biomedicines

MDPI

Article

How Resilient Are Deep Learning Models in Medical Image Analysis? The Case of the Moment-Based Adversarial Attack (Mb-AdA)

Theodore V. Maliamanis, Kyriakos D. Apostolidis and George A. Papakostas *

MLV Research Group, Department of Computer Science, International Hellenic University, 65404 Kavala, Greece
* Correspondence: gpapak@cs.ihu.gr; Tel.: +30-25-1046-2321

Abstract: In the past years, deep neural networks (DNNs) have become popular in many disciplines such as computer vision (CV). One of the most important challenges in the CV area is Medical Image Analysis (MIA). However, adversarial attacks (AdAs) have proven to be an important threat to vision systems by significantly reducing the performance of the models. This paper proposes a new black-box adversarial attack, which is based on orthogonal image moments named Mb-AdA. Additionally, a corresponding defensive method of adversarial training using Mb-AdA adversarial examples is also investigated, with encouraging results. The proposed attack was applied in classification and segmentation tasks with six state-of-the-art Deep Learning (DL) models in X-ray, histopathology and nuclei cell images. The main advantage of Mb-AdA is that it does not destroy the structure of images like other attacks, as instead of adding noise it removes specific image information, which is critical for medical models' decisions. The proposed attack is more effective than compared ones and achieved degradation up to 65% and 18% in terms of accuracy and IoU for classification and segmentation tasks, respectively, by also presenting relatively high SSIM. At the same time, it was proved that Mb-AdA adversarial examples can enhance the robustness of the model.

Keywords: adversarial attack; medical image analysis; computer vision; deep learning; adversarial training; robustness; image moments

check for updates

Citation: Maliamanis, T.V.; Apostolidis, K.D.; Papakostas, G.A. How Resilient Are Deep Learning Models in Medical Image Analysis? The Case of the Moment-Based Adversarial Attack (Mb-AdA). *Biomedicines* **2022**, *10*, 2545. https://doi.org/10.3390/biomedicines10102545

Academic Editors: Yu-Te Wu and Wan-Yuo Guo

Received: 20 September 2022
Accepted: 10 October 2022
Published: 12 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Computer Vision has evolved dramatically with the introduction of Deep Learning (DL) models [1] as the accuracy and the effectiveness of deep convolutional neural networks [2] became impressive. DL has been applied with great success to nearly all CV tasks e.g., classification [2], semantic segmentation [3], object detection [4], pose estimation [5], quality assessment [6] and depth prediction [7]. The application of CV in Medical Image Analysis [8] is mostly referred to segmentation and classification problems. The application of CV to MIA compared to other CV applications, seems to have an advantage due to the fact that medical images are mostly ideally captured, lacking occlusion and miss-orientation problems and minimizing distortion, deformation and mis-illumination problems that other images generally present. Lately, the pandemic of COVID-19 showed that doctors and personnel can never be enough and that everything that can help them is valuable. This underlines the need for the implementation of computer vision-based automated medical image analysis in a safe way.

When Adversarial Attacks appeared on DL models [9], CV became seemingly unreliable. A great research field of Adversarial Computer Vision (AdCV) was born in order to create greater security in the models [10–12]. Driven by the success of adversarial attacks on natural images, researchers implement adversarial examples on medical images to investigate the MIA models' robustness [13]. The majority of studies have been done on classification and segmentation tasks. Dermoscopy images were tested in [14,15] for classification tasks. MRIs have been tested on classification tasks [16,17] and segmentation

tasks [18,19]. For X-ray images, researchers in [20,21] proved that general purpose attacks are able to importantly decrease the models' accuracy. Moreover, attacks were implemented in fundoscopy images in [22,23] presenting high attack accuracy. The aforementioned studies used several attacks such as Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), DeepFool, Jacobian-based Saliency Map Attack (JSMA), Universal Adversarial Perturbations (UAPs), Basic Iterative Method (BIM) and Carlini & Wagner (C&W), which significantly reduced the MIA models' accuracy.

Additionally, several custom medical imaging attacks were developed. In [24] created an attack for Ultrasound (US) images for classification tasks. This attack was tested on the InceptionResNetV2 model. The authors in [25] created an attack for segmentation on fundoscopy and dermoscopy images, using the U-Net model. In addition, [26] developed an attack for fundoscopy images, which can be implemented on segmentation and classification tasks. Kugler et al. [27] tried to lead five models into misclassification on dermoscopy images, (ResNet, InceptionV3, InceptionResNetV2, MobileNet, Xception). Vatian et al. [28] created an attack, which is based on the "natural" noise of medical imaging systems. Experiments were carried out on CT scans and brain MR images. Chen et al. [29] generate adversarial examples for medical image segmentation, which are tested on CT scans. Another interesting attack was created by Tian et al. [30] and it is based on a biased field phenomenon.

DL is more efficient than traditional techniques because it provides powerful models that can achieve very high accuracy in difficult medical problems just by retraining some pretrained models. That is why DL models should be resilient to adversarial attacks, as it is a critical research task that will help CV to become reliable. Trustworthiness together with the explainability of DL models, which means the ability to explain how and what features are modeled inside the DL model to extract a decision, are two qualities of the utmost importance to achieve the goal of integrating this technology into medical practice. The contribution of this work is to:

(1) Highlight the vulnerability of medical images to attacks and try to explain why some tasks are more vulnerable to attacks than others.
(2) Propose a generalized attack that significantly affects the operation of DL models.
(3) Investigate adversarial training using the proposed attack method as a universal defense method.

The proposed attack has the following features:

(1) Can be characterized as a black box attack since there is no need for any knowledge of the MIA task, the DL model structure, or the dataset used for its training.
(2) Has several degrees of freedom and can be adapted to any DL model and any image resolution.
(3) Its effect is adjustable.
(4) The way it affects DL models is fully explainable, adding clues to how we can get closer to the DL interpretability or explainability goal.

