cancers

MDPI

Article

Comparing Detection Schemes for Adversarial Images against Deep Learning Models for Cancer Imaging

Marina Z. Joel [1,2], Arman Avesta [2], Daniel X. Yang [2], Jian-Ge Zhou [3], Antonio Omuro [4], Roy S. Herbst [5], Harlan M. Krumholz [5,6] and Sanjay Aneja [2,6,*]

1 Department of Dermatology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
2 Department of Therapeutic Radiology, Yale School of Medicine, New Haven, CT 06510, USA
3 Department of Chemistry, Physics and Atmospheric Science, Jackson State University, Jackson, MS 39217, USA
4 Department of Neurology, Yale School of Medicine, New Haven, CT 06510, USA
5 Department of Medicine, Yale School of Medicine, New Haven, CT 06510, USA
6 Center for Outcomes Research and Evaluation (CORE), Yale School of Medicine, New Haven, CT 06510, USA
* Correspondence: sanjay.aneja@yale.edu; Tel.: +1-475-331-2202

Simple Summary: While deep learning has become a powerful tool in analysis of cancer imaging, deep learning models have potential vulnerabilities that pose security threats in the setting of clinical implementation. One weakness of deep learning models is that they can be deceived by adversarial images, which are manipulated images that have pixels intentionally perturbed to alter the output of the deep learning model. Recent research has shown that adversarial detection models can differentiate adversarial images from normal images to protect deep learning models from attack. We compared the effectiveness of different adversarial detection schemes, using three cancer imaging datasets (computed tomography, mammography, and magnetic resonance imaging). We found that that the detection schemes demonstrate strong performance overall but exhibit limited efficacy in detecting a subset of adversarial images. We believe our findings provide a useful basis in the application of adversarial defenses to deep learning models for medical images in oncology.

Abstract: Deep learning (DL) models have demonstrated state-of-the-art performance in the classification of diagnostic imaging in oncology. However, DL models for medical images can be compromised by adversarial images, where pixel values of input images are manipulated to deceive the DL model. To address this limitation, our study investigates the detectability of adversarial images in oncology using multiple detection schemes. Experiments were conducted on thoracic computed tomography (CT) scans, mammography, and brain magnetic resonance imaging (MRI). For each dataset we trained a convolutional neural network to classify the presence or absence of malignancy. We trained five DL and machine learning (ML)-based detection models and tested their performance in detecting adversarial images. Adversarial images generated using projected gradient descent (PGD) with a perturbation size of 0.004 were detected by the ResNet detection model with an accuracy of 100% for CT, 100% for mammogram, and 90.0% for MRI. Overall, adversarial images were detected with high accuracy in settings where adversarial perturbation was above set thresholds. Adversarial detection should be considered alongside adversarial training as a defense technique to protect DL models for cancer imaging classification from the threat of adversarial images.

Keywords: artificial intelligence; deep learning; cancer classification; medical imaging

Citation: Joel, M.Z.; Avesta, A.; Yang, D.X.; Zhou, J.-G.; Omuro, A.; Herbst, R.S.; Krumholz, H.M.; Aneja, S. Comparing Detection Schemes for Adversarial Images against Deep Learning Models for Cancer Imaging. *Cancers* **2023**, *15*, 1548. https://doi.org/10.3390/cancers15051548

Received: 4 February 2023
Revised: 27 February 2023
Accepted: 27 February 2023
Published: 1 March 2023

1. Introduction

Diagnostic imaging is a cornerstone of clinical oncology with an increasingly important role in cancer detection, treatment planning, and response assessment. With the increasing use of various diagnostic imaging modalities for cancer management, there has been a growing desire to leverage machine learning (ML) methods to improve diagnostic

image analysis [1]. Deep learning (DL) models in particular have shown significant promise in helping interpret various diagnostic imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI), and X-ray images across cancer types [2–5]. Recently, the US Food & Drug Administration (FDA) has approved multiple DL-based computer vision algorithms for medical imaging to be used by healthcare start-ups [6–8]. Recent examples are FDA-approved DL algorithms for breast cancer screening (e.g., Transpara and Mammoscreen) using digital mammography and digital breast tomosynthesis [8]. As these DL-based medical imaging systems are already approved for medical diagnosis without clinician input, DL models for medical imaging have imminent potential to be utilized in real-world cancer diagnostics. Some of the incentives for using DL models in clinical diagnostics to supplement or even replace human decision making include mitigating healthcare costs as well as human error [9].

Despite the success of DL models across various imaging tasks, there remain notable vulnerabilities which may hinder clinical its implementation. Specifically, DL models are vulnerable to adversarial images—images engineered with slight perturbations to cause DL models to give false predictions. The weakness of DL models against adversarial images stems from the fact that DL models are algorithmically unstable, producing significantly different outputs when the given inputs are subtly modified [10,11]. Clinically, this could lead to a misdiagnosis of non-cancerous lesions as cancerous, or worse, miss potential cancers present in diagnostic images. Although adversarial images are often difficult to distinguish visually from clean images, they have been shown to significantly decrease DL model classification accuracy [12–14]. Previously, it was thought that limiting access to training data—medical images for classification—to be publicly unavailable would prevent the security threat of adversarial images, as generation methods for adversarial perturbations usually require the use of original training data. However, Minagi et al. showed that transfer learning from non-medical images can be used to generate adversarial perturbations for medical images without using actual medical images as training data [15]. Thus, bad actors can potentially create adversarial images to deceive medical DL models and manipulate clinical decision making even without access to the medical images used for training, presenting opportunities for healthcare fraud and risks to patient safety [9,15]. For example, adversarial images could be used to distort patient diagnosis to generate false referrals or inappropriate treatments or medication prescriptions [16]. In light of these potential threats to the healthcare system from the manipulation of DL models, solely relying on DL algorithms to automate medical imaging tasks without human intervention can be dangerous and irresponsible despite its cost effectiveness.

Although adversarial training methods have been developed to create robust DL models which are more successful at classifying adversarial images, they have shown limited efficacy on oncologic images, and their improvement of model robustness against adversarial images comes at a tradeoff of decreasing their standard accuracy against clean images [17–19]. Furthermore, adversarial training is very computationally expensive as an iterative fine-tuning method [20]. An alternative solution to mitigate misclassification of adversarial images is developing methods which identify adversarial images before a DL model makes a prediction.

