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], James Duncan[3], Antonio Omuro[4], Roy Herbst[5], Harlan Krumholz[5,6], Sanjay Aneja[1,6]

1. Department of Therapeutic Radiology, Yale School of Medicine, New Haven, CT
2. Yale School of Medicine, New Haven, CT
3. Department of Biomedical Engineering, Yale University, New Haven, CT
4. Yale Brain Tumor Program, New Haven, CT
5. Department of Internal Medicine, Yale School of Medicine, New Haven, CT
6. 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 promise 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 images with small pixel-level perturbations designed to cause models to misclassify images. 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 examined adversarial attacks on DL classification models separately trained on three medical imaging modalities commonly used in oncology: computed tomography (CT), mammography, and magnetic resonance imaging (MRI). We investigated how iterative adversarial training could be employed to increase model robustness against three first-order attack methods.

Results: On unmodified images, we achieved classification accuracies of 75.4% for CT, 76.4% accuracy for mammogram, and 93.6% for MRI. Under adversarial attack, model accuracy showed a maximum absolute decrease of 49.8% for CT, 52.9% for mammogram, 87.3% for MRI. Adversarial training caused model accuracy on adversarial images to increase by up to 42.9% for CT, 35.7% for mammogram, and 73.2% for MRI.

Conclusion: Our results indicated that DL models for oncologic images are highly sensitive to adversarial attacks, as visually imperceptible degrees of perturbation are sufficient to deceive the model the majority of the time. Adversarial training mitigated the effect of adversarial attacks on model performance but was less successful against stronger attacks. 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 magnetic resonance imaging (MRI), computed tomography (CT), and X-ray images with classification accuracy often rivaling trained clinicians.[3-9] As widespread clinical implementation of DL models becomes a more realistic possibility, the safety and efficacy of such models in healthcare is becoming a topic of increasing importance.[10-12]

One concerning limitation of DL models that may hinder safe clinical implementation is their susceptibility to *adversarial attacks.* Adversarial attacks on DL models occur when data is manipulated with small pixel-level perturbations specifically designed to cause trained DL models to misclassify images.[13-16] The vulnerability of DL models to adversarial attacks stems from the fact that DL models possess less algorithmic stability, resulting in significantly different outputs when given inputs which vary only slightly.[17,18] Although images created from adversarial attacks are often visually imperceptible from unmodified images, they have been shown to significantly affect DL model performance.[19-21]

Previous work concerning adversarial attacks on DL models has largely focused on non-medical images, and the vulnerability of diagnostic images to such attacks is relatively unknown.[21,22] In addition, although techniques to defend against adversarial attacks have been proposed for non-medical images, their generalizability on medical imaging tasks is unclear. Given the current financial incentives for healthcare fraud and the recent rises in cyberattacks on hospital systems, adversarial attacks on clinically implemented DL models pose a unique threat to healthcare systems.[23-27] This suggests 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 three common imaging modalities in clinical oncology. Specifically, we investigated the relative susceptibility of DL models trained on three oncologic imaging modalities to adversarial attacks and the effectiveness of adversarial defense techniques in mitigating DL model performance loss when facing adversarial attacks.

Methodology

Datasets

We examined adversarial attacks on DL classification models separately trained on three medical imaging modalities commonly used in oncology—computed tomography (CT), mammography, and magnetic resonance imaging (MRI). Each DL classification model was trained to identify the presence or absence of malignancy when given an image. Each dataset was split into a training set and a testing set in a 2:1 ratio.

CT imaging data consisted of 2,600 lung nodules from the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) collection.[28] The dataset contains 1,018 thoracic CT scans collected from 15 clinical sites across the US. Lung nodules used for DL model training were identified by experienced thoracic radiologists. The presence of malignancy was based on associated pathologic reports. For patients without pathologic confirmation, malignancy was based on radiologist consensus.

Mammography imaging data consisted of 1696 lesions from the Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM).[29] The CBIS-DDSM contains mammograms from 1,566 patients at four sites across the US. Mammographic lesions used for DL model training were obtained based on algorithmically derived regions of interest based on clinical metadata. The presence of malignancy was based on verified pathologic reports.

MRI data consisted of brain MRIs from 831 patients from a single institution brain metastases registry.[30] The presence or absence of a malignancy was identified on 4,000 brain lesions seen on MRI. Regions of interest were identified by a multi-disciplinary team of radiation oncologists, neurosurgeons, and radiologists. Presence of cancer was identified based on pathologic confirmation or clinical consensus.