2. Materials and Methods

Image moments are image projections on a basis produced by monomials or polynomials [31]. The first set of moments introduced were "geometrical moments" that can be produced using the following equation:

$$M_{pq} = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x^p y^q f(x,y)\, \mathrm{d}x\mathrm{d}y \quad (1)$$

where $f(x,y)$ is the density distribution function of a single image channel and p, q are integer numbers. For color images, the moment set consists of three sets one for each image channel produced using the same equation.

The uniqueness theorem presented by Hu [32] proves that a piecewise continuous and fractured function $f(x,y)$, that doesn't have infinite non-zero values, can be uniquely deter-

mined by a sequence of moments $\{M_{pq}\}$ and conversely. Digital images are represented by a density distribution function $f(x, y)$, which obviously satisfies Hu theorem's conditions, except continuation. Such a function is discrete and therefore not continuous, that is the reason why we need discrete polynomials to produce the basis. In the above means the Equation (1) for discrete polynomials can be written as:

$$M_{pq} = NF(p, q) \sum_{i=1}^{N} \sum_{j=1}^{M} Kernel_{pq}(x_i, y_j) f(x_i, y_j) \quad (2)$$

where $Kernel_{pq}$ is the product of specific polynomials [33] and is the appropriate normalization factor for the polynomial family selected.

It can easily be concluded that the maximum order of the moments is equal to the maximum of the values of the image dimensions N, M, so the produced moment set has the same dimension as the image ($N \times M$).

Any discrete polynomial set that can produce a basis can be used to produce moments, but orthogonal polynomial families are mostly used. The orthogonality property simply means that the moments production procedure can be easily inverted using the same products of polynomials used for moment production.

When we collect the sequence of moments $\{M_{pq}\}$, we can use them to reconstruct the image in a simple way, due to the orthogonality property of the polynomials used in $Kernel_{pq}$ (2), using the equation:

$$\hat{f}(x, y) = \sum_{p=0}^{K-1} \sum_{q=0}^{L-1} Kernel_{pq}(x, y) M_{pq} \quad (3)$$

Noting that in (3) according to the uniqueness theorem, if the limits K and L are equal to image dimensions N and M, respectively, the estimated density distribution function $\hat{f}(x, y)$ is theoretically identical to the initial function $f(x, y)$. Moment sets can fully describe and reconstruct the image, which explains why they were broadly used as image descriptors.

The explanation of moments in physics and mathematics shows that lower order geometrical moments carry the greatest amount of information and higher order moments carry details. In image moments, lower order geometric moments represent some well-known image properties such as the total mass, the center of mass, the orientation and many others. In addition, moment functions which are invariant to scaling, translation (positioning), rotation and reflection, are called "moment invariants" [34], and they are generated using lower order moments. Lower order moment features describe the image, while on the other hand higher order moments contain information that is relatively useless for the task of image classification [35]. The way that the image information is distributed among image moments depends on the polynomial family that is used for moment production and on the family of expanding image moments. Specifically in image tasks, for the reasons mentioned earlier, discrete orthogonal image moments [31] are used. Most representative families are the Tchebichef, Krawtchouk, dual Hahn and Racah [32]. It is worth mentioning that moment set's production is a computationally "heavy" task that can have large approximation errors, especially when many orders are needed. Numerous computational strategies [33] have been proposed, to minimize approximation errors and ease computational load.

The exclusion of some moments M_{pq} of specific orders used for reconstruction by excluding some values of p or q from the sums of the reconstruction Equation (3), or practically substituting the specific moment values to be excluded with zeros, concludes in the production of an approximation of the estimated density distribution function $\hat{f}(x, y)$. Obviously, a good approximation is obtained using the optimal moments, those that carry most of the image's information.

The proposed adversarial attack that we call Moment based- Adversarial Attack (Mb-AdA) produces the adversarial examples as follows:

- The image is transformed in a moment set $\{M_{pq}\}$ using a discrete orthogonal image moments family.
- Excluding some moments of specific orders, the image is reconstructed producing an approximation.
- The product can be used as an adversarial example to attack every DL model.

The motivation behind the proposed attack is based on the fact that by eliminating some moment orders (possibly non-robust features) participating in the reconstructed image (attacked) the model will fail to detect this information and consequently will not compute the desired features within its deep structure.

Mb-AdA can be applied to any image form or resolution. DL models have a standard input resolution and that is not obvious to an image analysis system (IAS) user, because usually the system internally converts the image to the model's standard color space and resolution before feeding it to the model's input. In that case, every Mb-AdA can be used, even if it is constructed for the attacked model's input resolution or not. The effect of Mb-AdA remains after any conversion made in the image processing part of an IAS.

The above, together with the fact that the attack is based only on image features, makes the proposed attack to be applicable to any DL model, without the need for any knowledge of its structure or the dataset that was used for training. Moreover, the goal of the proposed attack is not to misclassify an image in a specific category. That is why Mb-AdA can be clearly characterized as an untargeted black box attack.

One other characteristic of Mb-AdA is that its effect is adjustable. As previously explained, the lower the moment's order is, the more significant information is carried, which is why the exclusion of the last ordered moments is preferred. Using a smaller subset of moments during image reconstruction can increase the attack's effect but with the cost of decreased image quality. On the other hand, acquiring better image quality of the adversarial examples excluding a smaller set of the last ordered moments, leads to a less efficient attack.

Adversarial Attack methods usually add an inexplainable noise to images, that is optimized in a different way in every attack method. Mb-AdA is the only that excludes specific known features of the image that makes its effect on DL models fully explainable. Moreover, studying the effect of several ordered Mb-AdAs we can explain in which moment features, is the DL model trained to base its decisions.

