Adversarial Attack Vulnerability of Deep Learning Models for Oncologic Images

Marina Z. Joel[1], Sachin Umrao[1], Enoch Chang[2], Rachel Choi[2], Daniel Yang[1], Antonio Omuro[3], Roy Herbst[4], Harlan Krumholz[4,5], Sanjay Aneja[1,5]

1. Department of Therapeutic Radiology, Yale School of Medicine, New Haven, CT
2. Yale School of Medicine, New Haven, CT
3. Yale Brain Tumor Program, New Haven, CT
4. Department of Internal Medicine, Yale School of Medicine, New Haven, CT
5. Center for Outcomes Research and Evaluation, Yale School of Medicine, New Haven, CT

Corresponding Author:

Sanjay Aneja, MD
Assistant Professor
Department of Therapeutic Radiology
Center for Outcomes Research and Evaluation (CORE)
Yale School of Medicine
Email: sanjay.aneja@yale.edu

Abstract

Background: Deep learning (DL) models have shown the ability to automate the classification of medical images used for cancer detection. Unfortunately, recent studies have found that DL models are vulnerable to adversarial attacks which manipulate models into making incorrect predictions with high confidence. There is a need for better understanding of how adversarial attacks impact the predictive ability of DL models in the medical image domain.

Methods: We studied the adversarial attack susceptibility of DL models for three common imaging tasks within oncology. We investigated how PGD adversarial training could be employed to increase model robustness against FGSM, PGD, and BIM attacks. Finally, we studied the utility of adversarial sensitivity as a metric to improve model performance.

Results: Our experiments showed that medical DL models were highly sensitive to adversarial attacks, as visually imperceptible degrees of perturbation (<0.004) were sufficient to deceive the model the majority of the time. DL models for medical images were more vulnerable to adversarial attacks compared to DL models for non-medical images. Adversarial training increased model performance on adversarial samples for all classification tasks. We were able to increase model accuracy on clean images for all datasets by excluding images most vulnerable to adversarial perturbation.

Conclusion: Our results indicated that while medical DL systems are extremely susceptible to adversarial attacks, adversarial training show promise as an effective defense against attacks. Adversarial susceptibility of individual images can be used to increase model performance by identifying images most at-risk for misclassification. Our findings provide a useful basis for designing more robust and accurate medical DL models as well as techniques to defend models from adversarial attack.

Keywords: Adversarial Attacks, Medical Imaging, Deep Learning, Classification

Introduction:

Deep learning (DL) models have demonstrated increasing utility within the field of oncology and particular promise in analyzing growing amounts of oncologic imaging [1, 2]. DL models have been validated across a variety of diagnostic imaging modalities including MRI, CT, and X-ray images with classification accuracy often rivaling trained clinicians [3-8]. As widespread clinical implementation of DL models becomes a more realistic possibility, the safety of such models in healthcare is becoming a topic of increasing importance [9-11].

One concerning limitation of DL models that may hinder safe implementation is their susceptibility to *adversarial attacks.* Adversarial attacks on DL models occur when data is manipulated with small perturbations specifically designed to cause a decrease in DL model performance [12-15]. The vulnerability of DL models to adversarial attacks stems from the fact that DL models are often unstable locally, generating significantly different outputs on inputs which vary only slightly [16, 17]. Although adversarial images created from adversarial attacks are often imperceptible from unmodified images to the human eye, they have been shown to significantly reduce the accuracy of DL models [18-20].

Previous work concerning adversarial attacks on DL models has largely focused on non-medical images, and the vulnerability of oncologic images to such attacks is relatively unknown [20, 21]. Although techniques to defend against adversarial attacks have been proposed for non-medical images, their generalizability on medical images is unclear. Recent rises in cyberattacks on hospital systems shed light on the weaknesses of healthcare systems and suggest that adversarial attacks on clinically implemented DL models have genuine potential for harm [22-26]. There is a pressing need to understand both the vulnerability of medical DL models to adversarial attacks and the relative effectiveness of current adversarial defense techniques in the medical setting.

In this study we explored adversarial attacks on DL models trained on common imaging modalities in clinical oncology. Specifically, we investigated the relative susceptibility of different oncologic imaging modalities to adversarial attacks and the effectiveness of adversarial defense techniques in mitigating DL model performance when facing adversarial attacks. Lastly, we explored the utility of adversarial attack susceptibility as a metric to improve the overall performance of DL models used on oncologic images.

Methodology

Datasets

We identified three medical imaging modalities commonly used in oncology—computed tomography (CT), mammography, and magnetic resonance imaging (MRI) —as the basis of our analysis.