In this study, we investigate the efficacy of five different methods using DL- and ML-based detection models to classify adversarial images across three oncologic imaging modalities: CT, mammography, and MRI. Additionally, we examine the utility of combining adversarial image detection with adversarial training methods to improve DL model robustness.

2. Materials and Methods

2.1. Ethics Declaration

Research was conducted in accordance with the Declaration of Helsinki guidelines and approved by the Yale University Institutional Review Board (Protocol ID: HIC#2000027592). Informed consent was obtained from all participants in this study.

2.2. Datasets

Experiments were conducted on three datasets of different imaging modalities: CT, mammography, and MRI. We used CT imaging data composed of 1018 thoracic CT scans and 2600 lung nodules from the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) collection [21]. Thoracic radiologists identified the lung nodules used for the DL model, and associated pathologic reports were used to determine the presence of malignancy. Radiologist consensus was used to determine malignancy for patients without a pathologic determination.

We used mammography imaging data consisting of 1696 lesions from 1566 patients from the Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) [22]. Regions of interest were algorithmically derived using clinical metadata and were used to determine mammographic lesions. Verified pathologic reports were used to determine the presence of malignancy.

We used brain MRI data from 831 patients from a single institution brain metastases registry [1]. A multi-disciplinary team of radiation oncologists, neurosurgeons, and radiologists identified regions of interest. For 4000 brain lesions that we identified, we determined the presence of malignancy based on pathologic confirmation or clinical consensus.

2.3. Models

The classification models had a VGG16 convolutional neural network architecture with pretrained weights [18,23]. We used data augmentation—horizontal and vertical flips, and random rotations—to train the classification models and optimized the models using stochastic gradient descent. DL classification models were fixed post training and used for adversarial detection experiments. Each model was trained to classify the presence or absence of a malignancy in an image. Each imaging dataset was divided into a training set and a validation set using a ratio of 2:1. For image processing, each image was center cropped, resized, and normalized. Classes were balanced for each dataset.

For adversarial detection, we used five different detection models. Two were ImageNet-pretrained convolutional neural networks with ResNet50 and DenseNet-121 architecture, respectively. We also used a DenseNet-121 model to extract deep features from images and separately used logistic regression (LR), random forest (RF), and support vector machine (SVM) as the detection classifiers based on the extracted deep features. Each detection model was trained on the combination of the original training set and adversarial images generated from the training set, and tested on the combination of the original test set and adversarial images generated from the test set.

Details regarding model architecture and hyperparameter selection for model training are provided in the Supplementary Tables S1–S5. For both classification and detection models, model performance was evaluated using accuracy—the percentage of images for which the model was able to predict to correct label.

2.4. Adversarial Image Generation

We considered three first-order adversarial attack methods: Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Basic Iterative Method (BIM). Using these attack methods, we crafted adversarial images on the medical image datasets (Figure 1). All the attacks considered are bounded under a predefined perturbation size ε, which represents the maximum change to each input image pixel.

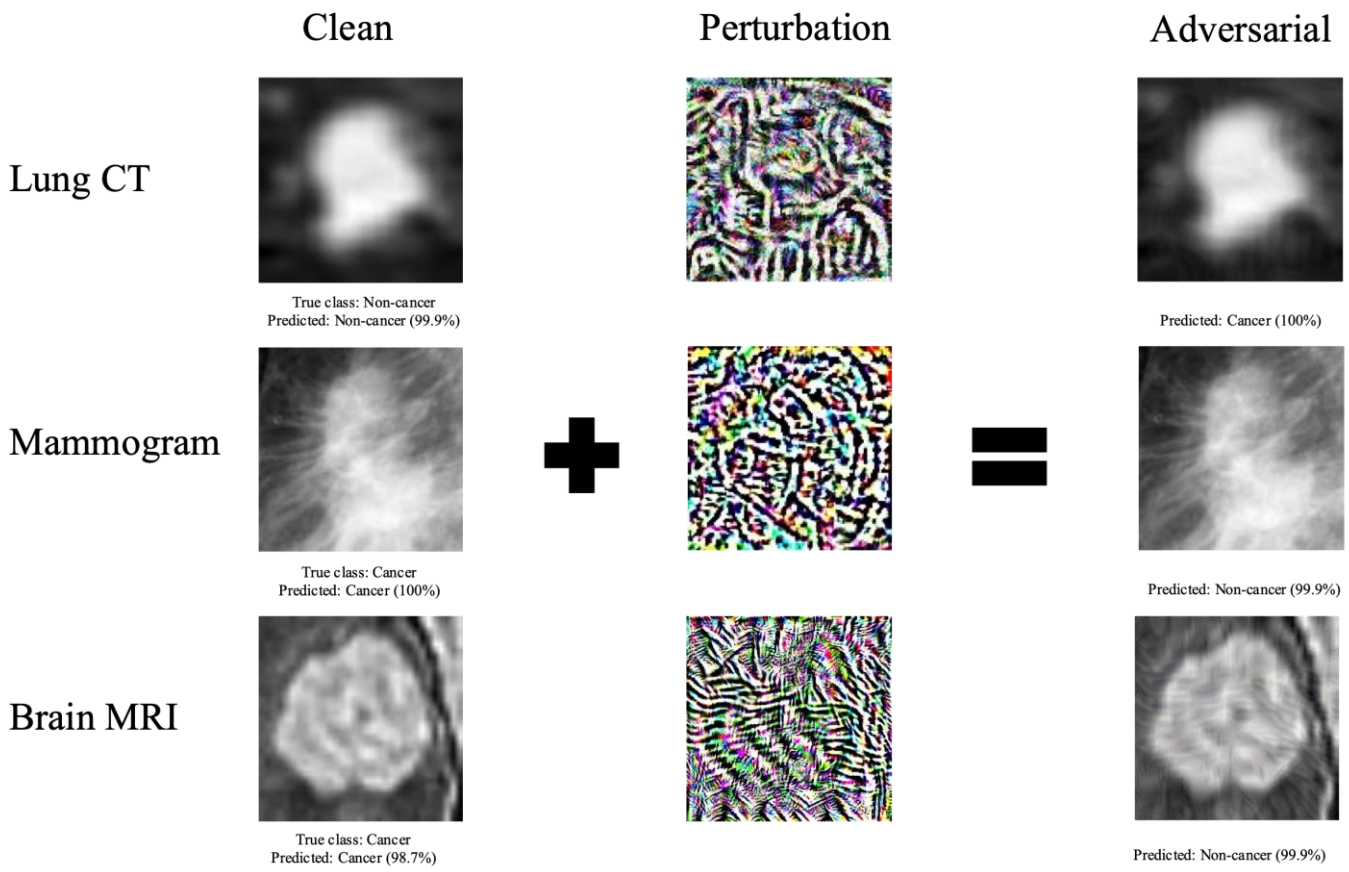


Figure 1. Examples of clean images, adversarial perturbations, and resulting adversarial images generated using PGD attack method. The percentage displayed represents the prediction confidence—the probability predicted by the DL model that the image is of a certain class. Adversarial perturbations cause a change in the DL model classification of the image.