Obviously, for preparing only a specific Mb-AdA there is no need actually to compute entire the moment sets of the images in the first step but only the chosen subset of the moments that will be used for the reconstruction, lowering the computational cost of the procedure.

The experiments following the proposed adversarial attack uses the Tchebichef image moments family, that are constructed using (2) and the image reconstruction can be done using (3). The used $Kernel_{pq}$ in both formulas is the product of *pth* and *qth* order of the scaled Tchebichef polynomials given in (4) below. The used normalization factor $NF(p,q)$, is the inverse of the product of the squared norms $\rho(p,N)$ given in (5).

$$\widetilde{t}_n(x) = (1/b(n,N)) \sum_{i=0}^{n} \left[\binom{N-1-i}{n-i} \binom{n+i}{n} \binom{x}{i} \right] \tag{4}$$

$$\rho(p,N) = \sum_{x=0}^{N-1} \left[\widetilde{t}_p(x)\right]^2 \tag{5}$$

where n is the order of the polynomial, $N-1$ is the maximum order and equal to the corresponding dimension value of the image and $b(n,N)$ is a scaling constant, usually set equal to N^n.

For constructing *K*-th order MB-AdA for the following experiments we used the following algorithm:

- Read image dimensions N, M.

- Select attack order $K \leq \min\{N, M\}$
- Compute the polynomials $\widetilde{t}_n(x)$ using (4) and add them to a matrix.
- Compute normalization factor $NF(p, q)$ using (5) and add them to another matrix.
- Compute moment set $\{M_{pq}\}$ for every image channel using Equation (2) with the above matrices. Compute the $Kernel_{pq}(x_i, y_j)$, in every iteration of the sum from 1 to the selected order K, using the appropriate values from the polynomial matrix and saving them in a new Kernel matrix.
- Reconstruct the image using the above moment set $\{M_{pq}\}$ and the saved Kernel matrix using (3).

It should be mentioned that the reason we do not compute the entire moment set in the above algorithm is that, as explained earlier, we exclude only the last ordered moments from the reconstruction procedure. So, their computation would be not only useless but also computationally expensive.

For measuring the quality of the reconstructed (attacked) image, mean structural similarity index measure (SSIM) [36,37] is used. SSIM is one of the most representative ways to measure the difference between two images as it is correlated with the quality and perception of the human visual system [38] and it is given by:

$$\mathrm{SSIM}(x, y) = (l(x, y))^{\alpha}(c(x, y))^{\beta}(s(x, y))^{\gamma} \tag{6}$$

where α , β *and* γ are weights (usually set to 1) of the terms luminance (l), contrast (c) and structure (s), that are given below as functions of mean values μ_A *and* μ_B, variance $\sigma_{\mathbf{A}}$ *and* $\sigma_{\mathbf{B}}$ and covariance $\sigma_{\mathbf{AB}}$ respectively:

$$l(x, y) = \frac{2\mu_A\mu_B + C_1}{\mu_A^2 + \mu_B^2 + C_1} \tag{7}$$

$$c(x, y) = \frac{2\sigma_{\mathbf{A}}\sigma_{\mathbf{B}} + C_2}{\sigma_{\mathbf{A}}^2 + \sigma_{\mathbf{B}}^2 + C_2} \tag{8}$$

$$s(x, y) = \frac{\sigma_{\mathbf{AB}} + C_3}{\sigma_{\mathbf{A}}\sigma_{\mathbf{B}} + C_3} \tag{9}$$

where C_1, C_2 *and* C_3 are small constants used to stabilize the metric for the case where the means and variances become very small.

3. Results

3.1. Attack Results

In order to measure the effectiveness of the proposed Mb-AdA to MIA tasks, we applied the attack, with several orders, in classification and segmentation tasks. The classification task was tested on X-ray [39] and histopathological [40] datasets. The first consists of three categories of lung X-rays (299 × 299 pixels size) while the second, which is more difficult, consists of four categories of cancer (2048 × 1536 pixels size). All datasets were divided into training, validation and test set. For the X-ray dataset, we used 3160 images for training 360 for validation and 365 for testing while in histopathological dataset we used 280/60/60 for training, validation and testing, respectively. These datasets, were used to train five DL models namely DenseNet 201 [41], Inception V3 [42], MobileNet V2 [43], DenseNet 169 [41] and Inception ResNet [44]. The segmentation task was tested on nuclei dataset [45] using the U-Net model [46]. For this task, we used 495 images (256 × 256 pixels size) for training and from 120 for training and validation. In Figure 1 examples of each dataset of the tasks are shown. The proposed attack was compared to FGSM [47], PGD [48] and Square Attack [49]. FGSM and PGD were trained in an irrelevant dataset in order for them to behave as black box attacks. All these attacks were created with Adversarial Robustness Toolbox (ART) [50]. All experiments were performed in Python with the Keras library. Additionally, we have experimented with several moment order values but for

space saving reasons we present the most significant. As metrics we used Accuracy for classification task and Intersection over Union (IoU) for segmentation task.

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

$$IoU = \frac{Area\ of\ Intersection\ of\ two\ masks\ (predicted\ and\ ground\ truth)}{Area\ of\ Union\ of\ two\ masks\ (predicted\ and\ ground\ truth)}$$

Figure 1. Examples from datasets were used. (**a**) Histopathology (scale bar—0.42 μm), (**b**) chest X-rays, (**c**) nuclei cells.

The FGSM attack extracts the adversarial gradient and decreases or increases the value of pixels so that the loss function increases. It perturbs a clean sample for a one-step update along the direction of gradient descend. Projected Gradient Descent (PGD) attack like an Iterative FGSM. Perturbations are constrained by projecting adversarial samples from each iteration into L_∞ or L_2 neighbor of the clean image. Square Attack (SA) is based on a randomized search scheme which selects localized square shaped updates at random positions so that at each iteration the perturbation is situated approximately at the boundary of the feasible set.