To compare the relative adversarial attack vulnerability of DL models trained on oncologic images compared to non-medical images, two additional DL classification models were trained on established non-medical datasets. The MNIST dataset consists of 70,000 hand written numerical digits.[31] The CIFAR-10 datasets includes 60,000 color images of ten non-medical objects.[32]

Models

For all DL classification models, we used a pre-trained convolutional neural network with the VGG16 architecture.[33] DL models were trained using data-augmentation and optimized using stochastic gradient descent. Details regarding model architecture and hyperparameter selection for DL model training are provided in the Supplement.

Adversarial Attack Vulnerability of DL Models

Three commonly employed first-order adversarial attack methods—Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), and Projected Gradient Descent (PGD)—were used to generate adversarial images on the medical and non-medical image datasets (Figure 1). Each attack method aims to maximize the DL model's classification error while minimizing the difference between the adversarial image and original image. All the attacks considered are bounded under a predefined perturbation size ε, which represents the maximum change to pixel values of an image.

The single step FGSM attack perturbs the original example by a fixed amount along the direction (sign) of the gradient of adversarial loss.[16]

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

BIM iteratively perturbs the normal example with smaller step size and clips the pixel values of the updated adversarial example after each step into a permitted range.[13]

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

Known as the strongest first-order attack, PGD iteratively perturbs the input with smaller step size and after each iteration, the updated adversarial example is projected onto the ε-ball of x and clipped onto a permitted range.[21]

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

Additional information regarding adversarial attacks methods is provided in the Supplement.

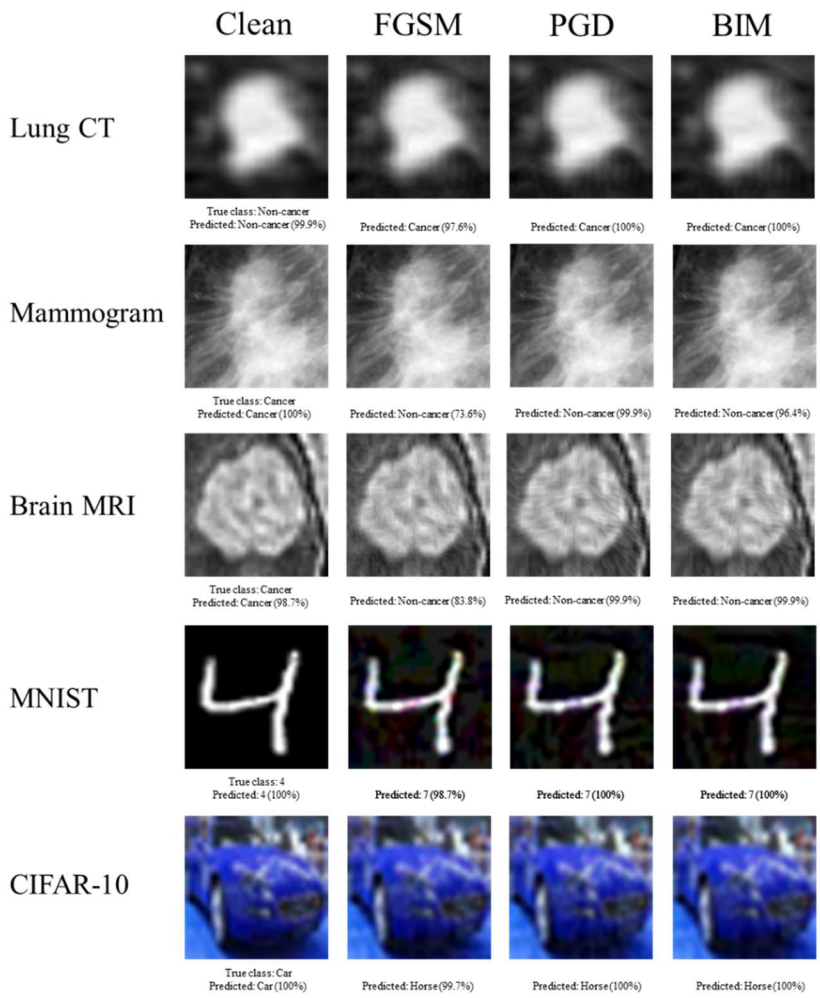


Figure 1. Examples of clean images and their adversarial counterparts generated using FGSM, PGD, and BIM attack methods. The percentage displayed represents the probability predicted by the model that the image is of a certain class.

Susceptibility to Adversarial Attacks

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 perturbation ε required for attacks to generally succeed. Given that larger image perturbations are more likely to be identified, images requiring a large pixel

perturbation for attack success are considered less susceptible to adversarial attacks. Conversely, images requiring a small pixel-level perturbation for attack success are more vulnerable to adversarial attacks.

