2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)
April 13-16, 2021, Nice, France

DEFENDING AGAINST ADVERSARIAL ATTACKS ON MEDICAL IMAGING AI SYSTEM, CLASSIFICATION OR DETECTION?

Xin Li *Deng Pan* *Dongxiao Zhu*

Department of Computer Science
Wayne State University
5057 Woodward Ave., Detroit, MI 48202

ABSTRACT

Medical imaging AI systems such as disease classification and segmentation are increasingly inspired and transformed from computer vision based AI systems. Although an array of defense techniques have been developed and proved to be effective in computer vision, defending against adversarial attacks on medical images remains largely an uncharted territory due to their unique challenges: 1) label scarcity limits adversarial generalizability; 2) vastly similar and dominant fore- and background make it difficult for learning the discriminating features; and 3) crafted adversarial noises added to a highly standardized medical image can make it a hard sample for model to predict. In this paper, we propose a novel robust medical imaging AI framework based on Semi-Supervised Adversarial Training (SSAT) and Unsupervised Adversarial Detection (UAD), followed by a new measure for assessing systems adversarial risk. We systematically demonstrate the advantages of our robust medical imaging AI system over the existing adversarial defense techniques under diverse real-world settings of adversarial attacks using a benchmark OCT imaging data set.

Index Terms— Adversarial Training, Adversarial Samples, Robust AI System, Medical Image Classification, OCT images

1. INTRODUCTION

Deep neural networks (DNNs) have achieved significant advancement in various tasks of medical imaging [1, 2, 3]. To deploy DNN-based AI systems to support disease diagnosis in those clinical applications, the robustness of DNN models increasingly arises as a great importance. Recent studies [4, 5] have specifically explored the reliability of DNN models in both classification and segmentation tasks of medical imaging. They show that these medical DNN models can be even more vulnerable to adversarial samples compared to natural image models. With human imperceptible perturbations added to clean images, adversarial samples can completely fool the trained DNN model into making incorrect predictions. To generate adversarial samples, various types of methods have been proposed, such as Fast Gradient Sign Method (FGSM) [6] and its variant with stronger attacks Projected Gradient Descent (PGD) [7], and optimization-based attack Carlini & Wagner (C&W) [8]. For segmentation tasks, Ozbulak et al. [5] propose an adaptive segmentation mask attack (ASMA), which produces a crafted mask to fool the trained DNN model. The vulnerability of DNNs to adversarial samples has raised substantial safety concerns on the deployment of medical imaging AI systems at scale.

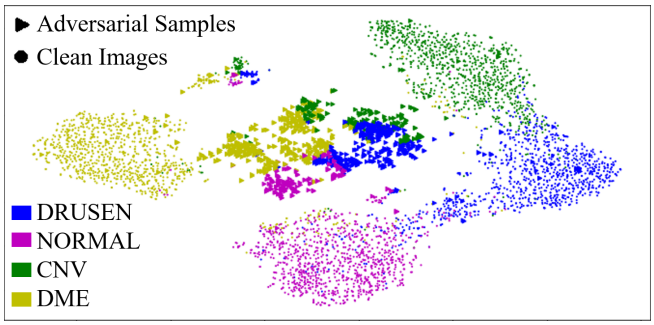

Fig. 1: T-SNE visualization of penultimate layer activations of the model trained on the OCT dataset [9]. The clean images are represented by solid circles with each color represents a true class. The adversarial samples (triangles) are crafted by PGD with a perturbation budget $\epsilon = 0.005$ where each color represents a predicted class. For each class, UAD is capable of filtering out the majority of adversarial samples (center) and SSAT enables the model to correctly predict the rest of adversarial samples (close to clean images).