In Table 1, nuclei segmentation results are presented with the SSIM (Structure Similarity Index Measure) and IoU (Intersection over Union) percentage that measures the accuracy of the segmentation for every attack. The indicating order number corresponds to the maximum order up to which is used to reconstruct the image (e.g., "Order 200" means that moments from 0 order up to 200 order, 201 × 201 moments in total are participating in the image reconstruction).

In Tables 2 and 3, the results accuracy percentage for X-rays and histopathology images are presented, respectively.

3.2. Defence Results

The impact of Adversarial Training with Mb-AdA examples, augmenting the training dataset of the previously used ML models, as a defensive technique against adversarial attacks, is a reasonable question. Training DL models with images that lack non-robust features, as Mb-AdA examples are, should theoretically make them highly resistible to adversarial attacks. Numerous experiments were performed with several combinations of orders of images but for space saving, we will show indicatively the influence of adversarial training on MobileNetV2 with the histopathological dataset, which was the most vulnerable model. Table 4 shows the MobileNetV2 results with adversarial training using the initial training set and several moment orders.

Table 1. Nuclei segmentation results under all attacks in terms of SSIM and IoU (in %).

ATTACK	SSIM	IoU
Attack Free	100	75.22
Mb-AdA Order 200	92.17	74.37
Mb-AdA Order 140	91.83	73.80
Mb-AdA Order 80	88.96	69.70
Mb-AdA Order 60	86.25	65.50
Mb-AdA Order 50	81.73	61.47
Mb-AdA Order 40	74.23	57.57
FGSM ϵ = 0.01	89.84	74.10
FGSM ϵ = 0.07	67.31	74.00
FGSM ϵ = 0.09	54.65	74.00
FGSM ϵ = 0.12	41.57	71.90
PGD ϵ = 0.01	89.73	74.12
PGD ϵ = 0.05	79.40	74.16
PGD ϵ = 0.07	70.45	74.15
PGD ϵ = 0.09	60.88	74.12
Square Attack ϵ = 0.01	90.14	74.00
Square Attack ϵ = 0.08	69.38	72.95
Square Attack ϵ = 0.1	65.00	72.28

Table 2. Image quality and accuracy in histopathological images under attacks (in %).

ATTACK	SSIM	DenseNet 201	Inception V3	MobileNet V2	DenseNet 169	Inception ResNet
Original	100	81.67	71.67	85.00	71.67	74.00
Mb-AdA Order 200	95.22	70.00	66.67	76.67	70.00	73.33
Mb-AdA Order 180	92.18	68.33	63.33	71.67	70.00	71.67
Mb-AdA Order 160	88.11	65.00	63.33	75.00	71.67	66.67
Mb-AdA Order 140	82.86	68.33	58.33	65.00	73.33	70.00
Mb-AdA Order 100	66.88	71.67	63.33	56.67	61.67	70.00
Mb-AdA Order 80	56.00	56.67	58.33	53.33	46.67	56.67
FGSM ϵ = 0.03	88.26	73.33	68.33	63.33	73.33	68.33
FGSM ϵ = 0.1	77.29	65.00	65.00	50.00	60.00	68.00
FGSM ϵ = 0.3	42.17	45.00	53.33	35.00	46.67	54.00
PGD ϵ = 0.01	89.46	75.00	66.67	76.67	75.00	76.67
PGD ϵ = 0.1	83.60	71.67	75.00	60.00	66.67	73.33
PGD ϵ = 0.2	65.32	63.33	58.33	40.00	55.00	66.67
PGD ϵ = 0.3	56.55	53.33	55.00	30.00	36.67	51.67
Square Attack ϵ = 0.01	89.70	76.67	70.00	81.67	73.33	73.33
Square Attack ϵ = 0.05	87.30	78.33	65.00	75.00	68.33	73.33
Square Attack ϵ = 0.1	80.90	66.67	56.67	60.00	65.00	75.00
Square Attack ϵ = 0.2	65.00	61.67	43.33	51.67	61.67	61.67

Table 3. Image quality and accuracy in chest X-ray images under attacks (in %).

ATTACK	SSIM	DenseNet 201	Inception V3	MobileNet V2	DenseNet 169	Inception ResNet
Original	100	99.18	93.97	97.81	97.26	97.26
Mb-AdA Order 200	99.03	98.90	93.42	91.51	93.42	96.99
Mb-AdA Order 160	98.13	98.90	92.05	87.67	94.52	96.44
Mb-AdA Order 120	95.26	96.90	89.04	80.55	89.59	94.25
Mb-AdA Order 80	91.35	82.74	81.10	56.16	69.32	86.58
Mb-AdA Order 50	85.43	60.82	59.73	43.84	36.70	61.37
Mb-AdA Order 40	82.00	53.97	40.55	42.47	33.15	60.00
Mb-AdA Order 30	77.89	40.55	34.52	40.27	32.88	54.25
FGSM ϵ = 0.01	98.71	98.36	91.23	91.78	93.42	95.34
FGSM ϵ = 0.03	94.60	96.40	84.93	80.27	88.49	92.60
FGSM ϵ = 0.05	84.10	90.14	73.15	53.15	70.41	84.11
FGSM ϵ = 0.07	76.35	84.03	66.11	42.30	68.35	76.75
FGSM ϵ = 0.09	64.23	73.11	57.14	39.78	64.71	69.75
PGD ϵ = 0.01	98.15	98.90	91.67	90.00	94.20	95.90
PGD ϵ = 0.03	93.84	96.40	83.30	75.56	82.70	87.50
PGD ϵ = 0.05	86.79	91.67	76.11	55.83	75.28	77.78