The single-step FGSM attack perturbs the clean image by a fixed amount along the direction (sign) of the gradient of adversarial loss [24]:

$$x_{adv} = x + \varepsilon \, sign(\nabla_x \, J(x, \, y))$$

where J represents the loss function, x represents the original input image, and y represents the ground-truth label of input image.

PGD iteratively perturbs the clean image for a number of T steps with smaller step sizes; after each iteration, the updated adversarial image is projected onto the ε-ball of x [14]:

$$x^t = \prod_{\epsilon} \left(x^{t-1} + \alpha \, sign\left(\nabla_x J\left(x^t, \, y \right) \right) \right)$$

where α represents the step size, $\prod$ represents the projection function, and x^t is the adversarial image at the t-th step.

BIM is the iterative version of FGSM, essentially performing FGSM multiple times with a step size α. It also clips the pixel values of the updated adversarial image after each step into a permitted range [25].

$$x^t = Clip_{x, \, \epsilon} \left\{ x^{t-1} + \alpha \, sign\left(\nabla_x \, J\left(x^t, \, y \right) \right) \right\}$$

We evaluated the performance of our VGG16 classification models using FGSM, PGD, and BIM adversarial image generation methods across different levels of pixel perturbation.

Relative model sensitivity to adversarial images was assessed by the amount of perturbation ε required for adversarial images to substantially decrease model accuracy.

2.5. Adversarial Detection

We used the same VGG16 classification models as for above attack experiments. Each detector model was trained on the combination of the clean training set and corresponding adversarial training set generated by BIM attack. Detector model training hyperparameters are detailed in the Supplementary Table S6. For each classification task, we measure detection performance by reporting the classification accuracy of the detector model on the combination of the normal test set and the corresponding adversarial test set generated through FGSM, PGD, or BIM attack. To assess the detectability of adversarial examples, we report the detection accuracies for the detector models against all three types of attacks of varying perturbation sizes across the datasets.

2.6. Comparison of Approaches on Improving Classification Accuracy

We compared the efficacy of adversarial detection, adversarial training, and the combination of adversarial detection and adversarial training on improving classification accuracy of the DL model. Each scheme was evaluated on the combination of a clean test set and the corresponding adversarial test set generated via BIM attack with a fixed perturbation size of 0.004. We first evaluated the baseline accuracy of the original DL model on the combined test set. We then evaluated the accuracy of the adversarially trained DL model on the combined test set. For adversarial training, a multi-step PGD adversarial training method was used where for each batch of training images, half were normal images and half were adversarial images. For adversarial detection, we first used the ResNet detector to exclude images detected as adversarial and then evaluated the accuracy of the original DL model on the remaining dataset; we adjusted the accuracy by accounting for clean images wrongly excluded by the detector by including that number in the denominator of accuracy calculation. For the combined adversarial detection and adversarial training approach, we repeated the previous scheme but used the adversarially trained model instead of the original DL model for final accuracy evaluation.

The code was implemented in Python 2.7, with DL models using the TensorFlow v.1.15.3 framework and ML models using the scikit-learn 1.2.0 package [26,27]. Adversarial images were generated with the Adversarial Robustness Toolbox v.1.4.1 [28].

2.7. Code Availability

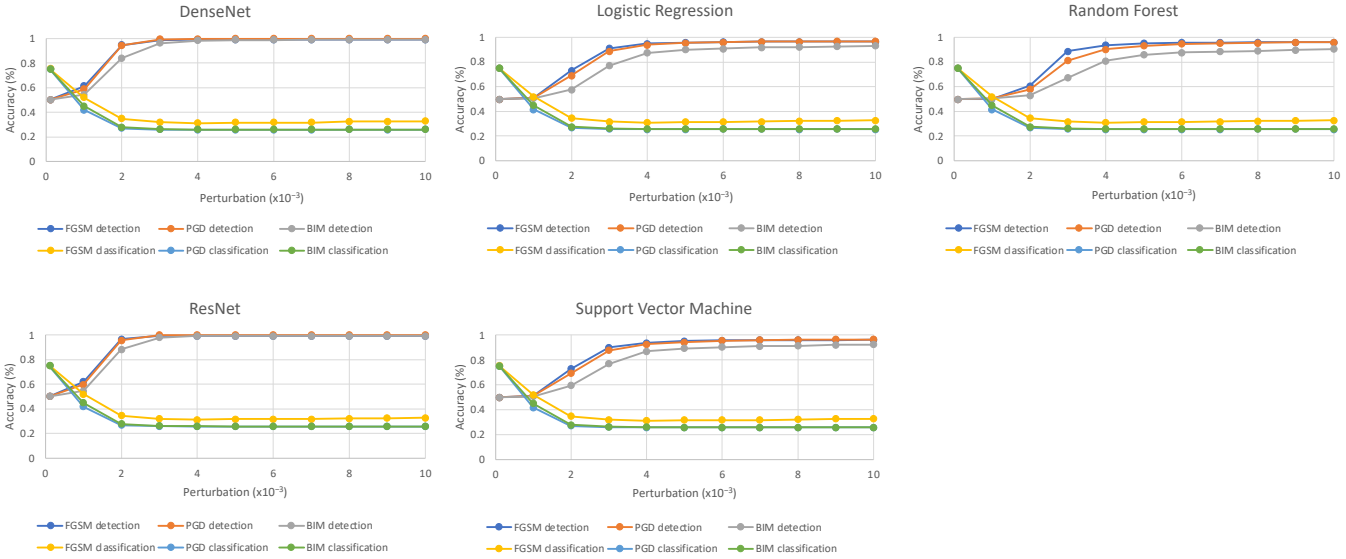
The source code for implementation of this paper is available online at Github: https://github.com/Aneja-Lab-Yale/Aneja-Lab-Public-Adversarial-Detection (accessed on 1 February 2023).

