medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

# Using Adversarial Images to Assess the Stability of Deep Learning Models Trained on Diagnostic Images in Oncology

Marina Z. Joel BS[1], Sachin Umrao PhD[1], Enoch Chang BA[1], Rachel Choi BA[1], Daniel X. Yang MD[1], James S. Duncan PhD[2], Antonio Omuro MD[3], Roy Herbst MD PhD[4], Harlan M. Krumholz MD[4,5] SM, Sanjay Aneja MD[1,5]

1. Department of Therapeutic Radiology, Yale School of Medicine
2. Department of Biomedical Engineering, Yale University
3. Department of Neurology, Yale School of Medicine
4. Department of Medicine, Yale School of Medicine
5. Center for Outcomes Research and Evaluation at Yale (CORE)

**Corresponding Author:**

Sanjay Aneja, MD

Assistant Professor, Department of Therapeutic Radiology

Center for Outcomes Research and Evaluation (CORE)

Yale School of Medicine

330 Cedar Street, CB326

New Haven, CT 06510

Telephone: 475.331.2202

Email: sanjay.aneja@yale.edu



NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

**Translational Relevance:**

Deep learning (DL) models have demonstrated great promise for classifying diagnostic images within oncology. An on-going challenge to DL models on diagnostic images is generalizability and classification stability. DL models trained on diagnostic images have shown dramatic differences in model outputs in response to even small pixel level changes to images. Adversarial images are manipulated images which such small pixel-level perturbations and are specifically designed to test the stability of DL models. In this study, we use adversarial images to demonstrate the instability of DL models trained across common oncologic imaging modalities. We also highlight the utility of an iterative adversarial training approach as a method to improve the stability of DL models trained on diagnostic images. Our study provides insight into the development of more stable and robust DL algorithms for diagnostic image classification which can safely implemented into clinical practice.

**Abstract**

**Purpose:** Deep learning (DL) models have rapidly become a popular and cost-effective tool for image classification within oncology. A major limitation of DL models is output instability, as small perturbations in input data can dramatically alter model output. The purpose of the study is to investigate the robustness of DL models in the oncologic image domain through the application of *adversarial images*: manipulated images with small pixel-level perturbations designed to assess the stability of DL models.

**Experimental Design:** We examined the impact of adversarial images on the classification accuracies of DL models trained to classify cancerous lesions across three common oncologic imaging modalities (CT, mammogram, and MRI). The CT model was trained to classify malignant lung nodules using the LIDC dataset. The mammogram model was trained to classify malignant breast lesions using the DDSM dataset. The MRI model was trained to classify brain

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

metastases using an institutional dataset. We also explored the utility of an iterative adversarial training approach to improve the stability of DL models to small pixel-level changes.

**Results:** Oncologic images showed instability with small pixel-level changes. A pixel-level of perturbation of .004 resulted in a majority of oncologic images to be misclassified by their respective DL models (CT 25.64%, mammogram 23.93%, MRI 6.36%). Adversarial training mitigated improved the stability and robustness of DL models trained on oncologic images compared to naive models [(CT 67.72% vs 26.92%), mammogram (63.39% vs 27.68%), MRI (87.20% vs 24.32%)].

**Conclusions:** DL models naively trained on oncologic images exhibited dramatic instability to small pixel-level changes resulting in substantial decreases in accuracy. Adversarial training techniques improved the stability and robustness of DL models to such pixel-level changes. Prior to clinical implementation, adversarial training should be considered to proposed DL models to improve overall performance and safety.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

**Introduction:**

Deep learning (DL) algorithms have the promise to improve the quality of diagnostic image interpretation within oncology (1,2). Models generated from DL algorithms 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). However, the success of DL models depends, in part, on their generalizability and stability. DL algorithms, in particular, have been shown to vary output based on small changes in the input data (10,11). Such variability in response to minor changes can signal an instability in the algorithm that could lead to misclassification and problems with generalizability to different patients and settings.

