{M}$ to classify an image of a given image $\mathcal{I}$ into class $\mathcal{C}_1$, although its correct class is $\mathcal{C}_2$. An untargeted attack aims to deceive the model $\mathcal{M}$ by giving it a distorted image $\mathcal{I}$ so that it is assigned to a class other than $\mathcal{C}_1$. Untargeted attacks are not as good as targeted attacks and take much less time. Targeted attacks are more effective at changing the model's predictions, but they come at a price: time.

Problem Formulation

An adversarial attack is a malicious attempt to perturb a data point $x_0 \in \mathbb{R}^d$ to another point $x \in \mathbb{R}^d$ so that x belongs to a particular adversarial target class.

Suppose a binary classifier classifies photographs of the anterior segment of the eye for cataracts.

For example, if x_0 is a feature vector of an anterior eye segment photograph with a correct classification of `cataract`. With an adversarial attack, we want to create another feature vector x classified as `healthy` (or another class specified by the attacker). In some scenarios, the goal may not be to assign x_0 to a particular target class $\mathcal{C}_t$, but to push it away from its original class $\mathcal{C}_i$.

Let us define a data set as $\mathbb{D} = \{x_i, y_i\}_{i=1}^{N}$, where x_i is a data sample labelled y_i, and N is the size of the data set. A trained model is denoted as $\mathcal{M}$ with input x and prediction $\mathcal{M}(x)$. A loss function is denoted by ℓ. In a soft machine learning system, training attempts to minimise the loss between the target label and the predicted label:

Definition 1. *Given data $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ and target labels $\mathcal{Y} = \{y_1, y_2, \ldots, y_n\}$ find a hypothesis $\mathcal{H}$ such that*

$$\arg\min_{\mathcal{H}} \sum_{x_i, y_i \in \mathbb{D}} \ell(\mathcal{H}(x_i), y_i).$$

The trained model is tested to see how well it can predict the predicted label. One of the methods is to calculate the error by summing the losses between the target label and the predicted label.

Let us define a test dataset as $\mathbb{D}' = \{x_i, y_i\}_{i=1}^{M}$, where x_i is a data sample with label y_i, and M is the size of the dataset.

Definition 2. *Given test data $\mathcal{X}' = \{x_1, x_2, \ldots, x_m\}$ and test labels $\mathcal{Y}' = \{y_1, y_2, \ldots, y_m\}$, the error is*

$$\sum_{x_i, y_i \in \mathbb{D}'} \ell(\mathcal{H}(x_i), y_i).$$

The adversarial attack algorithm is divided into two main parts:

1. The query input is changed from the benign input x to x'.
2. An attack target is set so that the prediction result $\mathcal{H}(x)$ is no longer y. The loss is changed from $\ell(\mathcal{H}(x_i), y_i)$ to $\ell(\mathcal{H}(x_i), y'_i)$, where $y'_i \neq y_i$.

Definition 3 (Adversarial Attack (AA))**.** *Let $x_0 \in \mathbb{R}^d$ be a data point belonging to class $\mathcal{C}_i$. Define a target class $\mathcal{C}_t$. An adversarial attack is a mapping $\mathcal{A} : \mathbb{R}^d \to \mathbb{R}^d$ such that the perturbed data*

$$x = \mathcal{A}(x_0)$$

3. Methodology

As a starting point for our discussion, we used the study by Ma et al. [21].

We began our search of open-access research papers available in Google Scholar's primary search. We searched for papers that consisted of entering various terms in three types of search engines, including "ophthalmology", "ophthalmic image", "retina", and "retinal image", "fundus", and "fundus image", along with the words "adversarial attacks" and "attacks". The exact search was performed in the Google search and PubMed engines. We considered papers published after 2018.

4. Challenges in Adversarial Attacks in Medicine

Before the paper by Ma et al. [21], adversarial machine learning analysis focused on natural images, and medical image AAs were still unclear. Unlike natural images, medical images may contain domain-specific features. AAs can affect medical deep-learning systems. AAs can arbitrarily change diagnoses and outcomes. Ma et al. emphasised the importance of the fact that the healthcare system is heavily resourced. It inevitably creates risks where potential attackers can attempt to profit from manipulating the healthcare system. They gave an example of an attacker who can manipulate the health system's investigative reports to commit insurance fraud or submit a fraudulent claim for medical reimbursement, as first presented by Paschali et al. [22].

In addition, an attacker may attempt to make a misdiagnosis by manipulating an image without detection. It may significantly affect patient decisions. DNNs operate on the black box principle, so such misclassification would be nearly impossible to detect. Other techniques can explain the DNNs' decision, but consulting an expert/doctor is necessary for the diagnosis to be certain. This was shown by Avramidis et al. [23], Kind et al. [24], Chetoui et al. [25], and Ribeiro et al. [26].

Decision models and medical imaging techniques are increasingly used in medical diagnostics. Safe and robust medical deep-learning systems have become essential, as shown by by Finlayson et al. [27] and Paschali et al. [22]. The work by Ma et al. [21] was an essential step toward developing a comprehensive understanding of AAs in this field. It provides a broad knowledge of AAs in medical images, from the perspectives of both generating and detecting these attacks. However, Finlayson et al. [27] and Paschali et al. [22] have studied AAs on medical images, primarily focusing on testing the robustness of the deep models developed for medical image analysis.