CT images were acquired from the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) collection [27], which contains thoracic CT scans from 1018 cases collected from 15 clinical sites across the US. The dataset included XML annotations from 4 experienced radiologists outlining the nodules, which were rated from 1-5 (1 being least likely for malignancy and 5 being most). Some nodules had associated pathological reports which classified the nodules as cancer or non-cancer, and for those without pathology information, nodules with a rating of 4 or 5 were assumed as cancer. We used 1702 lung CT images from the LIDC-IDRI dataset, and the images were divided into a training and test set containing 1140 images and 562 images, respectively.

Mammograms were obtained from the Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM) [28], which contains mammograms from 1522 cases (753 calcification cases and 891 mass cases, both benign and malignant) from 4 sites across the US. The dataset, a standardized and updated subset of the original DDSM, contains algorithmically obtained segmentations of regions of interest (ROI) and metadata information for each case. We obtained 1696 scanned film mammograms from the CBIS-DDSM and split the images into two subsets for training (1318 images) and testing (378 images).

MRI images were obtained from a MRI dataset of 831 patients with 3,596 brain metastases treated with primary stereotactic radiosurgery (SRS) at our institution between 2000-2018 [29]. Patients with prior resection or prior radiation treatment were excluded along with metastases <5 mm. We sampled 2000 slices of normal brain tissue and 2000 slices containing tumor. The final dataset of 4000 samples contains two classes ("tumor" and "non-tumor") and is split into 2680 training images and 1320 test images.

For all medical datasets, the primary outcome of interest was the presence of cancer in each image. Additionally, we used two common non-medical imaging datasets, MNIST [30] and CIFAR-10 [31], as a control for our study as a means to compare adversarial susceptibility between DL models for medical and non-medical images. The MNIST dataset is composed of 70,000 square 28x28 pixel grayscale images of handwritten single digits (60,000 training and 10,000 testing images). The images are divided into 10 classes representing the digits 0-9. The CIFAR-10 dataset consists of 60,000 32x32 color images in 10 classes (50,000 training images and 10,000 test images). For each dataset, the classes were balanced. All images were center-cropped and resized, and pixels were normalized to have unit variance.

Models

For all classification tasks, we used an Imagenet pretrained VGG16 [32] as the base network, followed by a dropout layer of rate 0.5, a flatten layer, a dense layer of 4096 neurons, another dropout layer of rate 0.5, a dense layer of 1024 neurons, and a K-neuron dense layer for classification. Details regarding model architecture and hyperparameters for model training are shown in the Supplement. We also used simple data augmentations including random rotations and horizontal/vertical flips.

Adversarial Attack Vulnerability

Three first-order adversarial attack methods—Fast Gradient Sign Method (FGSM) [15], Projected Gradient Descent (PGD) [20], and Basic Iterative Method (BIM) [12]—were used to generate adversarial images on the medical and non-medical image datasets. For each attack method, the aim is to maximize the model's classification error while minimizing the difference between the adversarial sample and original sample. All the attacks considered are bounded under a predefined maximum perturbation ε with respect to the L_∞ norm. Full descriptions of the adversarial attacks are detailed in the Supplement. Example images from each dataset, along with their adversarial counterparts, are shown in Figure 1.

We investigated whether adversarial attacks on DL models for medical images were more effective than those on DL models for non-medical images. We measured attack difficulty by determining the smallest maximum perturbation ε required for attacks to generally succeed.