Effectiveness of Adversarial Training Defense

One successful defense mechanism against adversarial attacks is *adversarial training*, which aims to improve model robustness by integrating adversarial samples into the training stage.[21,22] By training on both adversarial and normal images, the DL model learns to classify adversarial samples with higher accuracy compared to models trained on only normal samples. We investigated the effectiveness of an iterative adversarial training approach on the DL models trained on medical images. We measured the effectiveness of adversarial training by comparing model accuracy on adversarial samples of varying perturbation size before and after adversarial training. Details regarding our adversarial training protocol are detailed in the Supplement.

Image Level Adversarial Sensitivity and Model Performance

We examined each individual image's adversarial sensitivity, as measured by the level of pixel-level perturbation necessary for DL model prediction to change as compared to an unperturbed image. Images requiring smaller perturbation to change DL model predictions would be considered more sensitive to adversarial attacks. We hypothesized that images most sensitive to adversarial attacks were also the images most likely to be misclassified by the DL model. By excluding images most sensitive to adversarial perturbation, we aimed to improve model performance on the remaining dataset. 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.

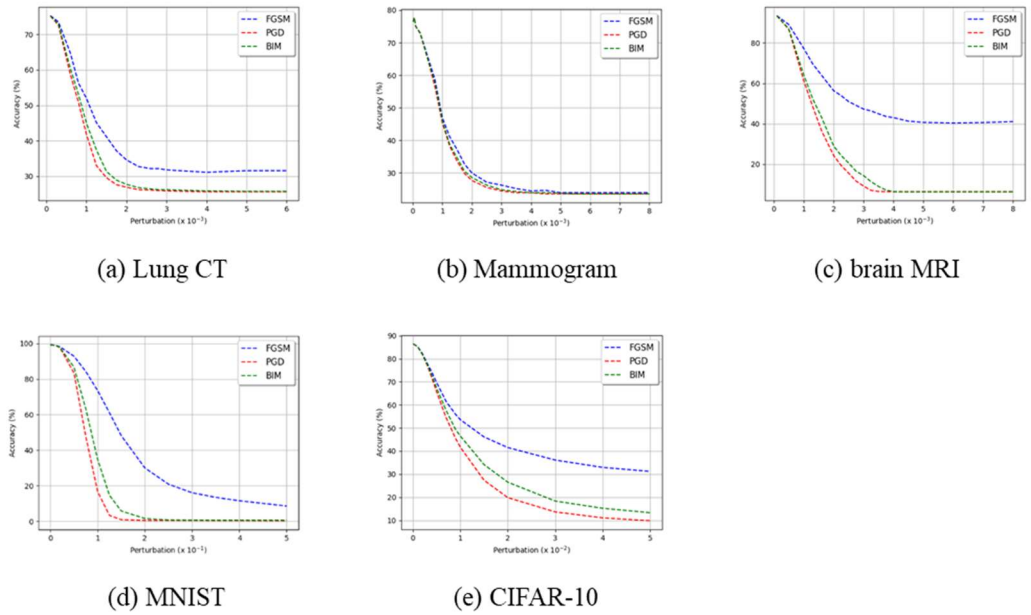
The proposed networks were implemented in Python 2.7 using TensorFlow v1.15.3 framework.[34] Adversarial attacks were created using the Adversarial Robustness Toolbox v1.4.1.[35] The code to reproduce the analyses and results is available online at https://github.com/Aneja-Lab-Yale/Aneja-Lab-Public-Adversarial-Imaging.

Results:

Adversarial Attack Vulnerability of DL Models

Both medical and non-medical DL models were highly susceptible to adversarial attacks, which resulted in dramatic decreases in model accuracy for all datasets. Prior to applying adversarial attacks, we achieved classification accuracies of 75.4% for CT, 76.4% accuracy for mammogram, 93.6% for MRI, 99.1% for MNIST, and 86.1% for CIFAR-10. Under PGD attack, model accuracy showed a maximum absolute decrease of 49.8% for CT, 52.9% for mammogram, 87.3% for MRI, 99.1% for MNIST, and 77.3% for CIFAR-10 (Figure 2).

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 example, under PGD attack with a perturbation size of 0.004, model accuracy was 25.6% for CT, 23.9% for mammogram, 6.36% for MRI, 99.0% for MNIST, and 71.9% for CIFAR-10 (Table 1). For the medical datasets, strong attacks (PGD and BIM) succeeded most of the time with tiny perturbations ($\varepsilon < 0.004$), while the non-medical datasets required much larger perturbations ($\varepsilon > 0.07$ for MNIST, $\varepsilon > 0.01$ for CIFAR-10) for attacks to be majority effective.



(a) Lung CT (b) Mammogram (c) brain MRI

(d) MNIST (e) CIFAR-10

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. *Note that the horizontal axis (ε) was scaled to 10^{-3} for graphs (a) to (c), to 10^{-1} for (d), and to 10^{-2} for (e).