The authors emphasised that AAs can succeed faster on medical images than natural ones. It means that fewer perturbations are required to execute a successful attack. They then pointed to two main reasons why deep neural networks that classify medical images are more exposed to AAs:

1. Typically, medical images have complex biological textures, leading to areas with stronger gradients that are sensitive to small perturbations from attackers.
2. State-of-the-art deep neural networks designed for large-scale natural image processing can be reparameterised for medical imaging tasks, resulting in a severe loss of landscape and high susceptibility to AAs.

According to the analysis by Ma et al. [21], medical imaging AAs generated using attack methods developed from natural images are not really "detrimental" in the medical sense. The authors noted that caution should be used when using these AAs to evaluate the performance of DNN models for medical images. Their study also shed light on the future development of more effective attacks on medical images. Future research should focus on developing specific AAs for different types of ophthalmic images.

The Danger of Adversarial Attacks in the Healthcare System

According to Finlayson et al. [28], the U.S. healthcare system favours AAs. They summarised parts of the healthcare system that can give an attacker a reason and opportunity to perform an AA.

The paper also discusses the technical reasons why medical machine-learning systems are accessible to AAs. The authors noted that the truth must often be clarified and id more contentious. End users change images that are difficult to diagnose. In this case, they can make it very hard for medical experts to figure out how much influence they have. Deep learning is most useful in edge cases.

They also noted that medical imaging is highly standardised. They found that medical AA attacks do not require invariance. Medical AA attacks are not immune to lighting or position changes.

They also considered the fact that standard network architectures are frequently used. Most of the best-published methods in medical computer vision have the same basic structure. This lack of architectural variety makes it easier for attackers to attack multiple medical systems. Models should be made public to provide transparency and enable more targeted attacks.

Finlayson et al. [28] provided a hypothetical but intriguing example of adversarial examples in ophthalmology. According to the authors, government agencies often issue diagnostic guidelines requiring coverage for specific procedures, if requirements are met. One of the criteria could be that a patient with diabetic retinopathy should undergo a specific surgery from an insurance company. Even if the insurer cannot control the policy, it can control surgery frequency using noise instead of slightly positive photos. On the other hand, an ophthalmologist may place a universal enemy image on the lens of its image capture system. A third-party image processing system would mistakenly label all images as positive cases without changing the image in the information system.

5. Adversarial Attacks in Ophthalmology

Artificial intelligence has a significant role to play in the field of ophthalmic imaging. Colour fundus imaging and optical coherence tomography play a substantial role in ophthalmology in terms of their diagnostic accuracy and usefulness in segmenting disorders.

This section provides a summary of AAs on ophthalmology-related models. However, the original issue is partitioned into two separate ones. The first is the usage of classical adversarial algorithms, and the second, given in the following section, is the application of ophthalmology-specific adversarial algorithms. In chronological order, we began our examination in 2018.

Shah et al. [29] studied the effect of AAs on retinal imaging. To detect diabetic retinopathy, they evaluated image-based CNN-0, closely influenced by Alexnet by Krizhevsky et al. [30]; CNN-1, inspired by Enet by Paszke et al. [31]; and hybrid-lesion-based medical image analysis models by Abràmoff et al. [32]. The CNN-1 and hybrid-lesion-based models were evaluated using CNN-0 and the iterative fast gradient sign method (I-FGSM) by Kurakin et al. [33] to produce adversarial images. According to the experimental results, CNN models performed rather poorly, but hybrid-lesion-based models were more robust and achieved 45% and 0.6% reductions in accuracy, respectively (Figure 2). They collected the image from the Eyepacs dataset by Cuadros and Bresnick [34]. The authors noted that the explainability of the decision process underlying medical-image-based diagnostic tasks is of utmost importance.