3. Results

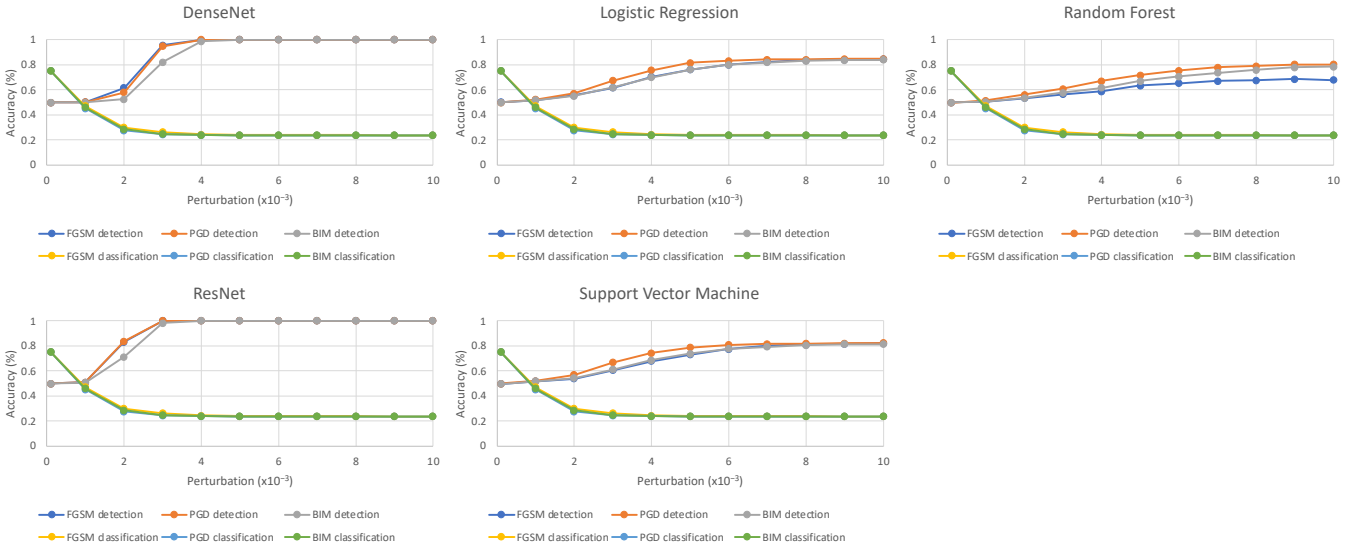
All three DL models for CT, mammogram, and brain MRI datasets were highly susceptible to adversarial attacks. Before the application of adversarial attacks, our DL models achieved baseline classification accuracies of 75.4% for CT, 76.4% for mammogram, and 92.4% for MRI. Adversarial images generated using PGD with a perturbation size of 0.004 resulted in dramatic decreases in performance: a DL model accuracy of 25.6% for CT, 23.9% for mammogram, and 7.65% for MRI.

Our adversarial detection models showed strong performance for all attacks across all datasets for attacks of perturbation sizes larger than 0.004 (Figure 2, Table 1). In all cases, the detection accuracy increases as the maximum perturbation (ε) of the attack is increased. This is expected, as adversarial images with larger perturbation sizes are more easily distinguished from normal images due to greater differences in feature distribution. Adversarial images generated using PGD with a perturbation size of 0.004 were detected by ResNet detection model with an accuracy of 100% for CT, 100% for mammogram, and 90.0% for MRI, and were detected by the DenseNet detection model with an accuracy of 99.7% for

CT, 99.9% for mammogram, and 80.5% for MRI. In contrast, the images were detected by the RF model with an accuracy of 90.6% for CT, 67.1% for mammogram, and 86.9% for MRI. Overall, our detection models showed stronger performance on the CT and mammogram datasets than on the MRI dataset. Out of the studied adversarial detection schemes, the DenseNet and ResNet models showed the best performance, while the Random Forest model showed the poorest ability to identify adversarial images.

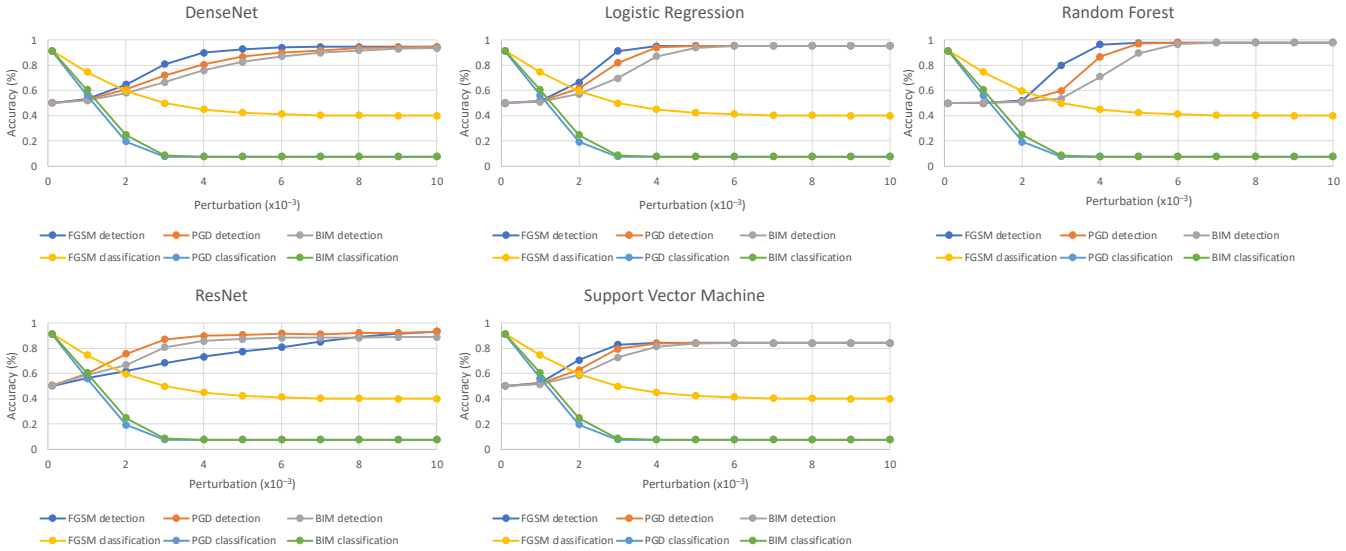


(**A**)