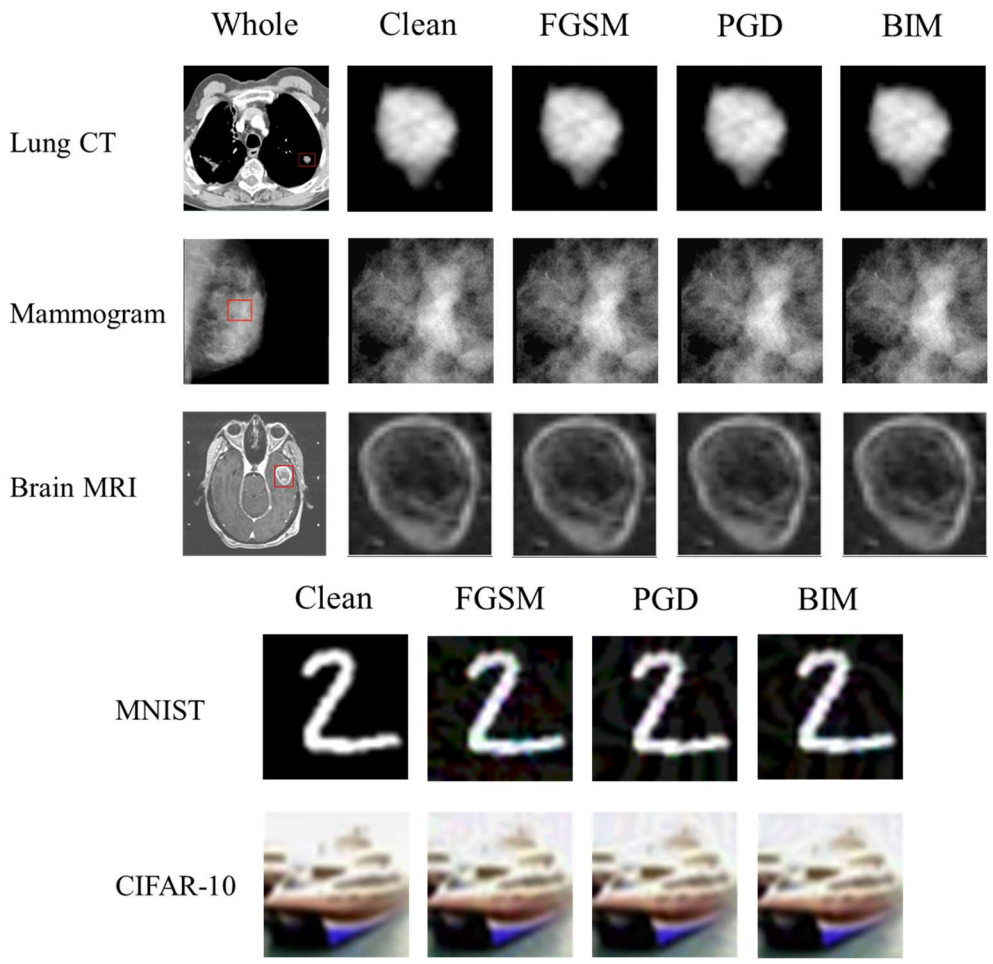

Figure 1. Example images and adversarial counterparts from the CT, mammogram, MRI, MNIST, and CIFAR-10 datasets. Adversarial images were generated using FGSM, PGD, and BIM attack methods. The maximum perturbation size used to bound the attacks were 0.003 for CT, 0.003 for mammogram, 0.004 for MRI, 0.1 for MNIST, and 0.02 for CIFAR-10.

Effectiveness of Adversarial Training

One successful defense mechanism against adversarial attacks is *adversarial training*, which aims to improve model robustness by integrating adversarial samples into the training stage [20, 21]. In particular, PGD adversarial training has been shown to achieve state-of-the-art robustness against L_∞ attacks on MNIST, CIFAR-10, and ImageNet datasets [20, 33, 34]. When Athalye et

al. broke an ensemble of adversarial defenses using novel attack techniques, PGD adversarial training was one of the few defenses which remained effective [35].

While adversarial training is a promising approach to defend models against attacks, its effectiveness varies by classification task. The success of adversarial training is correlated with the distance from the test image to the manifold of training images [36]. When the test images are far away from the training data, adversarial training becomes less effective at increasing model robustness. The selection of hyperparameters also influences the success of adversarial training [36]. Additionally, adversarial training for DL models have been primarily designed for non-medical image datasets with large labeled training sets, whereas medical datasets usually have small labeled training sets [37]. Due to these limitations of adversarial training, it's important to study the usefulness of adversarial training on a case-by-case basis.

Here, we proposed using multi-step PGD adversarial training to increase the robustness of our DL models against adversarial attacks. Under this *K*-PGD adversarial training algorithm, the inner loop constructs adversarial examples by PGD-*K* and the outer loop updates the model using minibatch SGD on the generated samples [20, 38]. The hyperparameters for adversarial training are detailed in the Supplement. We measured the effectiveness of adversarial training by comparing model accuracy on adversarial samples of varying maximum perturbation size before and after adversarial training.

Image Level Adversarial Vulnerability and Model Performance

We examined the effect of individual images' adversarial vulnerability (i.e., level of adversarial perturbation necessary for model prediction on the adversarial image to differ from that on the clean image) on model performance. We hypothesized that images most sensitive to adversarial

attack are more likely to be misclassified by the model in the absence of adversarial perturbation. By excluding images most sensitive to adversarial perturbation, we aimed to improve model performance on the remaining dataset under normal conditions (no attack). For each image, we calculated the minimum perturbation size necessary to flip the model prediction on the adversarial image (from the model prediction on the clean image). We identified the 20% of images most vulnerable to adversarial attack and excluded them from the test set. We then tested the performance of the original model on the reduced test set.

The proposed networks were implemented in Python 2.7 using TensorFlow 1.15.3 framework [39]. Adversarial attacks and training experiments were conducted via the Adversarial Robustness Toolbox (ART) version 1.4.1, a Python library for machine learning security (https://github.com/Trusted-AI/adversarial-robustness-toolbox) [40]. The code to reproduce the analyses and results is available online at https://github.com/Aneja-Lab-Yale/Aneja-Lab-Public-Adversarial-Imaging.