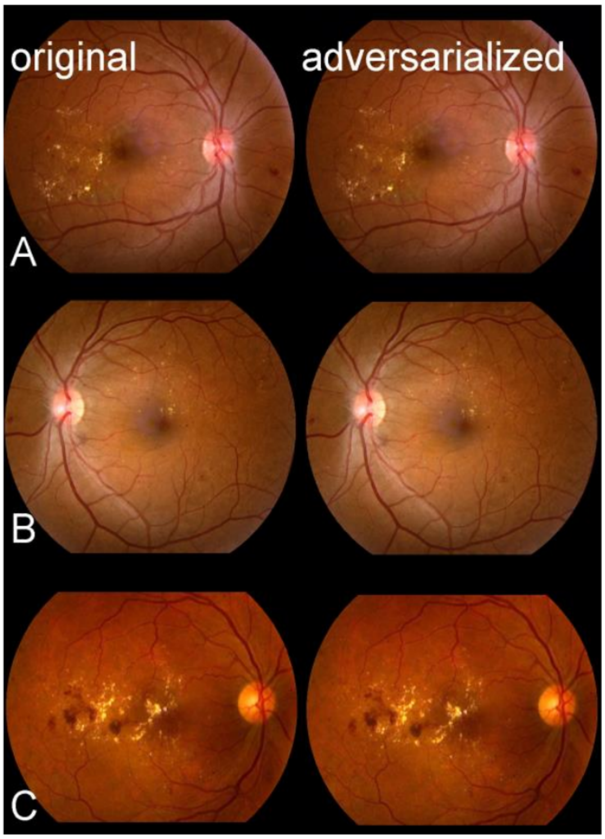


Figure 2. The AAs of three different patients (**A**–**C**) images that had previously been classified as a disease (referable diabetic retinopathy). The original images containing the disease are in the left column, while the right column displays the AA variants. Pixel differences are difficult to distinguish, even at high magnification (Shah et al. [29]).

Medical AAs were demonstrated to be feasible by Finlayson et al. [28]. The authors created a set of medical classifiers based on state-of-the-art clinical deep-learning algorithms. White- and black-box attacks against each other were the two types of AAs they carried out.

They created classification models for referable diabetic retinopathy based on fundoscopy images of the retina (similar to Gulshan et al. [35]). Data from the Kaggle Diabetic Retinopathy Dataset were used to train the models. Instead of the retinopathy grade, the Kaggle dataset sought to predict referable (grade 2 or worse) diabetic retinopathy using the model by Gulshan et al. [35].

Two different sorts of attacks, human undetected and patch attacks, were used to show how vulnerable their models were to AAs under various threat models. The white- and black-box projected gradient descent (PGD) attack methods were used in the undetectable by human attacks. The FGSM attack, first presented by Madry et al. [36], has evolved into the PGD attack, an iterative evolution of the FGSM attack. They adapted the approach outlined by Buckman et al. [37] for adversarial patch attacks.

The experimental results (Figure 3) showed that even very accurate medical classifiers can produce AAs, regardless of whether potential attackers have direct access to the model or must attack humans invisibly. PGD attacks require digital access to specific photographs provided to the model, while adversarial patch attacks can be used on any image, according to the research. They described realistic AAs against their systems.

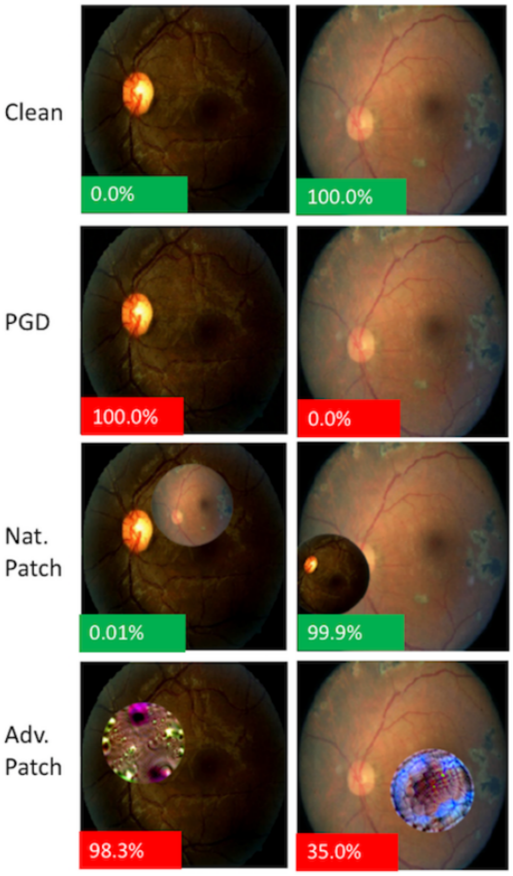


Figure 3. A demonstration that AAs can be feasible even for extremely accurate medical classifiers (Finlayson et al. [28]).

Ma et al. [21] examined four attacks on DNNs trained on five medical image datasets. They presented attack parameters, results, and analyses. FGSM, BIM, PGD, and the strongest optimisation-based Carlini and Wagner attack (L_∞ version) were performed. These attacks were limited by a maximum perturbation ϵ in the L_∞ norm on each input pixel.

They benchmarked the difficulty of classifying diabetic retinopathy fundus images. They initially categorised the image dataset into ‘no DR’ and ‘referable’. They focused on medical photos’ AA difficulty, which matched ImageNet’s nature images. Medical photos were easier to attack than ImageNet photos, the findings showed. This unexpected finding was clarified by a saliency map for two ImageNet and medical images from different classes. Based on classification loss gradients, the saliency map of an input picture shows the regions that change model output most from classification loss gradients. However, comparing the model procedures with the explanations of ImageNet and medical image output from different classes would be interesting, valuable, and intriguing.

As a result, we concluded that we require explainable models in addition to unique AAs for characteristic medical images.

Yoo and Choi [38] investigated whether AAs can confound deep-learning systems based on imaging techniques such as fundus photography (FP), ultra-widefield photography (FP), and optical coherence tomography (OCT). Based on the basic FGSM algorithm, they developed a binary classifier to identify cases of diabetic retinopathy and diabetic macular oedema. Again, the unique qualities of medical imagery were ignored in this study.