Attack	Perturbation size *ε*	CT Accuracy (%)	Mammogram Accuracy (%)	MRI Accuracy (%)	MNIST Accuracy (%)	CIFAR-10 Accuracy (%)
No attack		75.41	76.43	93.64	99.13	86.13
FGSM	0.001	51.98	46.96	77.27	99.05	85.32
	0.002	34.62	30.00	56.36	99.05	82.39
	0.004	31.12	24.46	43.03	99.04	74.29
	0.006	31.59	23.93	40.38	98.96	65.27
PGD	0.001	41.84	45.36	61.36	99.06	85.29
	0.002	26.92	27.68	24.32	99.05	81.93
	0.004	25.64	23.93	6.36	99.01	71.90
	0.006	25.64	23.57	6.36	98.92	59.98
BIM	0.001	44.99	46.07	64.24	99.06	85.32
	0.002	27.74	28.57	29.02	99.05	82.06
	0.004	25.87	23.93	6.44	99.01	72.84
	0.006	25.76	23.57	6.36	98.93	62.28

Table 1. Effects of adversarial attacks of varying perturbation sizes on model classification accuracy. Adversarial samples were created by FGSM, BIM, and PGD with increasing L_∞ maximum perturbation size ε. Models for medical datasets (CT, mammogram, and MRI) required smaller attack perturbation sizes than models for non-medical datasets (MNIST, CIFAR-10) for attacks to be generally effective.

Measuring 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 absolute accuracy of the model on adversarial images to increase by up to 42.9% for CT, 35.7% for mammogram, and 73.2% for MRI. As perturbation size is increased for attacks used to generate adversarial test images, adversarial training becomes less effective at improving model performance.

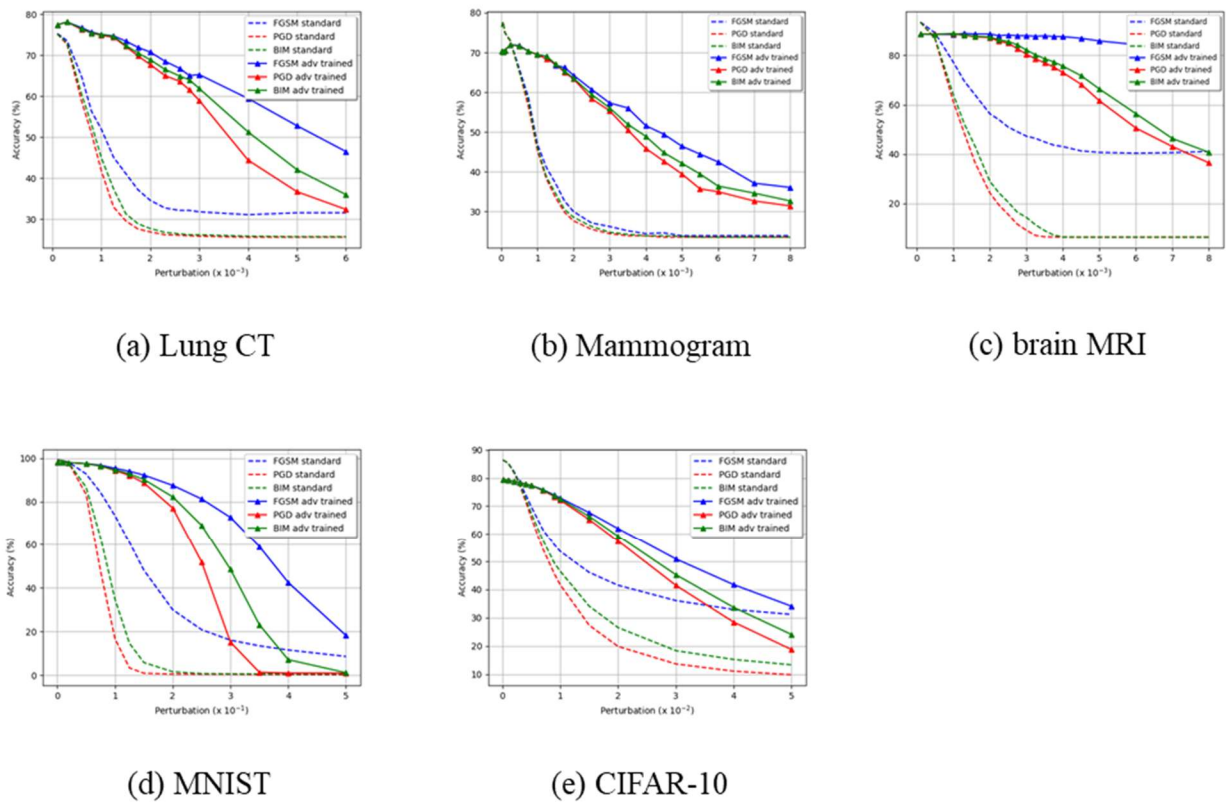


Figure 3. Comparison of model classification accuracy before and after adversarial training on adversarial samples crafted by FGSM, BIM, and PGD with increasing L_∞ maximum perturbation size ε. Adversarial training significantly increased model accuracy for all classification tasks and attack types. *Note that the horizontal axis (ε) was scaled to 10^{-3} for graphs (a) to (c), to 10^{-1} for (d), and to 10^{-2} for (e).