Table 3. *Cont.*

ATTACK	SSIM	DenseNet 201	Inception V3	MobileNet V2	DenseNet 169	Inception ResNet
PGD ϵ = 0.07	77.87	85.00	65.56	49.44	68.61	70.56
PGD ϵ = 0.09	68.21	76.94	59.72	46.39	63.98	60.83
Square Attack ϵ = 0.01	99.29	98.61	91.94	91.94	93.61	96.67
Square Attack ϵ = 0.05	88.00	90.28	72.22	61.39	83.89	92.50
Square Attack ϵ = 0.07	82.59	85.56	58.06	55.83	77.50	91.39
Square Attack ϵ = 0.09	76.39	79.72	46.67	58.06	70.28	85.00

Table 4. Accuracy after adversarial training in histopathological images with MobileNetV2 (in %).

ATTACK	Normal Training Set	Augmented Orders 20–200	Augmented Orders100–200
Original	85	88.33	81.6
Mb-AdA Order 200	76.67	90	73.33
Mb-AdA Order 180	71.67	90	71.67
Mb-AdA Order 160	75	86.67	75
Mb-AdA Order 140	65	85	73.33
Mb-AdA Order 120	65	88.33	76.67
Mb-AdA Order 110	65	85	80
Mb-AdA Order 100	56.67	86.67	75
Mb-AdA Order 80	53.33	81.67	76.67
Mb-AdA Order 60	45	70	63.33
Mb-AdA Order 50	45	68.33	68.33
FGSM ϵ = 0.03	63.33	90	76.67
FGSM ϵ = 0.05	60	88.33	58.33
FGSM ϵ = 0.1	50	71.67	61.67
PGD ϵ = 0.01	76.67	88.33	76.67
PGD ϵ = 0.1	60	81.67	70
PGD ϵ = 0.3	30	58.33	35
Square Attack ϵ = 0.02	80	86.67	85
Square Attack ϵ = 0.05	75	85	70
Square Attack ϵ = 0.1	60	68.33	75
Square Attack ϵ = 0.2	51.67	60	36.33

4. Discussion

The main parameter of the proposed attack is the order of moments. Mb-AdA is an attack family that does not add some specific or optimized noise to an image as usually other adversarial attacks do, but it rather removes a part of the image. The removed part is not abstract but it consists of the last ordered moment features. It is worth noting that the first moment features carry the main image information, and the last ordered ones less. Figures 2 and 3 are presented some examples of images under the proposed attack.

According to the results on X-ray dataset, Mb-AdA gets stronger as the order limit drops, because more and more of the higher ordered moment features are excluded, but as it is rational the image quality drops also. DenseNet 201, Inception V3 and Inception ResNet seem to be more resistant to higher order Mb-AdA than MobileNet V2 and DenseNet 169.

Additionally, compared to the other attacks, Mb-AdA seems to keep higher the image quality when it starts to significantly drop the accuracy rates. For example, Mb-AdA order 80 drops the DenseNet 201's accuracy score to 82.74 keeping SSIM to 91.35 when FGSM ϵ = 0.07 drops accuracy to the slightly worse 84.03 with SSIM to 76.35. PGD ϵ = 0.007 drops accuracy to the slightly worse 85 and also drops the SSIM to 77.87. The same happens with Square Attack ϵ = 0.07, which drops accuracy to 85.56 dropping the SSIM to 82.59, which is better than FGSM and PGD, nevertheless it can only be compared to Mb-AdA with order 40 and SSIM equals to 82 that drops accuracy score to 53.97. The proposed attack performs globally better classification than the other attacks in X-ray. On the other hand, the classification of histopathological images behaves differently. While SSIM drops, PGD, FGSM and Square Attack perform better than Mb-AdA. For example, Mb-AdA with order 100 and SSIM 66.88, achieved 71.67% accuracy while PGD ϵ = 0.3 with SSIM 65.32 and Square Attack ϵ = 0.2 with SSIM 65 achieved 53.33% and 61.67% accuracy, respectively. However, in high SSIM, the proposed attack outperforms the others attacks, which is our

main goal. The segmentation problem, strongly highlights the superiority of the proposed attack, as it globally shows a significantly greater reduction in terms of IoU compared to the other attacks, presenting a much higher image quality at the same time. The other attacks, achieved 3.5% IoU degradation even with very low SSIM, while our proposed attack achieved nearly 20% IoU degradation with significantly higher SSIM.

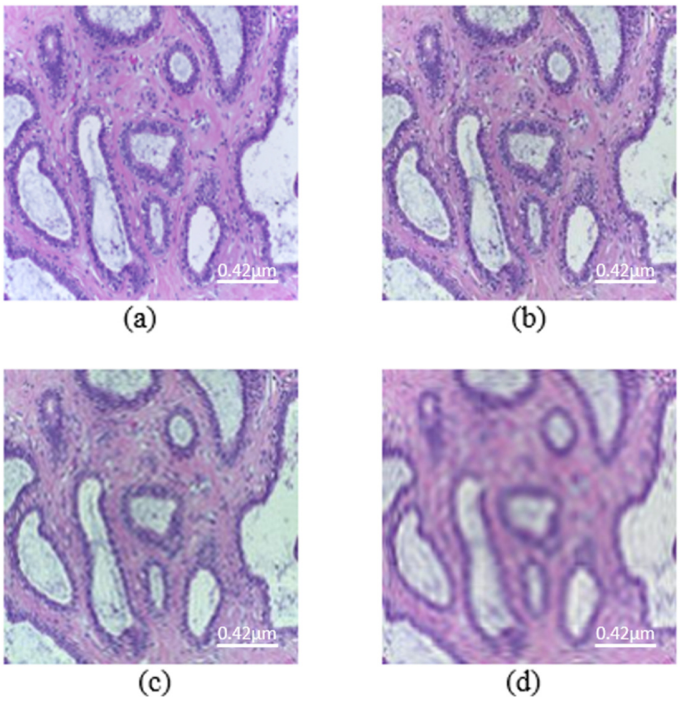

Figure 2. Images from the histopathological dataset (scale bar—0.42 μm). (**a**) Original image. Mb-Ada with order, (**b**) 200, (**c**) 120, (**d**) 50.