The results exhibited the same characteristics as those reported by Ma et al. [21]. Using FGSM attacks helps deep learning models perform better.

Nonetheless, the authors adopted the same stance on attacks as Ma et al. [21]. Attackers may disrupt hospitals' and insurance firms' medical billing and reimbursement systems. Defensive approaches, including adversarial training, denoising filters, and generative adversarial networks, can successfully decrease the impact of AAs.

Lal et al. [39] worked on establishing a framework for diabetic retinopathy detection-specific attack detection and defence. Keeping algorithms safe and reliable is a major concern in the fast-paced growth of artificial intelligence (AI) and deep learning (DL) techniques. The main conclusion of the study was a framework that provides a protective model against the adversarial speckle-noise attack and adversarial training. The authors then proposed a feature fusion technique that maintains categorisation with correct labelling. They assessed and studied the AAs and defences created by their framework on retinal fundus images for the problem of diabetic retinopathy identification. However, the authors concentrated on only two types of attacks: FGSM and speckle-noise attack (SN). The most significant disadvantage is that these attacks are detectable at first glance. The authors correctly noted that medical images are significantly degraded because, for example, noise is unavoidable in data acquisition, contrast is low due to illumination variations, and various other factors can produce random pixel values for individual pixels in an image by multiplying speckle noise (Figure 4).

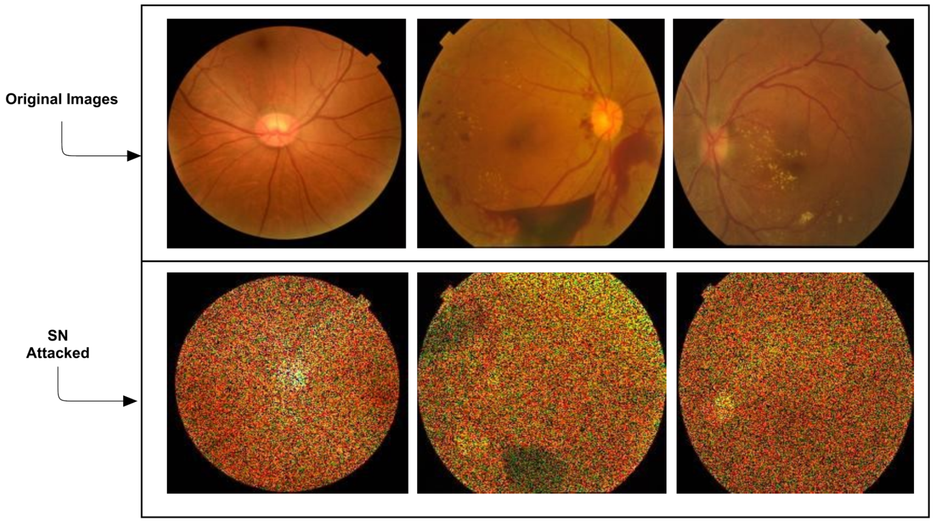


Figure 4. An example of the SN attacked images from [39] by Lal et al.

Hirano et al. [40] emphasised the need for more concentrated research on the vulnerability of DNNs to AAs. They focused on a single, modest, image-agnostic perturbation known as a universal adversarial perturbation (UAP), which can cause DNN failure in the majority of image classification tasks. Because such minute changes have little effect on data distributions, UAPs are challenging to find. Nonetheless, the authors found that UAP-based AAs are more accurate for adversaries to implement in real-world contexts. In medical imaging diagnostics, UAPs are prone to security risks.

Nonetheless, defence tactics against UAPs in DNN-based medical image classification still need to be studied, despite the sensitivity of DNNs to AAs, demonstrating the need for solutions to address security concerns.

Researchers focused on a representative medical image classification: the classification of diabetic retinopathy based on OCT images. Based on previous research, they built DNN models with different topologies for medical image recognition and investigated their susceptibility to untargeted and targeted attacks based on UAPs. They also studied the defence against attackers. They used adversarial training to test the increased resistance of DNNs to untargeted and targeted UAPs.

6. Specified Adversarial Attacks in Ophthalmology

The studies presented in the previous section have indicated a need to build algorithms for deception attacks explicitly targeted at medical images. In this section, we concentrate on attacks specifically designed for ophthalmic images in this area.

6.1. Adaptive Segmentation Mask Attack

Ozbulak et al. [41] demonstrated that deep-learning-based medical image segmentation models are susceptible to AAs. They concentrated on optic disc segmentation in glaucoma and targeted attacks. They employed the U-Net model, one of the best-recognised models for medical image segmentation. They introduced the adaptive mask segmentation attack (AMSA), a unique approach to creating AAs with realistic prediction masks. When misclassified, the algorithm modifies the masks in largely imperceptible ways to human sight. The authors posted the AA's source code on GitHub at https://github.com/utkuozbulak/adaptive-segmentation-mask-attack/ (accessed on 25 April 2023).