Image Level Adversarial Sensitivity and Model Performance

By using image level 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 the model's absolute accuracy by 5.9% for CT, 3.7% for mammogram, and 5.2% for MRI (Table 2).

	Model Accuracy (%) Original Test Set	Model Accuracy (%) Adversarially Aware Test Set	Change in Model Accuracy (%)
CT	75.41	81.31	5.90
Mammogram	76.43	80.13	3.70
MRI	93.64	98.82	5.18

Table 2. Classification accuracy (%) of VGG16 model on the original test set and the 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 0.0003 for CT, 0.00025 for mammogram, and 0.0006 for MRI.

Discussion:

To our knowledge, our study is the first to investigate both the adversarial vulnerability of DL models for multiple imaging modalities in clinical oncology and the use of an iterative adversarial training approach to defend such models against first-order adversarial attacks. As the role of diagnostic imaging increases throughout clinical oncology, deep learning has shown to be incredibly powerful for medical image analysis tasks and represent a cost-effective tool to supplement human decision-making.[36-38] However, vulnerability to adversarial attacks remains a potential barrier to fulfilling the promise of DL models in oncology. It is crucial 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 much smaller perturbation sizes to generally succeed compared to attacks on non-medical images. Furthermore, we found that adversarial training methods commonly used on non-medical imaging datasets mitigate the effects of adversarial attacks on DL models for oncological images. Finally, we showed that identifying images most susceptible to adversarial attacks maybe helpful in improving overall performance of DL models on medical images.

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.[10,15,39-42] Our work extends the findings of previous studies by showing that DL models for oncologic images across various imaging modalities (CT, mammography, MRI) exhibit extreme vulnerability to adversarial attack, with small perturbations (< 0.004) being associated with sharp drops in model performance. We also show that DL models exhibited different levels of sensitivity to adversarial attack across different imaging modalities. Furthermore, while most studies used only one fixed perturbation size for adversarial attack, we varied perturbation size along a broad range to examine the relationship between model performance and attack strength.

In addition, our results corroborate previous work which showed that DL models for medical images are more vulnerable to adversarial attack than DL models for non-medical images (based on the minimum perturbation size necessary for attacks to generally succeed).[15,43] By using MNIST and CIFAR-10 as a control and applying the same attack settings to DL models

for all datasets, we determined that DL models for medical images were much more susceptible to adversarial attacks than DL models for non-medical images. One reason for this behavior could be that medical images are highly standardized and small adversarial perturbations dramatically distort their distribution in the latent feature space.[44,45] 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.[15]

In the past, adversarial training on medical DL models have shown mixed results. In some studies, adversarial training improved DL model robustness for multiple medical imaging modalities like lung CT and retinal OCT.[44,46,47] On the other hand, Hirano et al. found that adversarial training generally did not increase model robustness for classifying dermoscopy, OCT, and chest X-ray images.[48] The difference in effectiveness of adversarial training can be attributed to the different adversarial training protocols used (e.g. single-step vs. multi-step, attack(s) used to generate training samples). It's important to note that even in studies where adversarial training showed success in improving model robustness, the results were still not close to ideal, as some misclassification occurs even after adversarial training. This is expected as adversarial training, while capable of improving model accuracy on adversarial examples, has limits in effectiveness against strong attacks even on non-medical image datasets.[21]

Our work applied an iterative adversarial training approach to DL models for lung CTs, mammograms, and brain MRIs, demonstrating substantial improvement in model robustness for all imaging modalities. The effectiveness of adversarial training was highly dependent on the hyperparameters of adversarial training, especially the perturbation size for 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 demonstrated 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 adversarial training is effective in defending against weaker attacks with smaller perturbation magnitudes, it showed less success with strong attacks. While adversarial training proved successful at improving model performance on adversarial examples, our results were still far from satisfactory. One contributing factor is that medical images have fundamentally differently properties than non-medical images.[15,44] Thus, adversarial defenses well-suited for non-medical images may not be generalizable to medical images.

We also showed that image level adversarial sensitivity, defined by the level of adversarial perturbation necessary to change image class predicted by model, is a useful metric for identifying normal images most at-risk for misclassification even under no attack. This has potentially useful clinical implications as we can improve the accuracy of the DL model by excluding such at-risk images from the rest of the dataset and providing them to a radiologist rather than the DL model for examination.