(**B**)

Figure 2. *Cont.*



(**C**)

Figure 2. Detection accuracy (%) of detector models (DenseNet, Logistic Regression, Random Forest, ResNet, and Support Vector Machine) on the combination of normal and adversarial test samples along with classification accuracy (%) of VGG16 classification model on adversarial samples as L_∞ maximum perturbation size ε is increased. As ε was increased, detection accuracy increased while classification model accuracy decreased for all datasets and attack types. Results are shown for (**A**) lung CT, (**B**) mammogram, and (**C**) MRI. Detection accuracy is measured as the percentage of images in the combined test set correctly classified by detection model to be normal or adversarial, while classification accuracy is measured as the percentage of images in the normal test set correctly classified by the VGG16 classification model as malignancy or no malignancy.

Table 1. Accuracy score (%) of DenseNet, Logistic Regression, Random Forest, ResNet, and Support Vector Machine detector models on a combination of normal samples and adversarial samples crafted by designated attack (FGSM, PGD, or BIM) at a set L_∞ maximum perturbation of 0.004 or 0.008 for lung CT, mammogram, and brain MRI datasets.

		Detection Accuracy (%)					
		FGSM		**PGD**		**BIM**	
		ε = 0.004	**ε = 0.008**	**ε = 0.004**	**ε = 0.008**	**ε = 0.004**	**ε = 0.008**
CT	DenseNet	99.0	99.1	99.7	99.8	98.4	99.1
	Logistic Regression	95.1	96.7	94.1	96.9	87.6	92.3
	Random Forest	93.9	96.2	90.6	95.8	81.2	89.1
	ResNet	99.3	99.3	100.0	100.0	99.2	99.3
	SVM	93.5	96.0	92.6	96.2	86.9	91.4
Mammogram	DenseNet	99.7	100.0	99.9	100.0	98.7	100.0
	Logistic Regression	70.4	83.8	75.6	84.2	69.8	83.0
	Random Forest	58.8	67.7	67.1	78.9	61.7	75.9
	ResNet	100.0	100.0	100.0	100.0	100.0	100.0
	SVM	67.9	81.0	74.4	82.0	68.7	80.6
MRI	DenseNet	90.0	94.4	80.5	93.5	75.8	91.7
	Logistic Regression	95.1	95.3	93.9	95.3	87.0	95.3
	Random Forest	96.5	97.9	86.9	97.9	70.9	97.9
	ResNet	73.3	89.1	90.0	92.2	85.8	88.4
	SVM	84.1	84.1	83.9	84.1	81.3	84.1

Overall, our detection models demonstrate strong performance against attacks with perturbation sizes above a certain threshold. Adversarial attacks with large perturbation sizes that dramatically decrease classification model performance were detected with high accuracy. On the other hand, weaker adversarial attacks with small perturbations were less likely to be detected. That being said, adversarial attacks with smaller perturbations are less likely to cause substantial changes to model classification. The perturbation threshold of detectability is heavily dependent on the perturbation size of adversarial images used to train the detection model. When the detection model is trained on adversarial images with very small perturbation sizes, the detection model is better at detecting adversarial attacks with small perturbations. However, when the perturbation sizes of adversarial attacks used to generate training images for the detection model become too small, the detection model does not train well because the differences in the features of adversarial and normal images become too miniscule for the detector to learn. Our detection schemes are strong in detecting adversarial attacks that pose powerful threats to DL classification models, as those attacks require a certain perturbation size to be effective.

When exploring the relationship between adversarial detection, adversarial training, or a combination approach on classification accuracy, we found that all three approaches significantly improved classification performance (Table 2). With adversarial images generated with BIM with a fixed perturbation size of 0.004, adversarial detection improved the classification accuracy from 50.58% to 75.63% for CT, from 50.18% to 76.43% for mammogram, and from 50.00% to 74.07% for MRI. Adversarial training improved the classification accuracy to 75.76% for CT, to 66.61% for mammogram, and to 87.88% for MRI. The combined approach improved the classification accuracy to 77.59% for CT, to 70.36% for mammogram, and to 79.99% for MRI.

Table 2. Accuracy score (%) of classification DL models after application of adversarial detection, adversarial training, or the combination of adversarial detection and adversarial training. The classification model was evaluated on a combination test set of normal and adversarial images. Adversarial images were generated with BIM with a fixed perturbation size of 0.004. For adversarial detection, the ResNet detector was used.

	Classification Accuracy (%)			
	Baseline	**Adv Detection**	**Adv Training**	**Adv Detection + Training**
LIDC	50.58	75.63	75.76	77.59
Mammogram	50.18	76.43	66.61	70.36
MRI	50.00	74.07	87.88	79.99

4. Discussion

Deep learning is a potentially powerful and inexpensive alternative or aid to human decision making for image analysis tasks [29–31]. However, as DL models are highly sensitive to adversarial attacks, protecting medical DL models against adversarial attacks is necessary for the safe and effective clinical implementation of DL models. In this study, we compared adversarial detection approaches to differentiate adversarial images from clean images. We found that adversarial attacks with perturbation sizes above a certain threshold can be detected with high accuracy using our detector models.

Previous studies that have found that adversarial images are highly dangerous to DL models for medical images, dramatically decreasing model accuracy [9,16,32,33]. We extended these findings by investigating the impact of generation methods and varying perturbation sizes of adversarial images on their efficacy at deceiving DL models for medical images [18]. We demonstrated that not all attacks are alike: PGD and BIM attacks are more effective than FGSM attacks, and adversarial images with greater perturbation sizes are more powerful than those with smaller perturbation sizes [18]. In this study, we showed that stronger adversarial images with larger perturbation sizes and a greater

impact on classification model performance can be detected with a higher accuracy than adversarial images with smaller perturbation sizes across all detection schemes.