To defend against these adversarial attacks, an array of strategies have been developed. One major line of those methods is adversarial training (AT) [6], which improves model's adversarial robustness by augmenting the training set with adversarial samples. However, AT for DNNs in medical imaging is problematic as they are primarily designed for natural images and require a large labeled training set [10] whereas medical data sets are usually with a small number of labeled samples. Another line of efficient defense approaches is to learn discriminative features for classifying natural and adversarial samples [11, 12]. With large inter-class separability and intra-class compactness in latent feature space, attacks

with a small perturbation budget are more difficult to succeed. However, medical imaging AI systems can be more susceptible to adversarial attacks [13] since medical images are highly standardized with well-established exposure and quality control, featuring a significant overlap in fore- and backgrounds, which is the so-called "hard sample problem" [14]. Thus a small adversarial perturbation on the entire clean images can significantly distort their distribution in the latent feature space, which can be detrimental to the model performance. As shown in Figure 1, the vast majority of adversarial samples deviate significantly from the distribution of clean samples, implying that they are hard samples for supervised classification. Consequently, the accurate prediction is not attainable yet unsupervised detection remains as a more promising path [13, 15].

Recently several techniques are proposed to improve the effectiveness of defensive methods for medical images. In segmentation tasks, He et al. [16] find that global contexts and global spatial dependencies are effective against adversarial samples, thus they propose a non-local context encoder in the medical image segmentation system to improve adversarial robustness. In classification tasks, Taghanaki et al. [17] use a radial basis mapping kernel to transform and separate features on a manifold to diminish the influence of adversarial perturbation. Based on features extracted from a trained DNN model, Ma et al. [13] attempt to distinguish adversarial samples from clean ones via density estimation in the subspace of deep features learned by a classification model. Although it achieves impressive performance, the so-called 'detection' methods rely on estimating the density of adversarial samples, e.g., via local [18] or Bayesian uncertainty [19] approaches, consequently the effectiveness is limited to the attack methods that are previously seen. To defend against diverse unseen attacks, Li et al. [15] treat the task as an unsupervised abnormal detection problem. They propose a detection strategy which models clean medical images as a unimodal multivariate density and trade adversarial images as outliers. Thus, their approach can effectively reject unknown adversarial attacks.

In order to train the AI system with a small set of labeled images to improve adversarial robustness against unseen and heterogeneous attacks, instead of performing supervised adversarial training, we take a different perspective via unsupervised detection of adversarial samples without the need for estimating density of the adversarial samples. We present a hybrid approach that enhances DNN defensive performance using semi-supervised adversarial training (SSAT) and unsupervised adversarial detection (UAD). Specifically, we utilize both labeled and unlabeled data to generate pseudo-labels for SSAT to improve the robustness of class prediction. To mitigate the distribution distortion of unseen adversarial samples, we employ UAD to screen out those adversarial samples in an effort to facilitate the correct prediction of the rest of adversarial samples by model enhanced with SSAT (Figure 1). Our method is tailor-designed for classifying medical imaging data sets with a limited number of labels and can robustly defend against various unseen attacks.

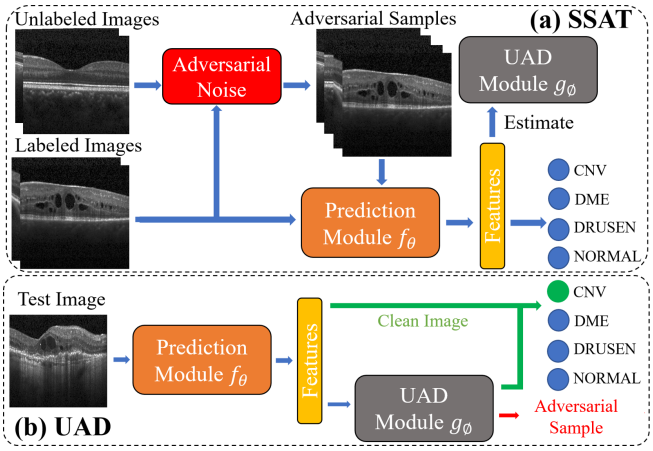

Fig. 2: The proposed robust OCT imaging classification system equipped with SSAT and UAD modules.