There are several limitations to our study which may bias our findings. First, we only used two-class medical imaging classification tasks. Thus, our findings might not generalize to multi-class or regression problems using medical images. Given that many medical diagnostic problems involve 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 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.[44,50,51] 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 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 BH, Thompson R, Thomas CR, Jr., Dicker A, Aneja S. Artificial Intelligence in Oncology: Current Applications and Future Directions. *Oncology (Williston Park).* 2019;33(2):46-53.
2. Aneja S, Chang E, Omuro A. Applications of artificial intelligence in neuro-oncology. *Current Opinion in Neurology.* 2019;32(6).
3. Siar M, Teshnehlab M. Brain Tumor Detection Using Deep Neural Network and Machine Learning Algorithm. Paper presented at: 2019 9th International Conference on Computer and Knowledge Engineering (ICCKE); 24-25 Oct. 2019, 2019.
4. Hashemzehi R, Mahdavi SJS, Kheirabadi M, Kamel SR. Detection of brain tumors from MRI images base on deep learning using hybrid model CNN and NADE. *Biocybernetics and Biomedical Engineering.* 2020;40(3):1225-1232.
5. Jain G, Mittal D, Thakur D, Mittal MK. A deep learning approach to detect Covid-19 coronavirus with X-Ray images. *Biocybernetics and biomedical engineering.* 2020;40(4):1391-1405.
6. Kann BH, Hicks DF, Payabvash S, 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):1304-1311.
7. Cao H, Liu H, Song E, et al. Dual-branch residual network for lung nodule segmentation. *Applied Soft Computing.* 2020;86:105934.
8. Tang Y-X, Tang Y-B, Peng Y, et al. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *npj Digital Medicine.* 2020;3(1):70.
9. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health.* 2019;1(6):e271-e297.
10. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science.* 2019;363(6433):1287-1289.
11. Park CW, Seo SW, Kang N, et al. Artificial Intelligence in Health Care: Current Applications and Issues. *J Korean Med Sci.* 2020;35(42):e379-e379.
12. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine.* 2019;17(1):195.
13. Kurakin A GI, Bengio S. Adversarial examples in the physical world. *arXiv:160702533.* 2016.
14. Yuan X, He P, Zhu Q, Li X. Adversarial Examples: Attacks and Defenses for Deep Learning. *arXiv e-prints.* 2017:arXiv:1712.07107.
15. Ma X, Niu Y, Gu L, et al. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition.* 2020:107332.
16. Goodfellow IJ SJ, Szegedy C. Explaining and Harnessing Adversarial Examples. *arXiv:14126572.* 2014.
17. Shaham U, Yamada Y, Negahban S. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing.* 2018;307:195-204.
18. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. arXiv 2013. *arXiv preprint arXiv:13126199.* 2013.
19. Shu H, Shi R, Zhu H, Chen Z. Adversarial Image Generation and Training for Deep Neural Networks. *arXiv e-prints.* 2020:arXiv:2006.03243.
20. Tabacof P, Valle E. Exploring the Space of Adversarial Images. *arXiv e-prints.* 2015:arXiv:1510.05328.
21. Madry A, Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv:170606083.* 2017.