Our study supports several works which have shown that it is feasible to develop strong approaches to detect adversarial images against DL models for medical images [7,16,33–37]. For example, Li et al. developed an effective unsupervised learning approach using a uni-modal multi-variate Gaussian model (MGM) to detect adversarial images on a deep learning model for chest X-rays [7]. Ma et al. used random forest, SVM, and logistic regression classifiers as detectors for deep features extracted by a neural network, finding high detection accuracy for each method given fixed settings for the adversarial images to be detected [32]. Our work extended this finding by comparing the performance of five detection models against adversarial images, finding some of these—ResNet and DenseNet—to be more consistently robust than others, such as the Random Forest classifier, on DenseNet-extracted features. These results show that detection model architecture is a key determinant of detection success. The detectability of adversarial medical images demonstrates underlying differences between properties of adversarial and clean medical images, as deep features of adversarial images are almost linearly separable from deep features of clean images when 2D embeddings of deep features are visualized using t-SNE [32]. In contrast with non-medical images, deep features for adversarial images closely resemble those for clean images [38,39]. Thus, medical adversarial images are easier to detect than non-medical adversarial images, even though DL models for medical images are more vulnerable to adversarial images than DL models for non-medical images [32].

To our knowledge, our work is the first to compare the effectiveness of adversarial detection, adversarial training, and the combination of adversarial detection and adversarial training to improve classification accuracy. We demonstrated that adversarial training and adversarial detection have comparable effectiveness. There are situations where one approach is superior to the other and vice versa. Furthermore, the use of adversarial training in addition to adversarial detection results in a classification performance that is intermediate to that of either approach alone. Thus, it might be helpful to use a combined approach to optimize classification performance for cases when one particular approach may be weak. This finding can be an important consideration when deciding how to best build robust image classification models for diagnostic use in clinical settings.

Unlike previous studies on adversarial imaging attacks on medical images, we found that our detection schemes underperformed when attempting to identify adversarial images with very small perturbation sizes [7,16,34]. The common limitation in many previous studies investigating adversarial detection for medical images was that they used adversarial images with a constant fixed perturbation size to evaluate the efficacy of the adversarial detector model. However, Shi et al. used an SVM classifier to detect adversarial images using chest x-ray and color fundus datasets and determined the maximum adversarial perturbations their model and human specialists cold detect, finding that detection models greatly outperformed human experts [35]. In our study, we investigated the relationship between varying adversarial perturbation sizes for adversarial images and detection performance, finding that adversarial perturbation size is positively correlated to detector model performance accuracy. Thus, while some adversarial images are more powerful and capable of wreaking havoc on the DL model, they are also more easily detectable. Adversarial images with very small perturbation sizes can fall through the cracks of standard detection schemes, but they are also less effective at decreasing DL model performance.

Our study has several limitations. First, we only tested on one classification model (VGG16), so our findings may not be applicable to other models. Additionally, some evidence suggests non-convolutional network-based models such as vision transformers maybe more robust to adversarial attacks [33]. Regardless, the VGG16 model shares behavioral similarities with other DL models which comprises a majority of clinically-employed models for image classification, so the findings from this work can be helpful to future works employing other models [40–42]. Additionally, some evidence suggests non-convolutional network-based models such as vision transformers maybe more robust

to adversarial attacks. Second, our approach only employs white-box attacks where the attack has prior knowledge of access to parameters. It would be helpful to extend the study to black-box attacks, where the attacker cannot see the model parameters, as black-box attacks may be common in real-world settings. Third, we used only first-order adversarial attacks to generate adversarial images, when higher-order attacks exist. Thus, there is a need to investigate the detectability of higher-order adversarial attacks on medical images.

While exciting progress has been made in the development of adversarial defenses, there is an arms race between the generation of novel adversarial defenses and the creation of adversarial image generation methods that circumvent these defenses [9]. We demonstrate that existing defenses against adversarial images, adversarial detection, and adversarial training cannot fully mitigate the impact of adversarial images against DL models for medical imaging classification. In the current state of DL models, the use of DL-based medical imaging algorithms should be heavily supervised by human clinicians to ensure protection against malicious interventions. Addressing the vulnerability of DL models against adversarial images should be prioritized to fully embrace widespread clinical implementation of DL systems in healthcare systems. Thus, further research into adversarial defense techniques and their effectiveness against medical adversarial images is essential.

5. Conclusions

In this work, we applied five different DL-based and ML-based adversarial detection models to compare their effectiveness at differentiating adversarial images from normal images in clinical oncology. We evaluated the performance of our detectors on three cancer imaging datasets of different diagnostic imaging modalities (CT, mammography, and MRI), finding that our detectors exhibit a high detection accuracy for adversarial images with perturbation sizes beyond a certain threshold. Our detection models can discern the adversarial images with larger perturbation sizes capable of dramatically decreasing DL classification model performance. We also demonstrated that the combination of adversarial detection and adversarial training may be a more secure method than the employment of either approach alone. However, we show that neither adversarial detection nor adversarial training can provide universal protection against adversarial images. Thus, future work should focus on detection methods capable of detecting adversarial images with a wider range of perturbation sizes. We believe that our work will facilitate the development of more robust adversarial image detection methods to defend medical deep learning models against adversarial images.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/cancers15051548/s1, Table S1: Model parameters of VGG16 model used for classification; Table S2: Training parameters for VGG16 classification models for all datasets; Table S3: Model Parameters of ResNet model used for detection; Table S4: Model Parameters of DenseNet model used for detection; Table S5: Model Training Parameters for Detection Models; Table S6: Equations and parameters for FGSM, PGD, and BIM attack methods.

Author Contributions: Conceptualization, S.A. and M.Z.J.; Methodology, S.A. and M.Z.J.; Software, S.A.; Validation, M.Z.J. and S.A.; Formal Analysis, M.Z.J., S.A. and J.-G.Z.; Investigation, M.Z.J. and S.A.; Resources, S.A.; Data Curation, S.A. and M.Z.J.; Writing—Original Draft Preparation, S.A. and M.Z.J.; Writing—Review and Editing, S.A., A.A., D.X.Y., J.-G.Z., A.O., R.S.H. and H.M.K.; Visualization, M.Z.J.; Supervision, S.A.; Project Administration, S.A.; Funding Acquisition, S.A. All authors have read and agreed to the published version of the manuscript.