2. THE PROPOSED MODEL

The medical image classification problem is to train a prediction function $f_\theta(\cdot)$ by minimizing the loss in mapping a clean image $\boldsymbol{x} \in \mathcal{X}$ to its true label $\boldsymbol{y} \in \mathcal{Y}$. Due to the existence of adversarial samples $\boldsymbol{x}' \in \mathcal{X}'$, it is necessary to have a detection function $g_\phi(\cdot)$ that can distinguish whether an input of f_θ is perturbed by an adversary. Ideally g_ϕ takes inputs from both $\mathcal{X}$ and $\mathcal{X}'$, rejects all $\boldsymbol{x}'$ from $\mathcal{X}'$, then f_θ only takes $\boldsymbol{x}$ from $\mathcal{X}$ to make predictions. A promising solution is to design a UAD function g_ϕ to reject all adversarial samples from $\mathcal{X}'$. However, it is a challenging task since some of adversarial samples are very close to clean images (Figure 1). As such, a supervised prediction function f_θ that is capable of correctly classifying those adversarial samples using a limited labeled training set is also indispensable for maximizing the defense effectiveness.

Figure 2 illustrates our adversarial defense approach. During training (Figure 2a), we learn the robust feature representation via SSAT for both prediction and UAD modules. During inference (Figure 2b), given an unseen test image, the system extracts the feature as the input for UAD module. The test image is rejected if it is detected as an adversarial sample, otherwise, it continues to the loss layer to predict a class label. We describe the technical details of SSAT and UAD modules in the following subsections.

2.1. Semi-supervised Adversarial Training

Adversarial training (AT) [6] is a powerful way to improve the adversarial robustness of a prediction module when the labeled training set is abundant. Recently adversarial samples generated from unlabeled data with pseudo labels

have been shown to be valuable for improving the adversarial robustness [10]. For training the prediction module f_θ with labeled images, we use the supervised AT, i.e., $\mathcal{L}_{\text{sup}}(\theta) = \mathop{\mathbb{E}}_{x\in\mathcal{X}} \sup_{x'\in\mathcal{N}_\epsilon(x)} \text{xent}\left(y, f_\theta\left(x'\right)\right)$, where xent is the cross-entropy loss, $\mathcal{N}_\epsilon(x)$ denotes the neighborhood of a clean image x and $||x - x'||_\infty < \epsilon$. The inner maximization can be approximated by any available attack method, such as PGD and FGSM. For training with unlabeled images, we first find their pseudo labels $\hat{y}(x)$ predicted by f_θ, followed by AT, i.e., minimizing $\mathcal{L}_{\text{unsup}}(\theta) = \mathop{\mathbb{E}}_{x\in\mathcal{X}} \sup_{x'\in\mathcal{N}_\epsilon(x)} \text{xent}\left(\hat{y}(x), f_\theta\left(x'\right)\right)$. We then minimize the loss function to perform SSAT in an effort to enhance model's adversarial robustness: $\mathcal{L}_{\text{semi-sup}}(\theta) = \mathcal{L}_{\text{sup}}(\theta) + \lambda\mathcal{L}_{\text{unsup}}(\theta)$, where λ is a hyper-parameter tunned based on relative abundances of labeled and unlabeled data.

2.2. Unsupervised Adversarial Detection

To filter out adversarial samples $\boldsymbol{x}'$ from being fed into f_θ, we design an UAD module g_ϕ with the goal to exclude the majority of adversarial samples $\boldsymbol{x}' \in \mathcal{X}'$, and simultaneously prevent $\boldsymbol{x} \in \mathcal{X}$ from being erroneously rejected. As shown in Figure 1, the clean images have a different distribution from adversarial samples classified into the same class (color). Inspired by this observation, we estimate a probability density only for clean images as the UAD module and reject images deviating away from this density as adversarial samples. Unlike the detection methods described in [18, 13, 19], our proposed UAD is completely unsupervised that does not need to estimate the adversarial density in whatever way. As a result, it is not limited to detecting the adversarial samples from the known attack types. Specifically, let $\boldsymbol{Z}$ be the latent feature extracted from the penultimate layer of f_θ using $\boldsymbol{x}$ as input and we employ a Gaussian mixture model (GMM) for UAD module g_ϕ. Let $\boldsymbol{\mu}_{ij} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma}_{ij} \in \mathbb{R}^{n\times n}$ represents the mean and covariance matrix of the jth Gaussian component of class i, respectively. For a single class, given all features extracted from clean training samples $\boldsymbol{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$, we can estimate parameters of the GMM using the EM algorithm. The high dimension of $\boldsymbol{Z}$ may cause numerical issues during training. Thus a small non-negative regularization is added to the diagonal of the covariance matrices to alleviate these issues [20].