Data scientists have developed strategies to quantify and mitigate the susceptibility of DL models to changes in their output in response to changes in the input using *adversarial images*. Adversarial images are manipulated images which undergo small pixel-level perturbations specifically designed to test the stability of the DL models (12-15). Pixel-level changes of adversarial images are often imperceptible to humans but can cause important differences in the model output (16-18). DL models which show stability in output when faced with adversarial images are likely the most robust and safe for clinical implementation. Previous work concerning adversarial images on DL models has largely focused on non-medical images, and the vulnerability of medical DL models is relatively unknown (18,19). Although techniques to defend against adversarial images have been proposed, the effectiveness of these methods on medical DL models is unclear.

Accordingly, we sought to test the effect of adversarial images on DL algorithms trained on three common oncologic imaging modalities. We established the performance of the DL

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

models and then tested model output stability in response to adversarial images with different degrees of pixel-level manipulation. We then tested the utility of techniques to defend the DL models against adversarial images. This research has direct application to the use of DL image interpretation algorithms, as it provides quantitative testing of their vulnerability to small input variations and determines if there are strategies to reduce this weakness.

## Methods

### *Datasets*

We examined the behavior of DL algorithm outputs in response to adversarial images across three medical imaging modalities commonly used in oncology—CT, mammography, and MRI. For each imaging modality a separate DL classification model was trained to identify the presence or absence of malignancy when given a diagnostic 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 (20). 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) (21). 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

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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 (22). 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 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 (23). The CIFAR-10 datasets includes 60,000 color images of ten non-medical objects (24).

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

*Models*

For all DL classification models, we used a pre-trained convolutional neural network with the VGG16 architecture (25). Models were fine-tuned in Keras using Stochastic Gradient Descent. Details regarding model architecture and hyperparameter selection for DL model training are provided (Table S2, Table S3).

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

*Adversarial Image Generation*

Three commonly employed first-order adversarial image generation methods—Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), and Projected Gradient Descent (PGD)—were used to create adversarial images on the medical and non-medical image datasets (Figure 1). Each method aims to maximize the DL model's classification error while minimizing the difference between the adversarial image and original image. All the adversarial image generation methods are bounded under a predefined perturbation size ε, which represents the maximum change to pixel values of an image. Vulnerability to adversarial images was assessed by comparing changes in model performance compared to baseline (without any adversarial images) under various perturbation sizes.

The single step FGSM attack perturbs the original example by a fixed amount along the direction (sign) of the gradient of adversarial loss (15).

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

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 (12).

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

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 (18).

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

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Additional information regarding adversarial image generation methods is provided (Table S1).

*Susceptibility of DL Models to Adversarial Images*

We investigated the DL model performance using FGSM, PGD, and BIM adversarial image generation methods across different levels of pixel perturbation. We measured relative susceptivity to adversarial images by determining the smallest perturbation ε required for adversarial images to generate a different output. DL models which required larger pixel level perturbations are likely to be more robust and have higher levels of stability suitable for clinical implementation. Conversely, models which change outputs in response to small pixel-level perturbations are inherently unstable and potentially less generalizable across different clinical settings and patient populations.

*Adversarial Training to Improve Model Robustness*

One proposed defense mechanism to prevent negative effects of adversarial images is *adversarial training*, which aims to improve model robustness by integrating adversarial samples into DL model training (18,19). 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 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 (Table S4). We investigated the effectiveness of our iterative adversarial training approach on the DL models trained on medical

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

images. We measured the effectiveness of adversarial training by comparing model accuracy on adversarial samples of varying perturbation size before and after adversarial training.

*Image Level Adversarial Image 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. We hypothesized that images requiring smaller pixel perturbations to change DL model predictions were also the images most likely to be misclassified by the model under normal conditions. 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 perturbation 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 (26). Adversarial images were created using the Adversarial Robustness Toolbox v1.4.1 (27). The code to reproduce the analyses and results is available online at https://github.com/Aneja-Lab-Yale/Aneja-Lab-Public-Adversarial-Imaging.