22. Ren K, Zheng T, Qin Z, Liu X. Adversarial Attacks and Defenses in Deep Learning. *Engineering.* 2020;6(3):346-360.
23. Argaw ST, Bempong N-E, Eshaya-Chauvin B, Flahault A. The state of research on cyberattacks against hospitals and available best practice recommendations: a scoping review. *BMC Med Inform Decis Mak.* 2019;19(1):10-10.
24. Singh Lallie H, Shepherd LA, Nurse JRC, 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:arXiv:2006.11929.
25. Williams CM, Chaturvedi R, Chakravarthy K. Cybersecurity Risks in a Pandemic. *J Med Internet Res.* 2020;22(9):e23692-e23692.
26. Finlayson SG, Chung HW, Kohane IS, Beam AL. Adversarial Attacks Against Medical Deep Learning Systems. *arXiv e-prints.* 2018:arXiv:1804.05296.
27. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science.* 2019;363(6433):1287.
28. Armato III SG, McLennan G, Bidaut L, 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):915-931.
29. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL. A curated mammography data set for use in computer-aided detection and diagnosis research. *Scientific Data.* 2017;4(1):170177.
30. Chang E, Joel M, Chang HY, et al. Comparison of Radiomic Feature Aggregation Methods for Patients with Multiple Tumors. *medRxiv.* 2020:2020.2011.2004.20226159.
31. Lecun Y. THE MNIST DATABASE of handwritten digits. *http://yannlecuncom/exdb/mnist/.*
32. Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. 2009.
33. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:14091556.* 2014.
34. Abadi M, Agarwal A, Barham P, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:160304467.* 2016.
35. Nicolae M-I, Sinn M, Tran MN, et al. Adversarial Robustness Toolbox v1.0.0. *arXiv e-prints.* 2018:arXiv:1807.01069.
36. Kyono T, Gilbert FJ, van der Schaar M. MAMMO: A Deep Learning Solution for Facilitating Radiologist-Machine Collaboration in Breast Cancer Diagnosis. *arXiv e-prints.* 2018:arXiv:1811.02661.
37. Park A, Chute C, Rajpurkar P, et al. Deep Learning–Assisted Diagnosis of Cerebral Aneurysms Using the HeadXNet Model. *JAMA Network Open.* 2019;2(6):e195600-e195600.
38. Sahiner B, Pezeshk A, Hadjiiski LM, et al. Deep learning in medical imaging and radiation therapy. *Med Phys.* 2019;46(1):e1-e36.
39. Paschali M, Conjeti S, Navarro F, Navab N. Generalizability vs. Robustness: Investigating Medical Imaging Networks Using Adversarial Examples. Paper presented at: Medical Image Computing and Computer Assisted Intervention – MICCAI 2018; 2018//, 2018; Cham.
40. Wetstein SC, González-Gonzalo C, Bortsova G, et al. Adversarial Attack Vulnerability of Medical Image Analysis Systems: Unexplored Factors. *arXiv e-prints.* 2020:arXiv:2006.06356.
41. Asgari Taghanaki S, Das A, Hamarneh G. Vulnerability Analysis of Chest X-Ray Image Classification Against Adversarial Attacks. *arXiv e-prints.* 2018:arXiv:1807.02905.
42. Yoo TK, Choi JY. Outcomes of Adversarial Attacks on Deep Learning Models for Ophthalmology Imaging Domains. *JAMA Ophthalmology.* 2020;138(11):1213-1215.
43. Shafahi A, Najibi M, Ghiasi A, et al. Adversarial Training for Free! *arXiv e-prints.* 2019:arXiv:1904.12843.

44. Li X, Pan D, Zhu D. Defending against adversarial attacks on medical imaging AI system, classification or detection? *arXiv e-prints.* 2020:arXiv:2006.13555.
45. Li X, Zhu D. Robust Detection of Adversarial Attacks on Medical Images. Paper presented at: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI); 3-7 April 2020, 2020.
46. Paul R, Schabath M, Gillies R, Hall L, Goldgof D. Mitigating Adversarial Attacks on Medical Image Understanding Systems. Paper presented at: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI); 3-7 April 2020, 2020.
47. Vatian A, Gusarova N, Dobrenko N, et al. Impact of Adversarial Examples on the Efficiency of Interpretation and Use of Information from High-Tech Medical Images. Paper presented at: 2019 24th Conference of Open Innovations Association (FRUCT); 8-12 April 2019, 2019.
48. Hirano H, Minagi A, Takemoto K. Universal adversarial attacks on deep neural networks for medical image classification. *BMC Medical Imaging.* 2021;21(1):9.
49. Li B, Chen C, Wang W, Carin L. Certified Adversarial Robustness with Additive Noise. *arXiv e-prints.* 2018:arXiv:1809.03113.
50. Chen C, Yuan W, Lu X, Ma L. Spoof Face Detection Via Semi-Supervised Adversarial Training. *arXiv e-prints.* 2020:arXiv:2005.10999.
51. Uesato J, Alayrac J-B, Huang P-S, Stanforth R, Fawzi A, Kohli P. Are Labels Required for Improving Adversarial Robustness? *arXiv e-prints.* 2019:arXiv:1905.13725.

Supplementary Material

Adversarial Attack Methods

Attack Type	Equation	Parameters
Fast Gradient Sign Method (FGSM)	$x_{adv} = x + \varepsilon\, sign(\nabla_x J(x, y))$	x_{adv} = adversarial image x = clean input image ε = perturbation size J = loss function y = target label
Projected Gradient Descent (PGD)	$x^0 = x,$ $x^t = Clip_{x,\epsilon}\{x^{t-1} + \alpha\, sign(\nabla_x J(x^t, y))$	x^0 = clean input image x^i = adversarial image at i^{th}step ε = maximum perturbation size α = perturbation step size J = loss function y = target label Clip{ } function limits updated adversarial sample to within range of ϵ ball ([x-ϵ, x+ϵ]) and the input space ([0,1] for pixel values)
Basic Iterative Method (BIM)	$x^0 = x,$ $x^t = \prod_\epsilon(x^{t-1} + \alpha\, sign(\nabla_x J(x^t, y))$	x^0 = clean input image x^i = adversarial image at i^{th}step ε = maximum perturbation size α = perturbation step size J = loss function y = target label y = target label ∏ { } function projects the adversarial example back onto the ϵ ball ([x-ϵ, x+ϵ])

Table S1. Equations and parameters for FGSM, PGD, and BIM attack methods. The number of perturbation steps for BIM and PGD are both set to 10, and the step sizes are set to ε/10 and ε/4 for BIM and PGD, respectively.