2.3. Adversarial Risk Evaluation

We propose a new adversarial risk evaluation measure for comparing systems performance in terms of adversarial defense. We assess the risk derived from clean images based on the following intuition: 1) a clean image incurs no risk if it can be correctly classified; 2) a clean image being rejected by the UAD incurs risk r^{uad}_{cln}; and 3) a clean image being accepted by the UAD but misclassified by prediction model incurs risk r^{prd}_{cln}. Assume that for clean images, the number of accepted images that incorrectly predicted is $N^{inc}_{cln}(f,g)$, and the number of clean images being rejected is $N^{rej}_{cln}(g)$, the risk derived from misclassifying (first term) and erroneously rejecting (second term) clean images is calculated as $R_{cln}(f,g) = N^{inc}_{cln}(f,g)\cdot r^{prd}_{cln} + N^{rej}_{cln}(g)\cdot r^{uad}_{cln}$. If only f is used to make predictions (without UAD), the second term is zeroed out. Lets denote $N^{inc}_{cln}(f)$ as the number of clean images being misclassified by f, then the risk is calculated as $R_{cln}(f) = N^{inc}_{cln}(f)\cdot r^{prd}_{cln}$.

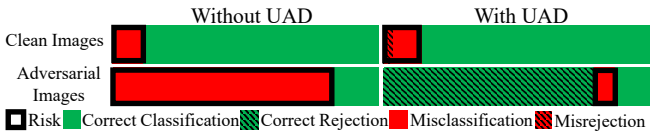

Fig. 3: An illustration of assessing systems adversarial risk. Note the system with UAD on the right exhibits a much lower risk represented by smaller red zones.

Similarly for adversarial samples, we have the following intuition: 1) being correctly rejected by UAD or bypassed but correctly classified incurs no risk; and 2) being erroneously accepted by UAD and misclassified incurs a risk r^{prd}_{adv}. Assume the number adversarial samples in 2) is $N^{inc}_{adv}(f,g)$, the risk derived from adversarial samples is calculated as $R_{adv}(f,g) = N^{inc}_{adv}(f,g)\cdot r^{prd}_{adv}$. When only f is used to make predictions (without UAD), since $N^{inc}_{adv}(f,g) = N^{inc}_{adv}(f)$ and $N^{inc}_{adv}(f)$ is the number of misclassified adversarial samples, the risk is calculated as $R_{adv}(f) = N^{inc}_{adv}(f)\cdot r^{prd}_{adv}$. The total risk, incurred by both clean and adversarial samples, thus can be calculated by $R = R_{cln} + R_{adv}$. The value of different risks ($r^{uad}_{cln}, r^{prd}_{cln}, r^{prd}_{adv}$.) are determined empirically, then we have the risk measures for AI systems with UAD as $R(f,g) = N^{inc}_{cln}(f,g)\cdot r^{prd}_{cln} + N^{rej}_{cln}(g)\cdot r^{uad}_{cln} + N^{inc}_{adv}(f,g)\cdot r^{prd}_{adv}$ and without UAD as $R(f) = N^{inc}_{cln}(f)\cdot r^{prd}_{cln} + N^{inc}_{adv}(f)\cdot r^{prd}_{adv}$.

These evaluation measures are illustrated in Figure 3. Using the above equations, we can assess and compare average adversarial risks between UAD based ($r(f,g) = R(f,g)/N$) and not UAD based ($r(f) = R(f)/N$) defense approaches.

3. EXPERIMENTS AND RESULTS

We use experiments to demonstrate that: 1): The SSAT module can significantly increase model's adversarial robustness without compromising classification performance of clean images. 2) The UAD module can detect and exclude a majority of successful adversarial examples. 3) Our medical imaging AI system (UAD + SSAT) minimizes adversarial risk compared to other existing AI systems. Our implementation and models are available at url:`https://github.com/xinli0928/Defending`.