**Results:**

*Susceptibility of DL Models to Adversarial Images*

Both medical and non-medical DL models were highly susceptible to misclassification of adversarial images resulting in decreases in model accuracy (Figure 2). Medical DL models appeared substantially more vulnerable to adversarial images compared to non-medical DL algorithms. All three medical DL models required smaller pixel-level perturbations to decrease model accuracy compared to non-medical DL models (Figure 2). For example, adversarial

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

images generated using the PGD method (perturbation = 0.002), resulted in DL model accuracy of 26.92% for CT (-48.48% from baseline), 27.68% for mammogram (-48.75% from baseline), and 24.32% for MRI (-61.81% from baseline). In contrast, adversarial images generated using the same methods/parameters did not cause substantial changes in performance for the MNIST (-.05% from baseline) or CIFAR-10 (-4.2% from baseline) trained models (Table 1). For the medical DL models, adversarial images generated using smaller pixel level perturbations ($\varepsilon < 0.004$) resulted in misclassification of a majority of images whereas non-medical DL models required much larger pixel perturbations ($\varepsilon > 0.07$ for MNIST, $\varepsilon > 0.01$ for CIFAR-10) for similar levels of misclassification (Table 1).

*Adversarial Training to Improve Model Robustness*

Adversarial training led to increased robustness of DL models when classifying adversarial images for both medical and non-medical images (Figure 3). Compared to baseline trained models, adversarial trained DL models caused absolute accuracy of the model on adversarial images to increase by 42.9% for CT (67.72% vs 26.92%), 35.7% for mammogram (63.39% vs 27.68%), and 73.2% for MRI (87.20% vs 24.32%) (Table S5). Despite adversarial training, DL models did not reach baseline accuracy suggesting adversarial training as only a partial solution to improve model robustness. Adversarial training became less effective when attempting to defend against adversarial images that possessed greater pixel perturbations.

*Image Level Adversarial Image Sensitivity and Model Performance*

Using image-level adversarial sensitivity, we were able to identify images most at risk for misclassification by the DL models and improve overall model performance across all diagnostic imaging modalities. Excluding the images in which the smallest pixel perturbations changed DL

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

model outputs increased the absolute accuracy of DL models by 5.9% for CT, 3.7% for mammogram, and 5.2% for MRI (Table 2).

**Discussion:**

As the role of diagnostic imaging increases throughout clinical oncology, deep learning represents a cost-effective tool to supplement human decision-making and aid in image analysis tasks (28-30). However, instability of DL model outputs can limit the performance and generalizability on large-scale medical datasets and hinder clinical utility. Evaluating a proposed DL model's susceptibility to adversarial images represents a way to identify the most robust DL models versus those at risk for erratic performance. In this study, we found that DL models trained on medical images were particularly unstable to pixel level changes from adversarial images resulting in significant decreases in expected performance. Moreover, we found diagnostic images within oncology to be more vulnerable such misclassification compared to DL models trained on non-medical images. Specifically, compared to non-medical images, all three diagnostic imaging modalities required substantially smaller pixel-level changes to reduce model performance. Furthermore, we found that adversarial training methods commonly used on non-medical imaging datasets are effective at improving DL model stability to such pixel-level changes. Finally, we showed that identifying images most susceptible to adversarial image attacks maybe helpful in improving overall robustness of DL models on medical images.

Several recent works have found that state-of-the-art DL architectures perform poorly on medical imaging analysis tasks when classifying adversarial images (14,31-35). Our work extends the findings of previous studies by evaluating performance across three common oncologic imaging modalities used for cancer detection. Additionally, we found that CT, mammography, and MRI images exhibit substantial vulnerability to adversarial images even

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

with small pixel-level perturbations (< 0.004). We also show that DL models exhibited different levels of sensitivity to adversarial images across different imaging modalities. Furthermore, while most previous studies used only one fixed perturbation size for adversarial image 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 trained on medical images are more vulnerable to misclassifying adversarial images compared to similar DL models trained on non-medical images (14,36). 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 misclassifying adversarial images 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 (37,38). 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).

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 optical coherence tomography (37,39,40). On the other hand, Hirano et al. found that adversarial training generally did not increase model robustness for classifying dermatoscopic images, optical coherence tomography images, and chest X-ray images (41). The difference in effectiveness of adversarial training can be attributed to differences in adversarial training protocols (e.g., single-step vs. iterative approaches). It's important to note that even in studies where adversarial training showed success in improving

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

model robustness, the results were still suboptimal, as the risk of misclassification increases with perturbation strength 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 (18).

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 attacks which altered pixels more substantially. 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 (14,37). 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. This has potentially useful

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

clinical implications as we can improve the robustness of DL models by excluding such 'high risk' images from DL model classification and instead providing them to a trained radiologist for examination.