Funding: S.A.: This work was supported by a Developmental Research Program Grant (PI: Aneja) from the Yale SPORE in Lung Cancer (P50CA196530), the William O. Seery Mentored Research Award for Cancer Research (PI: Aneja), and a Conquer Cancer Career Development Award (PI: Aneja), supported by Hayden Family Foundation. Any opinions, findings, and conclusions expressed in this material are those of the author(s) and do not necessarily reflect those of the American Society of Clinical Oncology®or Conquer Cancer®, or Hayden Family Foundation. J.-G.Z.: research funding by National Science Foundation NSF Award #HRD-2100971.

Institutional Review Board Statement: Research was conducted in accordance with the Declaration of Helsinki guidelines and approved by the Yale University Institutional Review Board (Protocol ID: HIC#2000027592).

Informed Consent Statement: Informed consent was obtained from all participants in this study.

Data Availability Statement: Data available upon reasonable request from authors.

Conflicts of Interest: Antonio Omuro: Consulting or Advisory Role: Merck, KIYATEC, Ono Pharmaceutical, BTG. Research Funding: Arcus Biosciences (Inst). Roy Herbst: Leadership: Jun Shi Pharmaceuticals, Immunocore. Consulting or Advisory Role: AstraZeneca, Genentech/Roche, Merck, Pfizer, AbbVie, Biodesix, Bristol Myers Squibb, Lilly, EMD Serono, Heat Biologics, Jun Shi Pharmaceuticals, Loxo, Nektar, NextCure, Novartis, Sanofi, Seattle Genetics, Shire, Spectrum Pharmaceuticals, Symphogen, TESARO, Neon Therapeutics, Infinity Pharmaceuticals, ARMO Biosciences, Genmab, Halozyme, Tocagen, Bolt Biotherapeutics, IMAB Biopharma, Mirati Therapeutics, Takeda, Cybrexa Therapeutics, eFFECTOR Therapeutics, Inc., Candel Therapeutics, Inc., Oncternal Therapeutics, STCube Pharmaceuticals, Inc., WindMIL Therapeutics, Xencor, Inc., Bayer HealthCare Pharmaceuticals Inc., Checkpoint Therapeutics, DynamiCure Biotechnology, LLC, Foundation Medicine, Inc., Gilead/Forty Seven, HiberCell, Inc., Immune-Onc Therapeutics, Inc., Johnson and Johnson, Ocean Biomedical, Inc, Oncocyte Corp, Refactor Health, Inc, Ribbon Therapeutics, Ventana Medical Systems, Inc. Research Funding: AstraZeneca, Merck, Lilly, Genentech/Roche. Harlan M. Krumholz: Employment: Hugo Health (I), FPrime, Stock and Other Ownership Interests: Element Science, Refactor Health, Hugo Health. Consulting or Advisory Role: UnitedHealthcare, Aetna. Research Funding: Johnson and Johnson. Expert Testimony: Siegfried and Jensen Law Firm, Arnold and Porter Law Firm, Martin/Baughman Law Firm. Sanjay Aneja: Consulting or Advisory Role: Prophet Consulting (I). Research Funding: The MedNet, Inc, American Cancer Society, National Science Foundation, Agency for Healthcare Research and Quality, National Cancer Institute, ASCO. Patents, Royalties, Other Intellectual Property: Provisional patent of deep learning optimization algorithm. Travel, Accommodations, Expenses: Prophet Consulting (I), Hope Foundation. Other Relationship: NRG Oncology Digital Health Working Group, SWOG Digital Engagement Committee, ASCO mCODE Technical Review Group. No other potential conflict of interest were reported.

References

1. Chang, E.; Joel, M.Z.; Chang, H.Y.; Du, J.; Khanna, O.; Omuro, A.; Chiang, V.; Aneja, S. Comparison of radiomic feature aggregation methods for patients with multiple tumors. *Sci. Rep.* **2021**, *11*, 9758. [CrossRef] [PubMed]
2. Hirano, H.; Koga, K.; Takemoto, K. Vulnerability of deep neural networks for detecting COVID-19 cases from chest X-ray images to universal adversarial attacks. *PLoS ONE* **2020**, *15*, e0243963. [CrossRef] [PubMed]
3. Zhao, W.; Jiang, W.; Qiu, X. Deep learning for COVID-19 detection based on CT images. *Sci. Rep.* **2021**, *11*, 14353. [CrossRef] [PubMed]
4. Akkus, Z.; Galimzianova, A.; Hoogi, A.; Rubin, D.L.; Erickson, B.J. Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions. *J. Digit. Imaging* **2017**, *30*, 449–459. [CrossRef]
5. Avesta, A.; Hossain, S.; Lin, M.; Aboian, M.; Krumholz, H.M.; Aneja, S. Comparing 3D, 2.5D, and 2D Approaches to Brain Image Auto-Segmentation. *Bioengineering* **2023**, *10*, 181. [CrossRef] [PubMed]
6. Benjamens, S.; Dhunnoo, P.; Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: An online database. *NPJ Digit. Med.* **2020**, *3*, 118. [CrossRef]
7. Li, X.; Zhu, D. Robust Detection of Adversarial Attacks on Medical Images. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020; pp. 1154–1158.
8. Potnis, K.C.; Ross, J.S.; Aneja, S.; Gross, C.P.; Richman, I.B. Artificial Intelligence in Breast Cancer Screening: Evaluation of FDA Device Regulation and Future Recommendations. *JAMA Intern. Med.* **2022**, *182*, 1306–1312. [CrossRef]
9. 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]
10. Shaham, U.; Yamada, Y.; Negahban, S. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing* **2018**, *307*, 195–204. [CrossRef]
11. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2013**, arXiv:1312.6199.
12. Shu, H.; Shi, R.; Zhu, H.; Chen, Z. Adversarial Image Generation and Training for Deep Neural Networks. *arXiv* **2020**, arXiv:2006.03243.
13. Tabacof, P.; Valle, E. Exploring the Space of Adversarial Images. *arXiv* **2015**, arXiv:1510.05328.
14. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv* **2017**, arXiv:1706.06083.