6.2. HFC

Yao et al. [42] presented a novel hierarchical feature constraint (HFC) to supplement current white-box attacks, allowing the adversarial representation to be concealed within the normal feature distribution. They examined the proposed approach using a fundoscopy image dataset. Their investigations revealed that the increased susceptibility of medical representations gives an adversary more opportunities for malevolent manipulation. Their premise was that medical AAs are easily detectable. They specified the HFC that can be used in attacks to evade detection. They performed comprehensive tests to determine the efficacy of the HFC. The experimental results demonstrated that HFC significantly outperformed other adaptive attacks, highlighting the inadequacies of existing approaches for identifying medical attacks in feature space.

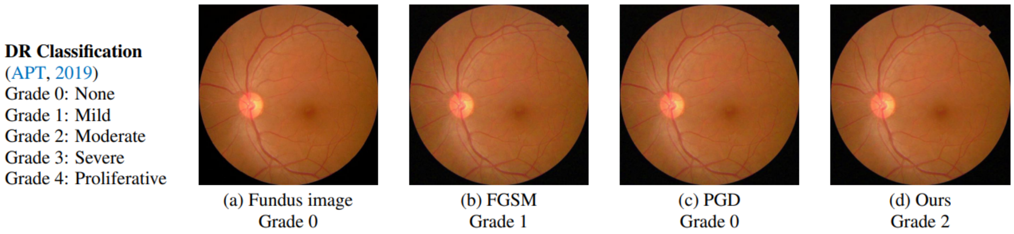
6.3. MSA

Shao et al. [43] reported a targeted attack against a biomedical image segmentation model built on multiscale gradients. Their approach integrates the attack on the adaptive segmentation mask with a perturbation of the feature space. They suggested a targeted attack approach employing multiscale gradients and applied the binary cross-entropy loss function for semantic image segmentation. The MSA approach introduces the gradient information of the loss functions to repeatedly calculate the adversarial perturbation. The original image is fitted with the adversarial perturbation and is subsequently transformed to an AA when the predicted mask of AA is close to the target mask. They conducted glaucoma-related studies utilising U-Net and an optic disc segmentation dataset. They demonstrated that their MSA method outperformed the standard method (ASMA).

6.4. SMIA

In order to diagnose pre-existing medical systems, Qi et al. [44] presented a method for medical image AA. By maximising the deviation loss term and minimising the loss stabilisation term, the approach iteratively builds AAs. The current iteration increases the difference between the CNN prediction and the matching ground truth label for AA. In comparison, the CNN predictions for the smoothed input are similar. Similar predictions ensure that the perturbations are created in a relatively flat region in the CNN feature space. The perturbations are updated from one flat region to another throughout subsequent iterations. Hence, the suggested approach cab explore perturbation space to smooth a single point and obtain a local optimum. Compared with the perturbation movement between single locations induced just by the loss deviation term, the loss stabilisation term enhances attack performance by stabilising the perturbation movement.

They used two datasets for diabetic retinopathy evaluation: APTOS-2019 and Kaggle-DR, and they used the ResNet-50 model combined with a graph convolutional network to classify fundus images. Under the same medical analytic settings, they compared the proposed SMIA technique with existing AA methods. In addition to the commonly utilised FGSM and PGD algorithms, they included DAG, which was proposed for natural image identification and segmentation, and DeepFool, which was designed for disease classification. The visualisation and experimental results proved the attack approach's success in determining the limitations of medical diagnostic systems and further enhancing them (Figure 5).



(a) Fundus image Grade 0

(b) FGSM Grade 1

(c) PGD Grade 0

(d) Ours Grade 2

Figure 5. An example of SMIA effectiveness (Qi et al. [44]).

6.5. Remarks

The aforementioned studies have illustrated specific attack tactics for medical images. However, no novel algorithms for attacks on various ophthalmic image types have yet to be developed, which is a task that must be accomplished. As a result, algorithms that explain the results of artificial intelligence models and that validate the computation of artificial intelligence models must be developed.

7. Regulations

7.1. The European Union

In April 2021, the European Commission has proposed REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS [45].

Note 51 says:

> *"Cybersecurity plays a crucial role in ensuring that AI systems are resilient against attempts to alter their use, behaviour, performance or compromise their security properties by malicious third parties exploiting the system's vulnerabilities. Cyberattacks against AI systems can leverage AI specific assets, such as training data sets (e.g., data poisoning) or trained models (e.g., adversarial attacks), or exploit vulnerabilities in the AI system's digital assets or the underlying ICT infrastructure. To ensure a level of cybersecurity appropriate to the risks, suitable measures should therefore be taken by the providers of high-risk AI systems, also taking into account as appropriate the underlying ICT infrastructure."*

Chapter 2, Article 15, point 4 says:

> *"High-risk AI systems shall be resilient as regards attempts by unauthorised third parties to alter their use or performance by exploiting the system vulnerabilities.*
>
> *The technical solutions aimed at ensuring the cybersecurity of high-risk AI systems shall be appropriate to the relevant circumstances and the risks.*
>
> *The technical solutions to address AI specific vulnerabilities shall include, where appropriate, measures to prevent and control for attacks trying to manipulate the training dataset ('data poisoning'), inputs designed to cause the model to make a mistake ('adversarial examples'), or model flaws."*

7.2. The United States of America

In January 2021, the U.S. Food and Drug Administration (FDA) presented "Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD)

Action Plan" [46]. Point 4 "4. Regulatory Science Methods Related to Algorithm Bias & Robustness" says:

> *"Support regulatory science efforts to develop methodology for the evaluation and improvement of machine learning algorithms, including for the identification and elimination of bias, and for the evaluation and promotion of algorithm robustness."*

7.3. Discussion

We could not find real-life adversarial attacks in the medicine descriptions in the literature. This shows how complicated and delicate the problem is. In the previous sections, we showed that two famous institutions are aware of the possibility of attacks and the lack of robustness of the models.