Results:

Adversarial Attack Sensitivity

Both medical and non-medical DL models were highly susceptible to adversarial attacks, resulting in dramatic decreases in DL model accuracy for all datasets. Prior to applying adversarial attacks, we achieved classification accuracies of 76.3% for CT, 75.4% accuracy for mammogram, 79.9% for MRI, 99.2% for MNIST, and 86.3% for CIFAR-10. Under PGD attack, model accuracy dropped to 24.7% for CT ($\varepsilon = 0.005$), 24.6% for mammograms ($\varepsilon = 0.005$),

20.5% for MRI ($\varepsilon = 0.005$), 0% for MNIST ($\varepsilon = 0.35$), 9% for CIFAR-10 ($\varepsilon = 0.05$) (Figure 2).

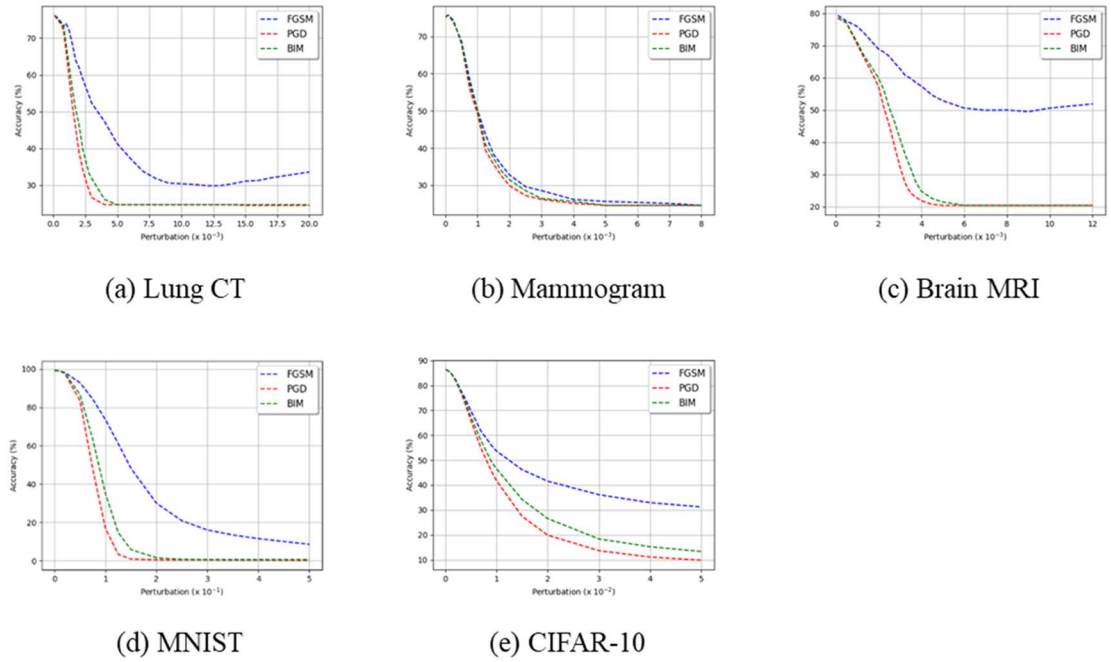

Figure 2. Classification accuracy of VGG16 model on adversarial examples generated by FGSM, BIM, and PGD attacks with increasing L_∞ maximum perturbation size ε. Model performance decreased as ε increased for all datasets and attack types.

Medical DL models appeared significantly more vulnerable to adversarial attacks compared to non-medical DL models. When examining the minimum perturbation size necessary for majority of attacks to be effective, a smaller ε was required for attacks to be successful on DL models for medical images compared to DL models for non-medical images. For the medical datasets, strong attacks (PGD and BIM) succeeded most of the time with tiny, visually imperceptible perturbations ($\varepsilon < 0.004$), while the non-medical datasets required much larger perturbations for attacks to be highly effective. Furthermore, the medical datasets themselves also varied in levels of susceptibility to adversarial attacks. For example, for a PGD attack with perturbation size of

0.002, model accuracy dropped to 39.1% for CT, 29.9% for mammogram, and 57.3% for MRI (Figure 2).

Effectiveness of Adversarial Training

By reducing the rate of misclassification on adversarial samples, adversarial training led to increased robustness of DL models against adversarial attacks. After adversarial training was applied to our original VGG16 model, classification accuracy on adversarial examples increased for all datasets (Figure 3). Adversarial training caused model accuracy to increase by up to 48% for CT, 29% for mammogram, and 53% for MRI (Table 1).