Dataset and Experiment Settings The experiments are conducted on a public retinal OCT image dataset, originally released in [9]. It contains 84,495 images taken from 4,686 patients with 4 classes: choroidal neovascularization (CNV),

diabetic macular edema (DME), drusen, and normal. To demonstrate the advantages of using unlabeled images for semi-supervised training, we randomly sample 4,000 images for training, 1,000 images for testing and additional 1,000 images as the unlabeled dataset for SSAT. The 4 classes are balanced in each data set. Following the standard prepossessing [6], all images are center-cropped to 224×224 and all pixels are scaled to [0,1]. For AT and SSAT, we augment the data set by generating adversarial samples for each mini-batch using FGSM with a uniformly sampling perturbation from the interval [0.001,0.003]. The number of adversarial and clean images remains $1:1$ within each mini-batch. We use ResNet-18 [21] pre-trained with ImageNet to learn robust feature representations against adversarial attacks. The networks are trained with the SGD optimizer for 10 epochs with a batch size of 64. We set $\lambda = 5$ for SSAT as in [10].

SSAT Performance We evaluate class prediction performance under the most challenging threat: 'white-box' setting [8]. Compared to the benign 'black-box' setting, the adversary possesses complete knowledge of the target model, including architecture and model parameters. We compare our SSAT with three baseline methods in terms of classification accuracy: natural training (NT) with cross-entropy loss, AT with cross-entropy loss [6] and NT with guided complement entropy (GCE) loss [11]. The 1,000 attacks are crafted by 1-step FGSM, 10-step PGD, and C&W.

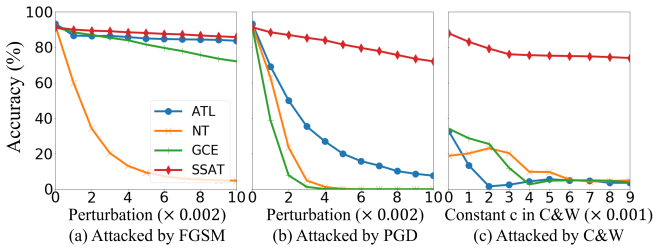

Fig. 4: The supervised prediction accuracy of the four trained models on 1000 adversarial examples crafted by FGSM, PGD, C&W with an increasing budget and constant c.

Figure 4 demonstrates that SSAT markedly outperforms other baselines in all white-box attack settings while maintaining a comparable or better performance on the clean image classification (when the perturbation budget is zero). The NT appears very susceptible to easy attacks generated using FGSM with a very small perturbation budget whereas GCE and AT demonstrate a solid performance against easy attacks but fail under strong attacks such as those generated using PGD and C&W. For AT, label scarcity has significantly limited its adversarial generalizability. For GCE, widening the gap in the manifold between different classes may not work well for medical images due to significant overlaps in both the fore- and backgrounds.

UAD Performance We use features extracted from 4000 clean images in the training set to estimate mixture model density for UAD. Then the 1000 images from test set and its successful adversarial counterparts are used for assessing performance of UAD. As shown in Table 1, UAD is effective in detecting and excluding adversarial samples evident by high area under the Precision-Recall curve (AUPRC) values among all settings. Furthermore, SSAT is more effective than other training strategies, i.e., NT, AT or GCE. Since the classes of clean images and successful attacks are highly imbalanced (136:1000), AUPRC is a suitable metric for evaluation [22]. The average AUPRC value of 0.973 shows the proposed UAD can correctly filter out a vast majority of adversarial samples.

Classes	CNV	DME	DRUSE	NORMAL	Average	# cases
NT	0.897	0.802	0.852	0.859	0.852	885
GCE	0.943	0.902	0.930	0.931	0.927	970
AT	0.890	0.932	0.841	0.903	0.892	580
SSAT	**0.965**	**0.987**	**0.967**	**0.974**	**0.973**	136

Table 1: UAD performance comparison using AUPRC under PGD attack with a perturbation $\epsilon = 0.005$. The last column shows the number of successful adversarial samples.

Method	NT	GCE	AT	SSAT	SSAT*
Adversarial Risk w/o UAD	0.483	0.529	0.324	0.112	0.456
Adversarial Risk w. UAD	**0.446**	**0.367**	**0.317**	**0.108**	**0.225**
Adv Samples Accuracy	11.5%	0.3%	42%	86.4%	17.5%

Table 2: Systems risk under PGD attack with a perturbation $\epsilon = 0.005$. SSTA* is the risk under a stronger PGD attack ($\epsilon = 0.01$).