There are several limitations to our study. 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. Our study employed only first-order adversarial image generation methods rather than higher-order methods, which have been shown to be more resistant against adversarial training (42). While the most commonly used adversarial image generation methods are first-order, there is still need for additional research on how to defend DL models for medical images against higher-order methods. 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 (37,43,44). 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 utilized adversarial images to explore the stability of DL models trained on three common diagnostic imaging modalities used in oncology. Our findings suggest that DL models trained on diagnostic images are vulnerable to pixel level changes which can substantially change expected performance. Specifically, we found vulnerability to adversarial

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

images can be a useful method to identify DL models which are particularly unstable in their classifications. Additionally, we found adversarial image training may improve the stability of DL models trained on diagnostic images. Lastly, we found that image-level adversarial sensitivity is a potential way to identify image samples which may benefit from human classification rather than DL model classification. By shedding light on the stability of DL models to small pixel changes, the findings from this paper can help facilitate the development of more robust and secure medical imaging DL models which can be more safely implemented into clinical practice.

## 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. 2019 24-25 Oct. 2019. p 363-8.
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):1225-32 doi https://doi.org/10.1016/j.bbe.2020.06.001.
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):1391-405 doi 10.1016/j.bbe.2020.08.008.
6. Kann BH, Hicks DF, Payabvash S, Mahajan A, Du J, Gupta V*, 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):1304-11 doi 10.1200/JCO.19.02031.
7. Cao H, Liu H, Song E, Hung C-C, Ma G, Xu X*, et al.* Dual-branch residual network for lung nodule segmentation. Applied Soft Computing **2020**;86:105934 doi https://doi.org/10.1016/j.asoc.2019.105934.
8. Tang Y-X, Tang Y-B, Peng Y, Yan K, Bagheri M, Redd BA*, et al.* Automated abnormality classification of chest radiographs using deep convolutional neural networks. npj Digital Medicine **2020**;3(1):70 doi 10.1038/s41746-020-0273-z.
9. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A*, 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-e97 doi https://doi.org/10.1016/S2589-7500(19)30123-2.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .


10. Shaham U, Yamada Y, Negahban S. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. Neurocomputing **2018**;307:195-204 doi https://doi.org/10.1016/j.neucom.2018.04.027.
11. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I*, et al.* Intriguing properties of neural networks. arXiv 2013. arXiv preprint arXiv:13126199 **2013**.
12. Kurakin A GI, Bengio S. Adversarial examples in the physical world. arXiv:160702533 **2016**.
13. Yuan X, He P, Zhu Q, Li X. Adversarial Examples: Attacks and Defenses for Deep Learning. arXiv e-prints **2017**:arXiv:1712.07107.
14. Ma X, Niu Y, Gu L, Wang Y, Zhao Y, Bailey J*, et al.* Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognition **2020**:107332 doi https://doi.org/10.1016/j.patcog.2020.107332.
15. Goodfellow IJ SJ, Szegedy C. Explaining and Harnessing Adversarial Examples. arXiv:14126572 **2014**.
16. Shu H, Shi R, Zhu H, Chen Z. Adversarial Image Generation and Training for Deep Neural Networks. arXiv e-prints **2020**:arXiv:2006.03243.
17. Tabacof P, Valle E. Exploring the Space of Adversarial Images. arXiv e-prints **2015**:arXiv:1510.05328.
18. Madry A, Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv:170606083 **2017**.
19. Ren K, Zheng T, Qin Z, Liu X. Adversarial Attacks and Defenses in Deep Learning. Engineering **2020**;6(3):346-60 doi https://doi.org/10.1016/j.eng.2019.12.012.
20. Armato III SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP*, 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):915-31.
21. 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 doi 10.1038/sdata.2017.177.
22. Chang E, Joel M, Chang HY, Du J, Khanna O, Omuro A*, et al.* Comparison of Radiomic Feature Aggregation Methods for Patients with Multiple Tumors. medRxiv **2020**:2020.11.04.20226159 doi 10.1101/2020.11.04.20226159.
23. Lecun Y. THE MNIST DATABASE of handwritten digits. http://yannlecuncom/exdb/mnist/.
24. Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. **2009**.
25. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556 **2014**.
26. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C*, et al.* Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:160304467 **2016**.
27. Nicolae M-I, Sinn M, Tran MN, Buesser B, Rawat A, Wistuba M*, et al.* Adversarial Robustness Toolbox v1.0.0. arXiv e-prints **2018**:arXiv:1807.01069.
28. 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.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .


29. Park A, Chute C, Rajpurkar P, Lou J, Ball RL, Shpanskaya K*, et al.* Deep Learning–Assisted Diagnosis of Cerebral Aneurysms Using the HeadXNet Model. JAMA Network Open **2019**;2(6):e195600-e doi 10.1001/jamanetworkopen.2019.5600.
30. Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH*, et al.* Deep learning in medical imaging and radiation therapy. Med Phys **2019**;46(1):e1-e36 doi 10.1002/mp.13264.
31. Paschali M, Conjeti S, Navarro F, Navab N. Generalizability vs. Robustness: Investigating Medical Imaging Networks Using Adversarial Examples. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, editors2018 2018//; Cham. Springer International Publishing. p 493-501.
32. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science **2019**;363(6433):1287-9 doi 10.1126/science.aaw4399.
33. Wetstein SC, González-Gonzalo C, Bortsova G, Liefers B, Dubost F, Katramados I*, et al.* Adversarial Attack Vulnerability of Medical Image Analysis Systems: Unexplored Factors. arXiv e-prints **2020**:arXiv:2006.06356.
34. 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.
35. Yoo TK, Choi JY. Outcomes of Adversarial Attacks on Deep Learning Models for Ophthalmology Imaging Domains. JAMA Ophthalmology **2020**;138(11):1213-5 doi 10.1001/jamaophthalmol.2020.3442.
36. Shafahi A, Najibi M, Ghiasi A, Xu Z, Dickerson J, Studer C*, et al.* Adversarial Training for Free! arXiv e-prints **2019**:arXiv:1904.12843.
37. 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.
38. Li X, Zhu D. Robust Detection of Adversarial Attacks on Medical Images. 2020 3-7 April 2020. p 1154-8.
39. Paul R, Schabath M, Gillies R, Hall L, Goldgof D. Mitigating Adversarial Attacks on Medical Image Understanding Systems. 2020 3-7 April 2020. p 1517-21.
40. Vatian A, Gusarova N, Dobrenko N, Dudorov S, Nigmatullin N, Shalyto A*, et al.* Impact of Adversarial Examples on the Efficiency of Interpretation and Use of Information from High-Tech Medical Images. 2019 8-12 April 2019. p 472-8.
41. Hirano H, Minagi A, Takemoto K. Universal adversarial attacks on deep neural networks for medical image classification. BMC Medical Imaging **2021**;21(1):9 doi 10.1186/s12880-020-00530-y.
42. Li B, Chen C, Wang W, Carin L. Certified Adversarial Robustness with Additive Noise. arXiv e-prints **2018**:arXiv:1809.03113.
43. Chen C, Yuan W, Lu X, Ma L. Spoof Face Detection Via Semi-Supervised Adversarial Training. arXiv e-prints **2020**:arXiv:2005.10999.
44. 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.