Data Preprocessing and Augmentation

For each medical dataset, the classes ("cancer" and "noncancer") were balanced. All images were center-cropped and resized, and pixels were normalized to have unit variance. For training the medical DL models, we used simple data augmentations: horizontal and vertical flips as well as random rotations with angles ranging between − 20° and 20°.

Model Architecture

Layer (type)	Output Shape	Param #
input_1 (Input Layer)	(None, 224, 224, 3)	0
block1_conv1 (Conv2D)	(None, 224, 224, 64)	1792
block1_conv2 (Conv2D)	(None, 224, 224, 64)	36928
block1_pool (MaxPooling2D)	(None, 112, 112, 64)	0
block2_conv1 (Conv2D)	(None, 112, 112, 128)	73856
block2_conv2 (Conv2D)	(None, 112, 112, 128)	147584
block2_pool (MaxPooling2D)	(None, 56, 56, 128)	0
block3_conv1 (Conv2D)	(None, 56, 56, 256)	295168
block3_conv2 (Conv2D)	(None, 56, 56, 256)	590080
block3_conv3 (Conv2D)	(None, 56, 56, 256)	590080
block3_pool (MaxPooling2D)	(None, 28, 28, 256)	0
block4_conv1 (Conv2D)	(None, 28, 28, 512)	1180160
block4_conv2 (Conv2D)	(None, 28, 28, 512)	2359808
block4_conv3 (Conv2D)	(None, 28, 28, 512)	2359808
block4_pool (MaxPooling2D)	(None, 14, 14, 512)	0
block5_conv1 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv2 (Conv2D)	(None, 14, 14, 512)	2359808
block5_conv3 (Conv2D)	(None, 14, 14, 512)	2359808
block5_pool (MaxPooling2D)	(None, 7, 7, 512)	0
dropout (Dropout)	(None, 7, 7, 512)	0
flatten (Flatten)	(None, 25088)	0
fc1 (Dense)	(None, 4096)	102764544
dropout_1 (Dropout)	(None, 4096)	0
fc2 (Dense)	(None, 1024)	4195328
predictions (Dense)	(None, 2)	2050
Total params: 121,676,610		

Table S2. Model parameters of VGG16 model.

Model Training Parameters

Parameters	CNN-1 (CT)	CNN-2 (Mammogram)	CNN-3 (MRI)	CNN-4 (MNIST)	CNN-5 (CIFAR-10)
Input Image Size	224 x 224	116 x 116	224 x 224	32 x 32	32 x 32
Batch Size	50	50	50	64	128
Max Epochs	200	100	100	20	60
Patience	50	20	5	0	10
Initial Learning Rate	0.0002	0.0002	0.0002	0.01	0.001
Learning rate decay	1e-6	1e-6	1e-6	0	1e-5
Momentum	0.9	0.9	0.9	0	0.9
Loss	Binary cross-entropy	Binary cross-entropy	Binary cross-entropy	Categorical cross-entropy	Categorical cross-entropy
Optimizer	SGD	SGD	SGD	SGD	SGD

Table S3. Training parameters for VGG16 models for all datasets.

Adversarial Training Protocol

We used a multi-step PGD adversarial training to increase the robustness of our DL models against adversarial attacks. In each batch, 50% of training samples were normal images, and the other 50% were adversarial images generated by PGD attack.

The hyperparameters for adversarial training are detailed below.

Parameters	CNN-1 (CT)	CNN-2 (Mammogram)	CNN-3 (MRI)	CNN-4 (MNIST)	CNN-5 (CIFAR-10)
Epsilon	0.003	0.002	0.004	0.2	0.02
Epsilon step	0.00075	0.0005	0.001	0.05	0.005
Max Iterations	7	7	7	7	7
# random initializations	1	1	1	1	1
Number of Epochs	200	100	200	50	100
Batch size	128	128	128	128	128

Table S4. Adversarial training parameters for multi-step PGD training of VGG16 models.

Attack	CT Accuracy (%)		Mammogram Accuracy (%)		MRI Accuracy (%)	
Model	Original	Adv trained	Original	Adv trained	Original	Adv trained
No attack	75.41	77.39	76.43	70.36	93.64	88.41
FGSM	31.82	65.27	26.25	57.32	47.27	87.95
PGD	25.87	58.97	24.46	55.36	9.17	80.45
BIM	26.22	61.89	24.82	56.07	14.39	82.12

Table S5. 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 0.003 (CT, mammogram, and MRI).