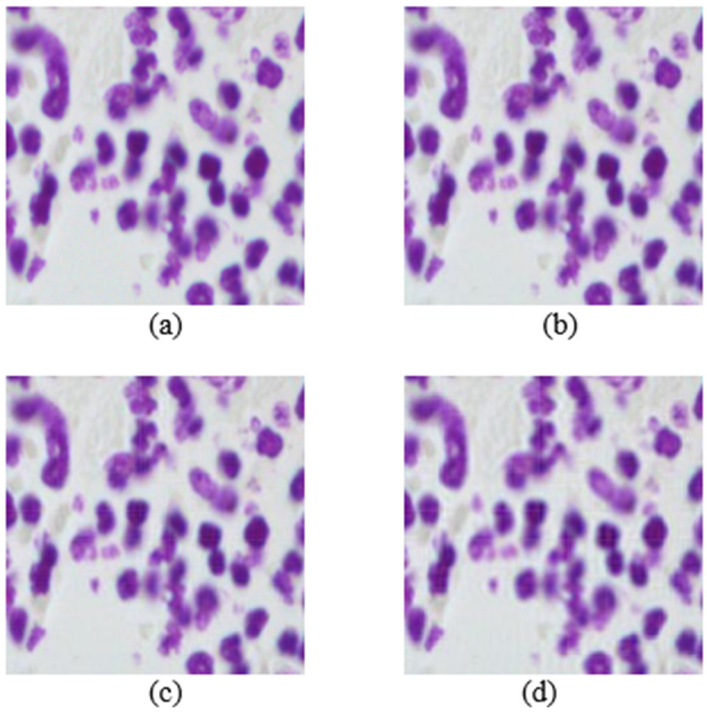

Figure 3. Images from the nuclei cell dataset. (**a**) Original image. Mb-Ada with order, (**b**) 200, (**c**) 120, (**d**) 80.

Mb-AdA is an attack that does not add some random or optimized noise to an image as usually other adversarial attacks do [13], but it rather removes a part of the image. The removed part is not abstract but it consists of the last ordered moment features. Removing from an image the last ordered non-robust features should have a small or no effect on DL models, since the image quality remains high. However, this effect is in most cases stronger than the other attacks. The reason can now be explained as the DL model during training, learning these non-robust features that are removed by the Mb-AdA, together with robust ones. The other attacks affect the non-robust and the robust features with the same weight factor, and that is why they importantly drop image quality in order to achieve the same scores with Mb-AdA.

Moreover, medical image analysis tasks are quite difficult because models learn to make decisions according to some small details in the pictures. Mb-AdA has the ability to remove features that describe these details without destroying the structure of the image, and that is why it achieves high model's performance degradation with high SSIM. The fact that the proposed attack does not destroy the structure of the image is also justified by the GradCAM algorithm [51] applied in Figure 4, which shows that models look at the almost same coordinates. This means that the attack preserves the structure of the images and does not disorient the model but removing useful information leads to misdiagnosis.

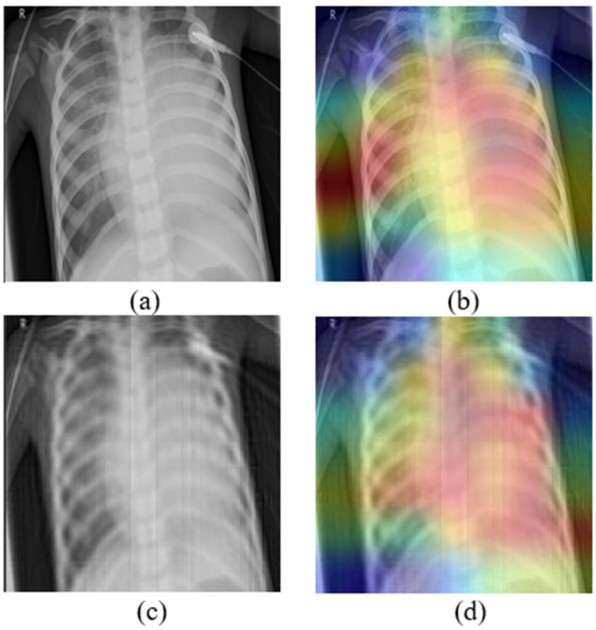

Figure 4. (**a**) Original image, (**b**) the corresponding GradCAM, (**c**) image under Mb-AdA with order 60, and (**d**) the corresponding GradCAM.

In addition to that, the data augmentation for adversarial training proves that it not only improves performance under Mb-AdA but improves performance even under other attacks. The MobileNetV2 model, which was the most vulnerable, has become quite robust after training with augmented data from the proposed attack in histopathological classification. The results are presented detailed in Table 4 but some representative examples are the following. The model trained with the Original Training Set (OTS) achieved 63.33% accuracy under FGSM $\epsilon = 0.03$, while the same model trained with Augmented Training Set (ATS) achieved 90% accuracy. Additionally, the model with OTS under PGD $\epsilon = 0.1$ achieved 60% accuracy while with ATS achieved 81.67%. These results come from the classification of histopathological images which was also the most difficult problem. It is clear that augmenting the data with examples of Mb-AdA significantly enhances the robustness of the model.