During our previous research, we used several commercial models. We proposed the algorithms' owners' joint research and the possibility of testing models using attacks several times. We also explained to companies that successful attacks will not discredit their product but show how their model can be improved. None of the companies expressed willingness to cooperate. .

Testing the models by sending perturbed images via the interface provided by the companies is possible. However, such attacking could be noticed by the fact that we would send a dozen or so, at first glance, photos for classification that did not differ in any way. This alsod involve high costs. The results obtained in this way could not be made available and published.

We are currently working on attacks against a model provided to us by another research group. However, again, this is not a commercially used model.

Soon, legal regulations requiring the verification of models and their explainability will come into force in the European Union. However, we should be aware that such a process may take a long time, and many misdiagnoses will occur before the law comes into force. Similar problems occurred with the rapid development of information systems in the 2000s. Many errors in the operation of the systems were detected after implementation, costing human lives and financial losses. Legal regulations regarding formal verification were introduced over time and in a few countries worldwide mainly because the cost of formal verification often exceeded the cost of software development.

8. Conclusions

Artificial intelligence algorithms may advance ophthalmic image processing and become a vital tool for clinicians and hospitals. However, AAs present a significant risk to patients as well as a barrier to the correct operation of AI models. The purpose of this study was to provide a summary of the studies focused on the use of AAs in ophthalmic image analysis. The analyses have demonstrated that these attacks can also mislead ophthalmic imaging models. Several researchers have presented novel attacks that have been created specifically for the field of ophthalmic image processing and have a very high level of accuracy. We provide a brief summary of the described attacks in Tables 1 and 2.

We think that computer scientists and the ophthalmology communities should closely cooperate and concentrate on tackling AAs to integrate AI algorithms into real-world problems.

Table 1. The standard AAs.

Type of the Attack	Characteristics
I-FGSM [29,33]	FGSM iteratively adds the noise (not random noise) whose direction is the same as the gradient of the cost function concerning the data.
PGD [28,36]	Initialises the attack to a random point in the ball of interest and performs random restarts.

Table 1. *Cont.*

Type of the Attack	Characteristics
FGSM [21,36,38]	Adds the noise (not random noise) whose direction is the same as the gradient of the cost function concerning the data.
BIM [21,33]	A simple extension of the FGSM where, rather than taking one significant step, it performs an iterative procedure by using FGSM numerous times on an image.
C&W [21,47]	A regularisation-based attack with some necessary modifications that can resolve the unboundedness issue.
SN [39]	The gritty salt-and-pepper pattern seen in radar imaging or a granular 'noise' that appears fundamentally in ultrasound.
UAP [40,48]	Leads to DNN failure in most image classification tasks.

Table 2. Specialised AAs.

Type of the Attack	Characteristics
AMSA [41]	Can produce AAs with realistic prediction masks.
HFC [42]	Enables hiding the adversarial representation in normal feature distribution.
SMIA [44]	The method iteratively generates AAs by maximising the deviation loss term and minimising the loss stabilisation term.
MSA [43]	Based on AMSA and multi-scale gradients.

In future work, we would like to compare all these attacks on a common database to conclude which attack is the strongest and evaluate the detection/defence methods, i.e., adversarial training, or explain the model susceptible to attacks using explainable methods.

Author Contributions: Conceptualisation, A.M.Z.; methodology, A.M.Z.; validation, A.M.Z. and A.E.G.; formal analysis, A.M.Z.; investigation, A.M.Z.; writing—original draft preparation, A.M.Z.; writing—review and editing, A.M.Z. and A.E.G.; supervision, A.E.G.; project administration, A.E.G., All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

References

1. Wang, S.Y.; Pershing, S.; Lee, A.Y. Big data requirements for artificial intelligence. *Curr. Opin. Ophthalmol.* **2020**, *31*, 318–323. [CrossRef]
2. Cheng, C.Y.; Soh, Z.D.; Majithia, S.; Thakur, S.; Rim, T.H.; Chung, T.Y.; Wong, T.Y. Big Data in Ophthalmology. *Asia-Pac. J. Ophthalmol.* **2020**, *9*, 291–298. [CrossRef] [PubMed]
3. Keenan, T.D.; Chen, Q.; Agrón, E.; Tham, Y.C.; Goh, J.H.L.; Lei, X.; Ng, Y.P.; Liu, Y.; Xu, X.; Cheng, C.Y.; et al. DeepLensNet: Deep Learning Automated Diagnosis and Quantitative Classification of Cataract Type and Severity. *Ophthalmology* **2022**, *129*, 571–584. [CrossRef] [PubMed]
4. Papadopoulos, A.; Topouzis, F.; Delopoulos, A. An Interpretable Multiple-Instance Approach for the Detection of referable Diabetic Retinopathy from Fundus Images. *Sci. Rep.* **2021**, *11*, 14326. [CrossRef]