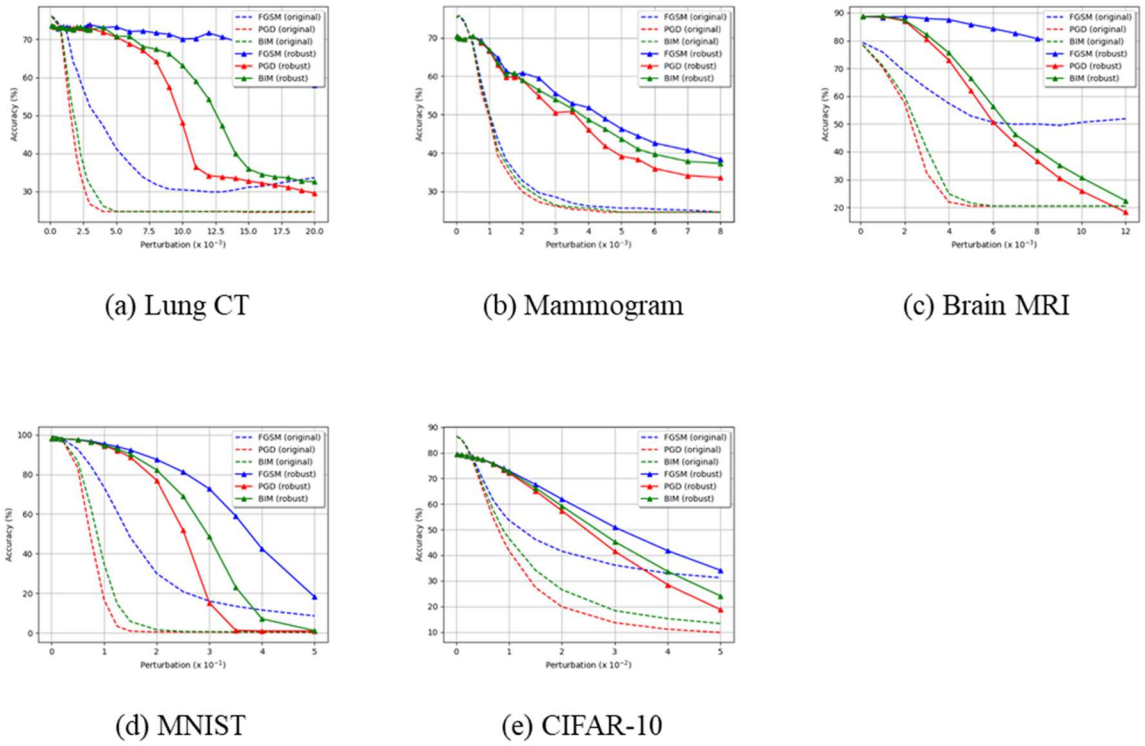

Figure 3. Classification accuracy of VGG16 model (before and after PGD adversarial training) on adversarial samples crafted by FGSM, BIM, and PGD with increasing L_∞ maximum perturbation size ε. Adversarial training significantly increased model accuracy for all classification tasks and attack types.

Attack	CT Accuracy (%)		Mammogram Accuracy (%)		MRI Accuracy (%)		MNIST Accuracy (%)		CIFAR-10 Accuracy (%)	
Model	Original	Adv trained	Original	Adv trained	Original	Original	Original	Adv trained	Original	Adv trained
No attack	76.33	73.67	75.40	70.37	86.33	86.33	99.16	98.11	86.33	79.36
FGSM	47.33	73.13	26.19	51.85	41.54	41.54	30.02	87.47	41.54	61.91
PGD	24.73	71.89	25.13	46.03	19.85	19.85	0.57	77.04	19.85	57.37
BIM	26.16	73.13	25.66	48.68	26.53	26.53	1.65	82.21	26.53	59.18

Table 1. Classification accuracy (%) of VGG16 model (before and after adversarial training) on clean test images and the 3 types of adversarial examples under maximum perturbation size of 0.004 (CT, mammogram, and MRI), 0.2 (MNIST), and 0.02 (CIFAR-10).

Image Level Adversarial Vulnerability and Model Performance

By using adversarial sensitivity as a metric to identify images most at risk for misclassification, we were able to improve overall DL model performance for all classification tasks. Test images were excluded if a PGD attack with perturbation size less than a certain threshold was sufficient to change the model prediction on the image. Excluding the images most susceptible to adversarial attack from the test set increased model accuracy by 5.3% for CT, 5.0% for mammogram, and 5.6% for MRI (Table 2).

	Model Accuracy (%) Original Test Set	Model Accuracy (%) Reduced Test Set	Change in Model Accuracy (%)
CT	76.33	81.64	5.31
Mammogram	75.40	80.39	4.99
MRI	79.85	85.48	5.63

Table 2. Classification accuracy (%) of VGG16 model on the original test set and the reduced test set excluding the 20% of test images most susceptible to adversarial attack. Images were excluded if PGD attack with perturbation size less than a certain threshold was sufficient to change the model prediction on the image. That threshold perturbation size was 0.0007 for CT, 0.00025 for mammogram, 0.0006 for MRI, 0.05 for MNIST, and 0.0025 for CIFAR-10.