Comparison of Adversarial Risks Finally, we demonstrate that UAD complementing with SSAT gives rise to the lowest adversarial risk in terms of the new measure proposed in Section 2.3. In Table 2, it is clear that UAD based systems have consistently lower risks compared to those are not, regardless of the training methods used. Note that the reduction of risk is not significant for SSAT against PGD attacks with a smaller budget ($\epsilon = 0.005$). The main reason is that these adversarial samples are relatively weak (highest class prediction accuracy of 86.4% in the last row) that SSAT can successfully predict their labels without the need for UAD. After we double the perturbation budget of PGD attack ($\epsilon = 0.01$), as shown in the last column, the adversarial risk decreases by half (from 0.456 to 0.225) with UAD, highlighting the striking robustness of our system against stronger PGD attacks compared with those without UAD.

4. CONCLUSIONS

We propose to enhance the robustness of medical image AI system via UAD complemented with SSAT. The former is to imbue the system with robustness against unseen adversarial samples whereas the latter mitigate the label scarcity problem in training a robust classifier for correctly predicting bypassed adversarial samples. Through experiments, our system demonstrates a superior performance in adversarial defense to competing techniques.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted using human subject data made available in open access by [9]. Ethical approval was not required as confirmed by the license attached with the open access data.

6. CONFLICTS OF INTEREST

No funding was received for conducting this study. The authors have no relevant financial or non-financial interests to disclose.

7. REFERENCES

[1] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al., "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[2] Xin Li, Rui Cao, and Dongxiao Zhu, "Vispi: Automatic visual perception and interpretation of chest x-rays," *arXiv preprint arXiv:1906.05190*, 2019.

[3] Xin Li, Chengyin Li, and Dongxiao Zhu, "Covid-mobilexpert: On-device covid-19 patient triage and follow-up using chest x-rays," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Los Alamitos, CA, USA, dec 2020, pp. 1063–1067, IEEE Computer Society.

[4] Magdalini Paschali, Sailesh Conjeti, Fernando Navarro, and Nassir Navab, "Generalizability vs. robustness: adversarial examples for medical imaging," *arXiv preprint arXiv:1804.00504*, 2018.

[5] Utku Ozbulak, Arnout Van Messem, and Wesley De Neve, "Impact of adversarial examples on deep learning models for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 300–308.

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[7] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[8] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57.

[9] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[10] Robert Stanforth, Alhussein Fawzi, Pushmeet Kohli, et al., "Are labels required for improving adversarial robustness?," *arXiv preprint arXiv:1905.13725*, 2019.

[11] Hao-Yun Chen, Jhao-Hong Liang, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Da-Cheng Juan, "Improving adversarial robustness via guided complement entropy," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4881–4889.

[12] Xin Li, Xiangrui Li, Deng Pan, and Dongxiao Zhu, "Improving adversarial robustness via probabilistically compact loss with logit constraints," *arXiv preprint arXiv:2012.07688*, 2020.

[13] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu, "Understanding adversarial attacks on deep learning based medical image analysis systems," *arXiv preprint arXiv:1907.10456*, 2019.

[14] Xiangrui Li, Xin Li, Deng Pan, and Dongxiao Zhu, "On the learning property of logistic and softmax losses for deep neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 4739–4746.

[15] Xin Li and Dongxiao Zhu, "Robust detection of adversarial attacks on medical images," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 1154–1158.

[16] Xiang He, Sibei Yang, Guanbin Li, Haofeng Li, Huiyou Chang, and Yizhou Yu, "Non-local context encoder: Robust biomedical image segmentation against adversarial attacks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 8417–8424.

[17] Saeid Asgari Taghanaki, Kumar Abhishek, Shekoofeh Azizi, and Ghassan Hamarneh, "A kernelized manifold mapping to diminish the effect of adversarial perturbations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11340–11349.

[18] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey, "Characterizing adversarial subspaces using local intrinsic dimensionality," *arXiv preprint arXiv:1801.02613*, 2018.

[19] Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner, "Detecting adversarial samples from artifacts," *arXiv preprint arXiv:1703.00410*, 2017.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] Takaya Saito and Marc Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, 2015.