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

**Funding:**

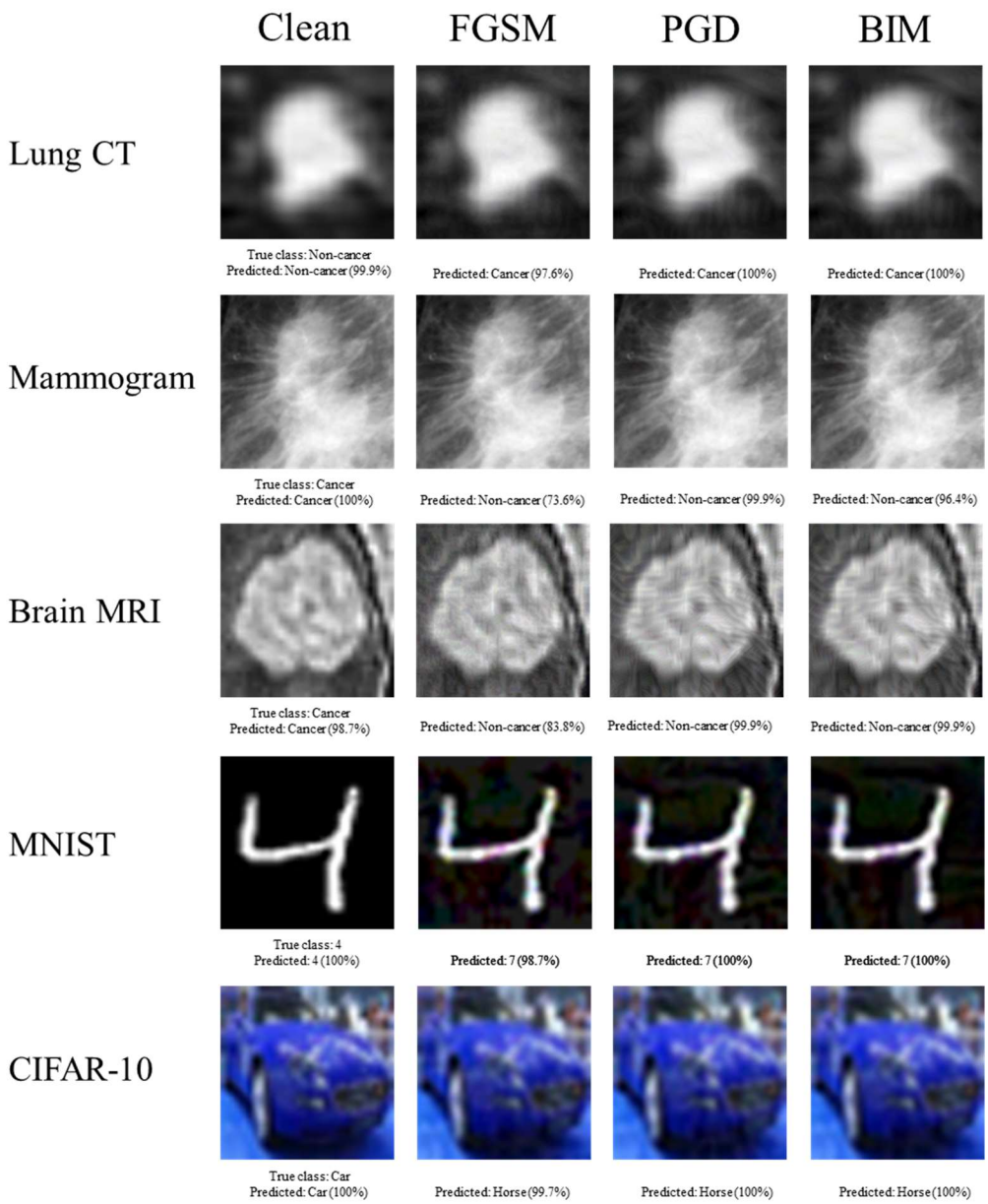
This work was supported by a Career Enhancement Program Grant (PI: Aneja) from the Yale SPORE in Lung Cancer (1P50CA196530); and by a Conquer Cancer Career Development Award (PI: Aneja), supported by Hayden Family Foundation. Any opinions, findings, and conclusions expressed in this material are those of the author(s) and do not necessarily reflect those of the American Society of Clinical Oncology or Conquer Cancer, or Hayden Family Foundation.

**Data Availability**

Data from the study is available from the corresponding author on reasonable request.