Discussion:

The role of diagnostic imaging is increasing throughout clinical oncology. DL models are an emerging technique which have shown promise in helping analyze diagnostic images and represent a cost-effective tool to supplement human-decision making [41, 42]. However, vulnerability to adversarial attacks remains a potential barrier to fulfilling the promise of DL models in oncology. It is increasingly important to understand how adversarial images can be crafted to deceive DL models in the medical domain and whether proposed defenses against adversarial attacks from the non-medical domain represent a viable solution. In this study, we found that DL models for medical images are more vulnerable to adversarial attacks compared to DL models for non-medical images. Specifically, attacks on medical images require smaller perturbation sizes to generally succeed compared to attacks on non-medical images. Furthermore, we found adversarial training methods commonly used on non-medical imaging datasets to have limited effectiveness against adversarial attacks on medical images commonly found in oncology. Finally, we showed that identifying images most susceptible to adversarial attacks maybe helpful in improving overall performance of DL models on medical images.

This study corroborates previous literature showing that medical DL models are highly susceptible to adversarial attacks. Several recent works have found that state-of-the-art DL architectures like Inception and UNet perform poorly on medical imaging analysis tasks when placed under adversarial attack [9, 14, 43-45]. Our findings demonstrate that DL models for medical images (CT, mammography, MRI) are extremely vulnerable to adversarial attacks, with small perturbations (< 0.004) being associated with sharp drops in model performance (>50%). Adversarial perturbations on such a scale were confirmed to be imperceptible to the human eye, despite being sufficient to fool the DL model with great success. Our work extends the findings

of previous studies by showing that DL models for medical images across various imaging modalities all exhibit vulnerability to adversarial attack and that there are different levels of sensitivity to adversarial attack between imaging modalities. Furthermore, while most studies used only one fixed perturbation size for adversarial attack, we varied perturbation size along a broad range in order to examine the relationship between model performance and attack strength.

In addition, our results reinforce previous work which showed that DL models for medical images are more vulnerable to adversarial attack than DL models for non-medical images. Ma et al. showed that for adversarial attacks to generally succeed on medical images (fundoscopy, dermoscopy, and chest X-ray), the perturbation size necessary was much smaller than what was reported for non-medical images in another study by Shafahi et al. [14, 38]. We compared the adversarial sensitivity of DL models for medical images to that of DL models for non-medical images by using two non-medical image datasets (MNIST, CIFAR-10) as a control. By applying the same attack settings to all datasets, we determined that DL models for medical images were significantly more susceptible to adversarial attacks than DL models for non-medical images. One reason for this behavior could be that highly medical images are highly standardized and small adversarial perturbations dramatically distort their distribution in latent feature space [37, 46]. Another factor could be the overparameterization of DL models for medical image analysis, as sharp loss landscapes around medical images lead to higher adversarial vulnerability [14].

Our findings also support previous studies which found adversarial training to be an effective defense against adversarial attacks on DL image classification models [20]. In recent works, adversarial training has shown to improve DL model robustness for multiple medical imaging modalities like lung CT [47] and retinal OCT [37]. Our work proposed applying a PGD adversarial training approach to DL models for lung CTs, mammograms, and brain MRIs,

demonstrating significant improvement in model robustness for each dataset. The effectiveness of adversarial training was highly dependent on the parameters of adversarial training, especially the maximum perturbation size for PGD attack. While too-small perturbation sizes limit the increase in model robustness post-adversarial training, increasing the perturbation size beyond a certain threshold prevents the model from learning during training, causing poor model performance on both clean and adversarial samples. Our work extended previous studies by demonstrating how the performance of the DL model post-adversarial training is inversely proportional to the perturbation size of the adversarial samples on which the model is evaluated. While PGD adversarial training is effective in defending against weaker attacks, it showed less success with strong attacks with greater perturbation sizes.

Our results differed from previous work which found very limited success with applying adversarial training to DL models for medical imaging. For instance, Hirano et al. found that adversarial training did not lead to any increased model robustness for classifying dermoscopy, OCT, and chest X-ray images [48]. We found significant increases in model performance (up to 50% increases in model accuracy) on adversarial samples due to adversarial training. The difference in effectiveness of adversarial training can be attributed to the different adversarial training protocols used. At the same time, our adversarial training results on our medical DL models were far from satisfactory. Because medical images have fundamentally differently properties than non-medical images, adversarial defenses well-suited for non-medical images may not be generalizable to medical images [14, 37]. Still, our study is novel in its application of Madry's PGD training algorithm to DL models for medical image classification.

There are a number of limitations to our study which may bias our findings. First, we only used 2-class medical imaging classification tasks. Thus, our findings might not generalizable to multi-

class or regression problems using medical images. Given a number of medical diagnoses problems involve two or a small number of classes, our findings are likely still widely applicable to a large portion of medical imaging classification tasks. Also of note is the fact that we focused on first-order adversarial attacks rather than higher-order attacks, which have been shown to be more resistant against PGD adversarial training [49]. While the most commonly used adversarial attacks are first-order attacks, there is still need for additional research on how to defend DL models for medical images against higher-order attacks. A final limitation is that we used traditional supervised adversarial training to improve model robustness, while other nuanced methods like semi-supervised adversarial training and unsupervised adversarial training exist [37, 50, 51]. While we demonstrated that supervised adversarial training is an effective method to improve model performance on adversarial examples, an interesting direction for future work would be to compare the utility of supervised adversarial training with that of semi-supervised or unsupervised adversarial training on DL models for medical images.