5. Conclusions

In this study, a black-box adversarial attack for medical images based on a moment exclusion strategy is proposed. The introduced attack was applied in classification and segmentation problems, achieving significant degradation in models' performance while maintaining good image quality. Compared to other black-box attacks, the Mb-AdA shows better attacking results with additional improved imperceptibility. Additionally, adversarial learning was applied as a defense, proving that it can significantly enhance the robustness of models even under other attacks. The proposed attack can be useful for developing more robust deep learning models towards enforcing the integration of Artificial Intelligence tools in critical applications such as medical image analysis.

This work is sparking the future investigation into several directions, such as the optimization of the excluded moment orders per case, the incorporation of other moment families (e.g., Krawtchouk, dual Hahn, Fractional Moments, etc.) in the same attacking scheme and the examination of other medical image modalities.

Author Contributions: Conceptualization, G.A.P.; methodology, T.V.M. and G.A.P.; software, T.V.M. and K.D.A.; validation, T.V.M. and K.D.A.; investigation, T.V.M. and K.D.A.; resources, K.D.A.; data curation, K.D.A.; writing—original draft preparation, T.V.M.; writing—review and editing, G.A.P. and T.V.M.; visualization, K.D.A.; supervision, G.A.P.; project administration, G.A.P. 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: Not applicable.

Acknowledgments: This work was supported by the MPhil program "Advanced Technologies in Informatics and Computers", hosted by the Department of Computer Science, International Hellenic University, Greece, Kavala.

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

References

1. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
2. Rawat, W.; Wang, Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Comput.* **2017**, *29*, 2352–2449. [CrossRef] [PubMed]
3. Wang, R.; Lei, T.; Cui, R.; Zhang, B.; Meng, H.; Nandi, A.K. Medical Image Segmentation Using Deep Learning: A Survey. *IET Image Process.* **2022**, *16*, 1243–1267. [CrossRef]
4. Zhao, Z.-Q.; Zheng, P.; Xu, S.; Wu, X. Object Detection with Deep Learning: A Review. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 3212–3232. [CrossRef]
5. Zheng, C.; Wu, W.; Chen, C.; Yang, T.; Zhu, S.; Shen, J.; Kehtarnavaz, N.; Shah, M. Deep Learning-Based Human Pose Estimation. A Survey. *arXiv* **2022**, arXiv:2012.13392v4.
6. Ryu, J. A Visual Saliency-Based Neural Network Architecture for No-Reference Image Quality Assessment. *Appl. Sci.* **2022**, *12*, 9567. [CrossRef]
7. Li, Q.; Zhu, J.; Liu, J.; Cao, R.; Li, Q.; Jia, S.; Qiu, G. Deep Learning based Monocular Depth Prediction: Datasets, Methods and Applications. *arXiv* **2020**, arXiv:2011.04123.
8. Rahim, T.; Usman, M.A.; Shin, S.Y. A survey on contemporary computer-aided tumor, polyp, and ulcer detection methods in wireless capsule endoscopy imaging. *Comput. Med. Imaging Graph.* **2020**, *85*, 101767. [CrossRef]
9. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2013**, arXiv:1312.6199.
10. Maliamanis, T.; Papakostas, G. Adversarial computer vision: A current snapshot. In Proceedings of the Twelfth International Conference on Machine Vision (ICMV 2019), Amsterdam, Netherlands, 16–18 November 2019; Osten, W., Nikolaev, D.P., Eds.; SPIE: Amsterdam, The Netherlands, 2020; p. 121.
11. Schneider, J.; Apruzzese, G. Concept-based Adversarial Attacks: Tricking Humans and Classifiers Alike. *arXiv* **2022**, arXiv:220310166.