15. Minagi, A.; Hirano, H.; Takemoto, K. Natural Images Allow Universal Adversarial Attacks on Medical Image Classification Using Deep Neural Networks with Transfer Learning. *J. Imaging* **2022**, *8*, 38. [CrossRef] [PubMed]
16. Bortsova, G.; González-Gonzalo, C.; Wetstein, S.C.; Dubost, F.; Katramados, I.; Hogeweg, L.; Liefers, B.; van Ginneken, B.; Pluim, J.P.W.; Veta, M.; et al. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. *Med. Image Anal.* **2021**, *73*, 102141. [CrossRef] [PubMed]
17. Apostolidis, K.D.; Papakostas, G.A. A Survey on Adversarial Deep Learning Robustness in Medical Image Analysis. *Electronics* **2021**, *10*, 2132. [CrossRef]
18. Joel, M.Z.; Umrao, S.; Chang, E.; Choi, R.; Yang, D.X.; Duncan, J.S.; Omuro, A.; Herbst, R.; Krumholz, H.M.; Aneja, S. Using Adversarial Images to Assess the Robustness of Deep Learning Models Trained on Diagnostic Images in Oncology. *JCO Clin. Cancer Inform.* **2022**, *6*, e2100170. [CrossRef] [PubMed]
19. Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; Madry, A. Robustness may be at odds with accuracy. *arXiv* **2018**, arXiv:1805.12152.
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. Armato III, S.G.; McLennan, G.; Bidaut, L.; McNitt-Gray, M.F.; Meyer, C.R.; Reeves, A.P.; Zhao, B.; Aberle, D.R.; Henschke, C.I.; Hoffman, E.A. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Med. Phys.* **2011**, *38*, 915–931. [CrossRef]
22. Lee, R.S.; Gimenez, F.; Hoogi, A.; Miyake, K.K.; Gorovoy, M.; Rubin, D.L. A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci. Data* **2017**, *4*, 170177. [CrossRef] [PubMed]
23. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
24. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. *arXiv* **2014**, arXiv:1412.6572.
25. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial examples in the physical world. *arXiv* **2016**, arXiv:1607.02533.
26. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv* **2016**, arXiv:1603.04467.
27. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
28. Nicolae, M.-I.; Sinn, M.; Tran, M.N.; Buesser, B.; Rawat, A.; Wistuba, M.; Zantedeschi, V.; Baracaldo, N.; Chen, B.; Ludwig, H. Adversarial Robustness Toolbox v1. 0.0. *arXiv* **2018**, arXiv:1807.01069.
29. Kyono, T.; Gilbert, F.J.; van der Schaar, M. MAMMO: A Deep Learning Solution for Facilitating Radiologist-Machine Collaboration in Breast Cancer Diagnosis. *arXiv* **2018**, arXiv:1811.02661.
30. Park, A.; Chute, C.; Rajpurkar, P.; Lou, J.; Ball, R.L.; Shpanskaya, K.; Jabarkheel, R.; Kim, L.H.; McKenna, E.; Tseng, J.; et al. Deep Learning–Assisted Diagnosis of Cerebral Aneurysms Using the HeadXNet Model. *JAMA Netw. Open* **2019**, *2*, e195600. [CrossRef]
31. Sahiner, B.; Pezeshk, A.; Hadjiiski, L.M.; Wang, X.; Drukker, K.; Cha, K.H.; Summers, R.M.; Giger, M.L. Deep learning in medical imaging and radiation therapy. *Med. Phys.* **2019**, *46*, e1–e36. [CrossRef]
32. 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]
33. Ghaffari Laleh, N.; Truhn, D.; Veldhuizen, G.P.; Han, T.; van Treeck, M.; Buelow, R.D.; Langer, R.; Dislich, B.; Boor, P.; Schulz, V.; et al. Adversarial attacks and adversarial robustness in computational pathology. *Nat. Commun.* **2022**, *13*, 5711. [CrossRef] [PubMed]
34. Li, X.; Pan, D.; Zhu, D. Defending against adversarial attacks on medical imaging AI system, classification or detection? *arXiv* **2020**, arXiv:2006.13555.
35. Yang, Y.; Shih, F.Y.; Chang, I.C. Adaptive Image Reconstruction for Defense Against Adversarial Attacks. *Int. J. Pattern Recognit. Artif. Intell.* **2022**, *36*, 2252022. [CrossRef]
36. Yang, Y.; Shih, F.Y.; Roshan, U. Defense Against Adversarial Attacks Based on Stochastic Descent Sign Activation Networks on Medical Images. *Int. J. Pattern Recognit. Artif. Intell.* **2022**, *36*, 2254005. [CrossRef]
37. Shi, X.; Peng, Y.; Chen, Q.; Keenan, T.; Thavikulwat, A.T.; Lee, S.; Tang, Y.; Chew, E.Y.; Summers, R.M.; Lu, Z. Robust convolutional neural networks against adversarial attacks on medical images. *Pattern Recognit.* **2022**, *132*, 108923. [CrossRef]
38. Feinman, R.; Curtin, R.R.; Shintre, S.; Gardner, A.B. Detecting adversarial samples from artifacts. *arXiv* **2017**, arXiv:1703.00410.
39. Ma, X.; Li, B.; Wang, Y.; Erfani, S.M.; Wijewickrema, S.; Schoenebeck, G.; Song, D.; Houle, M.E.; Bailey, J. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv* **2018**, arXiv:1801.02613.
40. Thompson, R.F.; Valdes, G.; Fuller, C.D.; Carpenter, C.M.; Morin, O.; Aneja, S.; Lindsay, W.D.; Aerts, H.; Agrimson, B.; Deville, C., Jr.; et al. Artificial Intelligence in Radiation Oncology Imaging. *Int. J. Radiat. Oncol. Biol. Phys.* **2018**, *102*, 1159–1161. [CrossRef]

41. Aneja, S.; Chang, E.; Omuro, A. Applications of artificial intelligence in neuro-oncology. *Curr. Opin. Neurol.* **2019**, *32*, 850–856. [CrossRef]
42. Thompson, R.F.; Valdes, G.; Fuller, C.D.; Carpenter, C.M.; Morin, O.; Aneja, S.; Lindsay, W.D.; Aerts, H.J.W.L.; Agrimson, B.; Deville, C.; et al. The Future of Artificial Intelligence in Radiation Oncology. *Int. J. Radiat. Oncol. Biol. Phys.* **2018**, *102*, 247–248. [CrossRef] [PubMed]