Conclusion

In this work, we explored the issue of adversarial attacks on DL models used for medical image analysis in clinical oncology. The paper first examined the sensitivity of DL models to adversarial attacks across increasing perturbation magnitudes for different medical and non-medical imaging datasets, demonstrating that DL models for medical images are more susceptible to attacks than DL models for non-medical images. We then showed that PGD adversarial training effectively improved model robustness against adversarial attack. Finally, we saw that adversarial sensitivity of individual images could be used as a metric to improve model performance. By shedding light on the behavior of adversarial attacks on medical DL systems in

oncology, the findings from this paper can help facilitate the development of more secure medical imaging DL models and techniques to circumvent adversarial attacks.

References

1. Kann, B.H., et al., *Artificial Intelligence in Oncology: Current Applications and Future Directions.* Oncology (Williston Park), 2019. **33**(2): p. 46-53.
2. Aneja, S., E. Chang, and A. Omuro, *Applications of artificial intelligence in neuro-oncology.* Current Opinion in Neurology, 2019. **32**(6).
3. Siar, M. and M. Teshnehlab. *Brain Tumor Detection Using Deep Neural Network and Machine Learning Algorithm*. in *2019 9th International Conference on Computer and Knowledge Engineering (ICCKE)*. 2019.
4. Hashemzehi, R., et al., *Detection of brain tumors from MRI images base on deep learning using hybrid model CNN and NADE.* Biocybernetics and Biomedical Engineering, 2020. **40**(3): p. 1225-1232.
5. Jain, G., et al., *A deep learning approach to detect Covid-19 coronavirus with X-Ray images.* Biocybernetics and biomedical engineering, 2020. **40**(4): p. 1391-1405.
6. Kann, B.H., et al., *Multi-Institutional Validation of Deep Learning for Pretreatment Identification of Extranodal Extension in Head and Neck Squamous Cell Carcinoma.* Journal of Clinical Oncology, 2019. **38**(12): p. 1304-1311.
7. Cao, H., et al., *Dual-branch residual network for lung nodule segmentation.* Applied Soft Computing, 2020. **86**: p. 105934.
8. Tang, Y.-X., et al., *Automated abnormality classification of chest radiographs using deep convolutional neural networks.* npj Digital Medicine, 2020. **3**(1): p. 70.
9. Finlayson, S.G., et al., *Adversarial attacks on medical machine learning.* Science (New York, N.Y.), 2019. **363**(6433): p. 1287-1289.
10. Park, C.W., et al., *Artificial Intelligence in Health Care: Current Applications and Issues.* Journal of Korean medical science, 2020. **35**(42): p. e379-e379.
11. Kelly, C.J., et al., *Key challenges for delivering clinical impact with artificial intelligence.* BMC Medicine, 2019. **17**(1): p. 195.
12. Kurakin A, G.I., Bengio S, *Adversarial examples in the physical world.* arXiv:1607.02533, 2016.
13. Yuan, X., et al., *Adversarial Examples: Attacks and Defenses for Deep Learning.* arXiv e-prints, 2017: p. arXiv:1712.07107.
14. Ma, X., et al., *Understanding adversarial attacks on deep learning based medical image analysis systems.* Pattern Recognition, 2020: p. 107332.
15. Goodfellow IJ, S.J., Szegedy C., *Explaining and Harnessing Adversarial Examples.* arXiv:1412.6572, 2014.
16. Shaham, U., Y. Yamada, and S. Negahban, *Understanding adversarial training: Increasing local stability of supervised models through robust optimization.* Neurocomputing, 2018. **307**: p. 195-204.
17. Szegedy, C., et al., *Intriguing properties of neural networks. arXiv 2013.* arXiv preprint arXiv:1312.6199, 2013.
18. Shu, H., et al., *Adversarial Image Generation and Training for Deep Neural Networks.* arXiv e-prints, 2020: p. arXiv:2006.03243.
19. Tabacof, P. and E. Valle, *Exploring the Space of Adversarial Images.* arXiv e-prints, 2015: p. arXiv:1510.05328.
20. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A., *Towards Deep Learning Models Resistant to Adversarial Attacks.* arXiv:1706.06083, 2017.
21. Ren, K., et al., *Adversarial Attacks and Defenses in Deep Learning.* Engineering, 2020. **6**(3): p. 346-360.