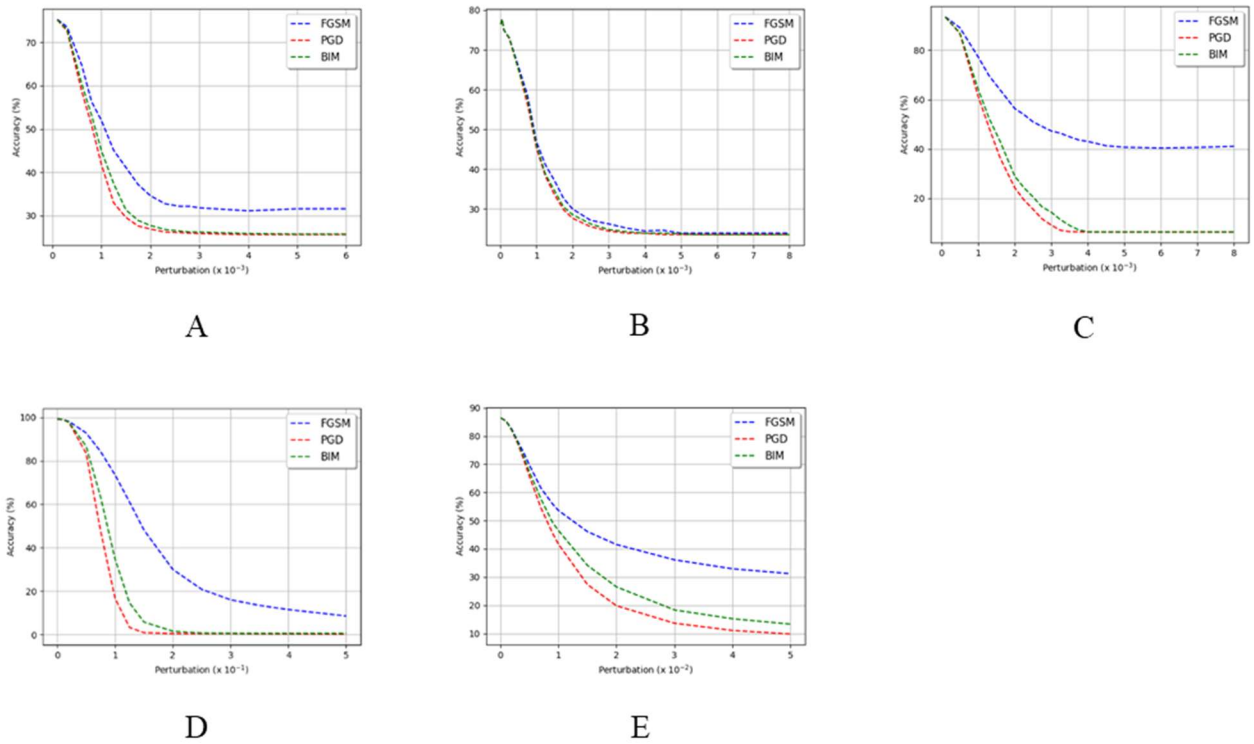
medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

**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.



medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

**Figure 2.** Classification accuracy of VGG16 model on adversarial examples generated by FGSM, BIM, and PGD attacks with increasing $L_\infty$ maximum perturbation size ε. Model performance decreased as ε increased for all datasets: A) lung CT; B) mammography; C) brain MRI; D) MNIST; and E) CIFAR-10. *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).



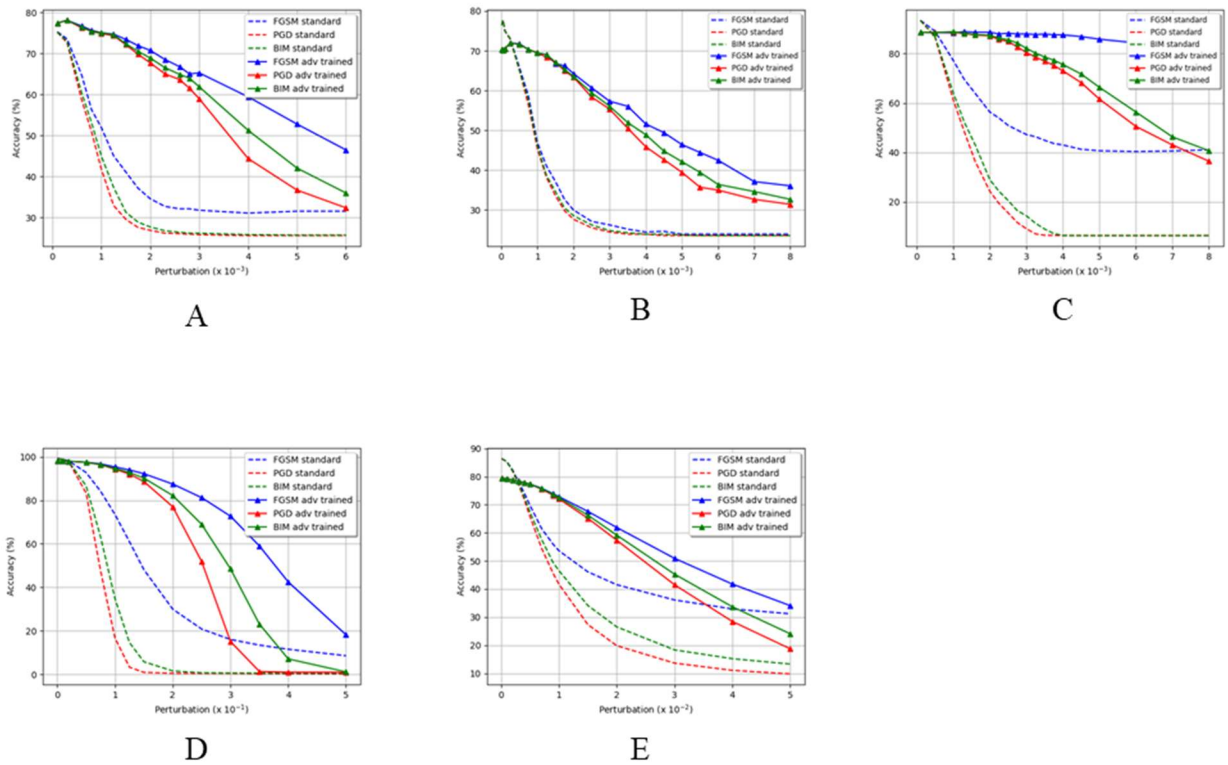
medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