5. Rampasek, L.; Goldenberg, A. Learning from Everyday Images Enables Expert-like Diagnosis of Retinal Diseases. *Cell* **2018**, *172*, 893–895. [CrossRef] [PubMed]
6. Ishii, K.; Asaoka, R.; Omoto, T.; Mitaki, S.; Fujino, Y.; Murata, H.; Onoda, K.; Nagai, A.; Yamaguchi, S.; Obana, A.; et al. Predicting intraocular pressure using systemic variables or fundus photography with deep learning in a health examination cohort. *Sci. Rep.* **2021**, *11*, 3687. [CrossRef]
7. Kermany, D.; Goldbaum, M.; Cai, W.; Valentim, C.; Liang, H.Y.; Baxter, S.; 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]
8. Liu, Y.; Yang, J.; Zhou, Y.; Wang, W.; Zhao, J.; Yu, W.; Zhang, D.; Ding, D.; Li, X.; Chen, Y. Prediction of OCT images of short-term response to anti-VEGF treatment for neovascular age-related macular degeneration using generative adversarial network. *Br. J. Ophthalmol.* **2020**, *104*, 1735–1740. [CrossRef]
9. Liu, T.; Farsiu, S.; Ting, D. Generative adversarial networks to predict treatment response for neovascular age-related macular degeneration: Interesting, but is it useful? *Br. J. Ophthalmol.* **2020**, *104*, 1629–1630. [CrossRef]
10. Burlina, P.; Paul, W.; Mathew, P.; Joshi, N.; Pacheco, K.; Bressler, N. Low-Shot Deep Learning of Diabetic Retinopathy With Potential Applications to Address Artificial Intelligence Bias in Retinal Diagnostics and Rare Ophthalmic Diseases. *JAMA Ophthalmol.* **2020**, *138*, 1070–1077. [CrossRef]
11. Cen, L.P.; Ji, J.; Lin, J.W.; Ju, S.; Lin, H.J.; Li, T.P.; Wang, Y.; Yang, J.F.; Liu, Y.F.; Tan, S.; et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nat. Commun.* **2021**, *12*, 4828. [CrossRef]
12. Zheng, C.; Xie, X.; Zhou, K.; Chen, B.; Chen, J.; Ye, H.; Li, W.; Qiao, T.; Gao, S.; Yang, J.; et al. Assessment of Generative Adversarial Networks Model for Synthetic Optical Coherence Tomography Images of Retinal Disorders. *Transl. Vis. Sci. Technol.* **2020**, *9*, 29. [CrossRef]
13. Shekar, S.; Satpute, N.; Gupta, A. Review on diabetic retinopathy with deep learning methods. *J. Med. Imaging* **2021**, *8*, 060901. [CrossRef]
14. Zhao, X.; Lv, B.; Meng, L.; Xia, Z.; Wang, D.; Zhang, W.; Wang, E.; Lv, C.; Xie, G.; Chen, Y. Development and quantitative assessment of deep learning-based image enhancement for optical coherence tomography. *BMC Ophthalmol.* **2022**, *22*, 139. [CrossRef] [PubMed]
15. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
16. Liu, X.; Glocker, B.; McCradden, M.M.; Ghassemi, M.; Denniston, A.K.; Oakden-Rayner, L. The medical algorithmic audit. *Lancet Digit. Health* **2022**, *4*, e384–e397. [CrossRef]
17. Mahmood, K.; Mahmood, R.; van Dijk, M. On the Robustness of Vision Transformers to Adversarial Examples. 2021. Available online: https://openaccess.thecvf.com/content/ICCV2021/papers/Mahmood_On_the_Robustness_of_Vision_Transformers_to_Adversarial_Examples_ICCV_2021_paper.pdf (accessed on 25 April 2023).
18. Hu, H.; Lu, X.; Zhang, X.; Zhang, T.; Sun, G. Inheritance Attention Matrix-Based Universal Adversarial Perturbations on Vision Transformers. *IEEE Signal Process. Lett.* **2021**, *28*, 1923–1927. [CrossRef]
19. Naseer, M.; Ranasinghe, K.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Intriguing Properties of Vision Transformers. 2021. Available online: https://proceedings.neurips.cc/paper/2021/file/c404a5adbf90e09631678b13b05d9d7a-Paper.pdf (accessed on 25 April 2023).
20. Wang, Z.; Ruan, W. Understanding Adversarial Robustness of Vision Transformers via Cauchy Problem. 2022. Available online: https://arxiv.org/abs/2208.00906 (accessed on 22 April 2023).
21. Ma, X.; Niu, Y.; Gu, L.; Wang, Y.; Zhao, Y.; Bailey, J.; Lu, F. Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems. *Pattern Recognit.* **2020**, *110*, 107332. [CrossRef]
22. Paschali, M.; Conjeti, S.; Navarro, F.; Navab, N. Generalizability vs. Robustness: Adversarial Examples for Medical Imaging. *arXiv* **2018**, arXiv:1804.00504.
23. Avramidis, K.; Rostami, M.; Chang, M.; Narayanan, S. Automating Detection of Papilledema in Pediatric Fundus Images with Explainable Machine Learning. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022. [CrossRef]
24. Kind, A.; Azzopardi, G. An Explainable AI-Based Computer Aided Detection System for Diabetic Retinopathy Using Retinal Fundus Images. In *Proceedings of the Computer Analysis of Images and Patterns*; Vento, M., Percannella, G., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 457–468.
25. Chetoui, M.; Akhloufi, M.A. Explainable Diabetic Retinopathy using EfficientNET. In Proceedings of the 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montréal, QC, Canada, 20–24 July 2020; pp. 1966–1969. [CrossRef]
26. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
27. 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]
28. Finlayson, S.G.; Kohane, I.S.; Beam, A.L. Adversarial Attacks Against Medical Deep Learning Systems. *arXiv* **2018**, arXiv:1804.05296.