22. Argaw, S.T., et al., *The state of research on cyberattacks against hospitals and available best practice recommendations: a scoping review.* BMC medical informatics and decision making, 2019. **19**(1): p. 10-10.
23. Singh Lallie, H., et al., *Cyber Security in the Age of COVID-19: A Timeline and Analysis of Cyber-Crime and Cyber-Attacks during the Pandemic.* arXiv e-prints, 2020: p. arXiv:2006.11929.
24. Williams, C.M., R. Chaturvedi, and K. Chakravarthy, *Cybersecurity Risks in a Pandemic.* Journal of medical Internet research, 2020. **22**(9): p. e23692-e23692.
25. Finlayson, S.G., et al., *Adversarial Attacks Against Medical Deep Learning Systems.* arXiv e-prints, 2018: p. arXiv:1804.05296.
26. Finlayson, S.G., et al., *Adversarial attacks on medical machine learning.* Science, 2019. **363**(6433): p. 1287.
27. Armato III, S.G., et al., *The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans.* Medical physics, 2011. **38**(2): p. 915-931.
28. Lee, R.S., et al., *A curated mammography data set for use in computer-aided detection and diagnosis research.* Scientific Data, 2017. **4**(1): p. 170177.
29. Chang, E., et al., *Comparison of Radiomic Feature Aggregation Methods for Patients with Multiple Tumors.* medRxiv, 2020: p. 2020.11.04.20226159.
30. Lecun, Y., *THE MNIST DATABASE of handwritten digits.* http://yann.lecun.com/exdb/mnist/.
31. Krizhevsky, A. and G. Hinton, *Learning multiple layers of features from tiny images.* 2009.
32. Simonyan, K. and A. Zisserman, *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556, 2014.
33. Xie, C., et al. *Feature denoising for improving adversarial robustness*. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
34. Xie, C., et al., *Feature Denoising for Improving Adversarial Robustness.* 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: p. 501-509.
35. Athalye, A., N. Carlini, and D. Wagner, *Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples.* arXiv e-prints, 2018: p. arXiv:1802.00420.
36. Zhang, H., et al., *The Limitations of Adversarial Training and the Blind-Spot Attack.* arXiv e-prints, 2019: p. arXiv:1901.04684.
37. Li, X., D. Pan, and D. Zhu, *Defending against adversarial attacks on medical imaging AI system, classification or detection?* arXiv e-prints, 2020: p. arXiv:2006.13555.
38. Shafahi, A., et al., *Adversarial Training for Free!* arXiv e-prints, 2019: p. arXiv:1904.12843.
39. Abadi, M., et al., *Tensorflow: Large-scale machine learning on heterogeneous distributed systems.* arXiv preprint arXiv:1603.04467, 2016.
40. Nicolae, M.-I., et al., *Adversarial Robustness Toolbox v1.0.0.* arXiv e-prints, 2018: p. arXiv:1807.01069.
41. Kyono, T., F.J. Gilbert, and M. van der Schaar, *MAMMO: A Deep Learning Solution for Facilitating Radiologist-Machine Collaboration in Breast Cancer Diagnosis.* arXiv e-prints, 2018: p. arXiv:1811.02661.
42. Park, A., et al., *Deep Learning–Assisted Diagnosis of Cerebral Aneurysms Using the HeadXNet Model.* JAMA Network Open, 2019. **2**(6): p. e195600-e195600.
43. Paschali, M., et al. *Generalizability vs. Robustness: Investigating Medical Imaging Networks Using Adversarial Examples*. in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. 2018. Cham: Springer International Publishing.
44. Wetstein, S.C., et al., *Adversarial Attack Vulnerability of Medical Image Analysis Systems: Unexplored Factors.* arXiv e-prints, 2020: p. arXiv:2006.06356.

45. Asgari Taghanaki, S., A. Das, and G. Hamarneh, *Vulnerability Analysis of Chest X-Ray Image Classification Against Adversarial Attacks.* arXiv e-prints, 2018: p. arXiv:1807.02905.
46. Li, X. and D. Zhu. *Robust Detection of Adversarial Attacks on Medical Images*. in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. 2020.
47. Paul, R., et al. *Mitigating Adversarial Attacks on Medical Image Understanding Systems*. in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. 2020.
48. Hirano, H., A. Minagi, and K. Takemoto, *Universal adversarial attacks on deep neural networks for medical image classification.* BMC Medical Imaging, 2021. **21**(1): p. 9.
49. Li, B., et al., *Certified Adversarial Robustness with Additive Noise.* arXiv e-prints, 2018: p. arXiv:1809.03113.
50. Chen, C., et al., *Spoof Face Detection Via Semi-Supervised Adversarial Training.* arXiv e-prints, 2020: p. arXiv:2005.10999.
51. Uesato, J., et al., *Are Labels Required for Improving Adversarial Robustness?* arXiv e-prints, 2019: p. arXiv:1905.13725.