12. Yao, Z.; Gholami, A.; Xu, P.; Keutzer, K.; Mahoney, M.W. Trust region based adversarial attack on neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019, Long Beach, CA, USA, 15–20 June 2019; pp. 11350–11359.
13. Apostolidis, K.D.; Papakostas, G.A. A Survey on Adversarial Deep Learning Robustness in Medical Image Analysis. *Electronics* **2021**, *10*, 2132. [CrossRef]
14. Huq, A.; Pervin, M.T. Analysis of Adversarial Attacks on Skin Cancer Recognition. In Proceedings of the 2020 International Conference on Data Science and Its Applications (ICoDSA), Bandung, Indonesia, 5–6 August 2020; IEEE: Bandung, Indonesia, 2020; pp. 1–4.
15. Paschali, M.; Conjeti, S.; Navarro, F.; Navab, N. Generalizability vs. Robustness: Investigating Medical Imaging Networks Using Adversarial Examples. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018*; Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2018; Volume 11070, pp. 493–501. ISBN 978-3-030-00927-4.
16. Li, Y.; Zhang, H.; Bermudez, C.; Chen, Y.; Landman, B.A.; Vorobeychik, Y. Anatomical context protects deep learning from adversarial perturbations in medical imaging. *Neurocomputing* **2020**, *379*, 370–378. [CrossRef] [PubMed]
17. Risk Susceptibility of Brain Tumor Classification to Adversarial Attacks | SpringerLink. Available online: https://link.springer.com/chapter/10.1007/978-3-030-31964-9_17 (accessed on 4 June 2021).
18. Cheng, G.; Ji, H. Adversarial Perturbation on MRI Modalities in Brain Tumor Segmentation. *IEEE Access* **2020**, *8*, 206009–206015. [CrossRef]
19. Anand, D.; Tank, D.; Tibrewal, H.; Sethi, A. Self-Supervision vs. Transfer Learning: Robust Biomedical Image Analysis Against Adversarial Attacks. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020; IEEE: Iowa City, IA, USA, 2020; pp. 1159–1163.
20. Hirano, H.; Minagi, A.; Takemoto, K. Universal adversarial attacks on deep neural networks for medical image classification. *BMC Med. Imaging* **2021**, *21*, 9. [CrossRef]
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.* **2021**, *110*, 107332. [CrossRef]
22. Finlayson, S.G.; Chung, H.W.; Kohane, I.S.; Beam, A.L. Adversarial Attacks Against Medical Deep Learning Systems. *arXiv* **2018**, arXiv:1804.05296.
23. Shah, A.; Lynch, S.; Niemeijer, M.; Amelon, R.; Clarida, W.; Folk, J.; Russell, S.; Wu, X.; Abramoff, 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; IEEE: Washington, DC, USA, 2018; pp. 1454–1457.
24. Byra, M.; Styczynski, G.; Szmigielski, C.; Kalinowski, P.; Michalowski, L.; Paluszkiewicz, R.; Ziarkiewicz-Wroblewska, B.; Zieniewicz, K.; Nowicki, A. Adversarial attacks on deep learning models for fatty liver disease classification by modification of ultrasound image reconstruction method. In Proceedings of the 2020 IEEE International Ultrasonics Symposium (IUS), Las Vegas, NV, USA, 7–11 September 2020.
25. 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.
26. Yao, Q.; He, Z.; Lin, Y.; Ma, K.; Zheng, Y.; Zhou, S.K. A Hierarchical Feature Constraint to Camouflage Medical Adversarial Attacks. *arXiv* **2021**, arXiv:2012.09501.
27. Kugler, D. Physical Attacks in Dermoscopy: An Evaluation of Robustness for clinical Deep-Learning. In Proceedings of the Medical Imaging with Deep Learning (MIDL), London, UK, 8–10 July 2019; pp. 1–11.
28. Vatian, A.; Gusarova, N.; Dobrenko, N.; Dudorov, S.; Nigmatullin, N.; Shalyto, A.; Lobantsev, A. Impact of Adversarial Examples on the Efficiency of Interpretation and Use of Information from High-Tech Medical Images. In Proceedings of the 2019 24th Conference of Open Innovations Association (FRUCT), Moscow, Russia, 8–12 April 2019; IEEE: Moscow, Russia, 2019; pp. 472–478.
29. Chen, L.; Bentley, P.; Mori, K.; Misawa, K.; Fujiwara, M.; Rueckert, D. Intelligent image synthesis to attack a segmentation CNN using adversarial learning. *arXiv* **2019**, arXiv:1909.11167.
30. Tian, B.; Guo, Q.; Juefei-Xu, F.; Chan, W.L.; Cheng, Y.; Li, X.; Xie, X.; Qin, S. Bias Field Poses a Threat to DNN-based X-ray Recognition. *arXiv* **2021**, arXiv:2009.09247.
31. Papakostas, G.A. Over 50 years of image moments and moment invariants. *Moments Moment Invariants-Theory Appl.* **2014**, *1*, 3–32.
32. Hu, M.-K. Visual pattern recognition by moment invariants. *IEEE Trans. Inf. Theory* **1962**, *8*, 179–187.
33. Flusser, J.; Zitova, B.; Suk, T. *Moments and Moment Invariants in Pattern Recognition*; John Wiley & Sons: Hoboken, NJ, USA, 2009.
34. Shu, H.; Luo, L.; Coatrieux, J.L. *Derivation of Moments Invariants*; Science Gate Publishing: Xanthi, Greece, 2014.
35. Papakostas, G.A. Improving the recognition performance of moment features by selection. In *Feature Selection for Data and Pattern Recognition*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 305–327.
36. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]

37. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale structural similarity for image quality assessment. In Proceedings of the The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, Pacific Grove, CA, USA, 9–12 November 2003; Volume 2, pp. 1398–1402.
38. Setiadi, D.R.I.M. PSNR vs. SSIM: Imperceptibility quality assessment for image steganography. *Multimed. Tools Appl.* **2021**, *80*, 8423–8444. [CrossRef]
39. Chowdhury, M.E.H.; Rahman, T.; Khandakar, A.; Mazhar, R.; Kadir, M.A.; Mahbub, Z.B.; Islam, K.R.; Khan, M.S.; Iqbal, A.; Emadi, N.A.; et al. Can AI Help in Screening Viral and COVID-19 Pneumonia? *IEEE Access* **2020**, *8*, 132665–132676. [CrossRef]
40. Aresta, G.; Araújo, T.; Kwok, S.; Chennamsetty, S.S.; Safwan, M.; Alex, V.; Marami, B.; Prastawa, M.; Chan, M.; Donovan, M.; et al. BACH: Grand challenge on breast cancer histology images. *Med. Image Anal.* **2019**, *56*, 122–139. [CrossRef] [PubMed]
41. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. *arXiv* **2018**, arXiv:1608.06993.
42. 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); IEEE: Las Vegas, NV, USA, 2016; pp. 2818–2826.
43. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. *arXiv* **2019**, arXiv:180104381.
44. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv* **2016**, arXiv:160207261. [CrossRef]
45. 2018 Data Science Bowl | Kaggle. Available online: https://www.kaggle.com/c/data-science-bowl-2018/data (accessed on 18 March 2022).
46. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
47. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. *arXiv* **2015**, arXiv:14126572.
48. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv* **2019**, arXiv:170606083.
49. Andriushchenko, M.; Croce, F.; Flammarion, N.; Hein, M. Square Attack: A Query-Efficient Black-Box Adversarial Attack via Random Search. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 484–501.
50. Nicolae, M.-I.; Sinn, M.; Tran, M.N.; Buesser, B.; Rawat, A.; Wistuba, M.; Zantedeschi, V.; Baracaldo, N.; Chen, B.; Ludwig, H.; et al. Adversarial Robustness Toolbox v1.0.0. *arXiv* **2019**, arXiv:180701069.
51. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.