**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 $L_\infty$ 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.

| **Attack** | Perturbation size $\varepsilon$ | CT Accuracy (%) | Mammogram Accuracy (%) | MRI Accuracy (%) | MNIST Accuracy (%) | CIFAR-10 Accuracy (%) |
|---|---|---|---|---|---|---|
| Baseline | | 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 |

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

**Figure 3.** Comparison of model classification accuracy before and after adversarial training on adversarial samples crafted by FGSM, BIM, and PGD with increasing $L_\infty$ maximum perturbation size ε. Adversarial training significantly increased model accuracy for datasets: A) lung CT; B) mammography; C) brain MRI; D) MNIST; and E) CIFAR-10. *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).



medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

**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.

| | 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 |

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

## Supplementary Material

**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.

### 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<br>x = clean input image<br>ε = perturbation size<br>J = loss function<br>y = target label |
| Projected Gradient Descent (PGD) | $x^0 = x,$<br>$x^t = Clip_{x,\epsilon}\{x^{t-1} + \alpha\, sign(\nabla_x J(x^t,y))\}$ | $x^0$ = clean input image<br>$x^i$ = adversarial image at $i^{th}$step<br>ε = maximum perturbation size<br>α = perturbation step size<br>J = loss function<br>y = target label<br>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) | $x^0 = x,$<br>$x^t = \prod_\epsilon\{x^{t-1} + \alpha\, sign(\nabla_x J(x^t,y))\}$ | $x^0$ = clean input image<br>$x^i$ = adversarial image at $i^{th}$step<br>ε = maximum perturbation size<br>α = perturbation step size<br>J = loss function<br>y = target label<br>y = target label<br>∏ { } function projects the adversarial example back onto the ϵ ball ([x-ϵ, x+ϵ]) |

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

**Table S2.** VGG16 model architecture and parameters.

**Model Architecture**

| Layer (type) | Output Shape | Parameter Number |
|---|---|---|
| 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 parameters: 121,676,610 | | |

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

**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 |

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

| 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 |

medRxiv preprint doi: https://doi.org/10.1101/2021.01.17.21249704; this version posted June 8, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

**Table S5.** Effectiveness of Adversarial Training on Classification accuracy (%) of VGG16 model (perturbation size of .002).

| Attack | CT Accuracy (%) | | | Mammogram Accuracy (%) | | | MRI Accuracy (%) | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Original | Adv trained | Difference | Original | Adv trained | Difference | Original | Adv trained | Difference |
| FGSM | 34.62 | 70.75 | 36.13 | 30.00 | 64.29 | 34.29 | 56.36 | 88.64 | 32.28 |
| PGD | 26.92 | 67.72 | 40.80 | 27.68 | 63.39 | 35.71 | 24.32 | 87.20 | 62.88 |
| BIM | 27.74 | 68.88 | 41.14 | 28.57 | 63.57 | 35.00 | 29.02 | 87.42 | 58.40 |