29. Shah, A.; Lynch, S.; Niemeijer, M.; Amelon, R.; Clarida, W.; Folk, J.; Russell, S.; Wu, X.; Abràmoff, M.D. Susceptibility to misdiagnosis of adversarial images by deep learning based retinal image analysis algorithms. In Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, USA, 4–7 April 2018; pp. 1454–1457. [CrossRef]
30. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the Advances in Neural Information Processing Systems*; Pereira, F., Burges, C., Bottou, L., Weinberger, K., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2012; Volume 25.
31. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation. *arXiv* **2016**, arXiv:1606.02147.
32. Abràmoff, M.D.; Lou, Y.; Erginay, A.; Clarida, W.; Amelon, R.; Folk, J.C.; Niemeijer, M. Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning. *Investig. Ophthalmol. Vis. Sci.* **2016**, *57*, 5200–5206. [CrossRef]
33. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial examples in the physical world. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017.
34. Cuadros, J.; Bresnick, G. EyePACS: An Adaptable Telemedicine System for Diabetic Retinopathy Screening. *J. Diabetes Sci. Technol.* **2009**, *3*, 509–516. [CrossRef]
35. Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M.C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **2016**, *316*, 2402–2410. [CrossRef]
36. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 30 April–3 May 2018.
37. Buckman, J.; Roy, A.; Raffel, C.; Goodfellow, I. Thermometer Encoding: One Hot Way To Resist Adversarial Examples. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
38. Yoo, T.K.; Choi, J.Y. Outcomes of Adversarial Attacks on Deep Learning Models for Ophthalmology Imaging Domains. *JAMA Ophthalmol.* **2020**, *138*, 1213–1215. [CrossRef] [PubMed]
39. Lal, S.; Rehman, S.U.; Shah, J.H.; Meraj, T.; Rauf, H.T.; Damaševičius, R.; Mohammed, M.A.; Abdulkareem, K.H. Adversarial Attack and Defence through Adversarial Training and Feature Fusion for Diabetic Retinopathy Recognition. *Sensors* **2021**, *21*, 3922. [CrossRef]
40. Hirano, H.; Minagi, A.; Takemoto, K. Universal adversarial attacks on deep neural networks for medical image classification. *BMC Med. Imaging* **2020**, *21*, 9. [CrossRef]
41. Ozbulak, U.; Van Messem, A.; De Neve, W. Impact of Adversarial Examples on Deep Learning Models for Biomedical Image Segmentation. *arXiv* **2019**, arXiv:1907.13124. https://doi.org/10.48550/ARXIV.1907.13124.
42. Yao, Q.; He, Z.; Lin, Y.; Ma, K.; Zheng, Y.; Zhou, S.K. A Hierarchical Feature Constraint to Camouflage Medical Adversarial Attacks. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2021*; de Bruijne, M., Cattin, P.C., Cotin, S., Padoy, N., Speidel, S., Zheng, Y., Essert, C., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 36–47.
43. Shao, M.; Zhang, G.; Zuo, W.; Meng, D. Target attack on biomedical image segmentation model based on multi-scale gradients. *Inf. Sci.* **2021**, *554*, 33–46. [CrossRef]
44. Qi, G.; Gong, L.; Song, Y.; Ma, K.; Zheng, Y. Stabilized Medical Image Attacks. 2021. Available online: https://arxiv.org/abs/2103.05232 (accessed on 25 April 2023).
45. Commision, E. Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. 2021. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN (accessed on 22 April 2023).
46. Food and Drug Administration Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. 2021. Available online: https://www.fda.gov/media/145022/download (accessed on 25 April 2023).
47. Carlini, N.; Wagner, D.A. Towards Evaluating the Robustness of Neural Networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, 22–26 May 2017; pp. 39–57. [CrossRef]
48. Moosavi-Dezfooli, S.; Fawzi, A.; Fawzi, O.; Frossard, P. Universal adversarial perturbations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.