Received: 6 April 2021 | Revised: 15 July 2021 | Accepted: 28 August 2021

DOI: 10.1002/mp.15208

RESEARCH ARTICLE

MEDICAL PHYSICS

Enhancing adversarial defense for medical image analysis systems with pruning and attention mechanism

Lun Chen[1] | **Lu Zhao**[2] | **Calvin Yu-Chian Chen**[1,3,4]

[1] Artificial Intelligence Medical Center, School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, China

[2] Department of Clinical Laboratory, The Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

[3] Department of Medical Research, China Medical University Hospital, Taichung, Taiwan

[4] Department of Bioinformatics and Medical Engineering, Asia University, Taichung, Taiwan

Correspondence
Calvin Yu-Chian Chen, School of Intelligent Systems Engineering, Director of Artificial Intelligence Medical Center, Sun Yat-sen University, Guangzhou 510275, China.
Email: chenyuchian@mail.sysu.edu.cn

Lun Chen and Lu Zhao contribute equally.

Funding information
National Natural Science Foundation of China, Grant/Award Number: 62176272; Guangzhou Science And Technology Fund, Grant/Award Number: 201803010072; Science, Technology & Innovation Commission of Shenzhen Municipality, Grant/Award Number: 20170818165305521; China Medical University Hospital, Grant/Award Numbers: DMR-107-067, DMR-108-132, DMR-110-097

Abstract

Purpose: Deep learning has achieved impressive performance across a variety of tasks, including medical image processing. However, recent research has shown that deep neural networks (DNNs) are susceptible to small adversarial perturbations in the image, which raise safety concerns about the deployment of these systems in clinical settings.

Methods: To improve the defense of the medical imaging system against adversarial examples, we propose a new model-based defense framework for medical image DNNs model equipped with pruning and attention mechanism module based on the analysis of the reason why existing medical image DNNs models are vulnerable to attacks from adversarial examples is that complex biological texture of medical imaging and overparameterized medical image DNNs model.

Results: Three benchmark medical image datasets have verified the effectiveness of our method in improving the robustness of medical image DNNs models. In the chest X-ray datasets, our defending method can even achieve up 77.18% defense rate for projected gradient descent attack and 69.49% defense rate for DeepFool attack. And through ablation experiments on the pruning module and the attention mechanism module, it is verified that the use of pruning and attention mechanism can effectively improve the robustness of the medical image DNNs model.

Conclusions: Compared with the existing model-based defense methods proposed for natural images, our defense method is more suitable for medical images. Our method can be a general strategy to approach the design of more explainable and secure medical deep learning systems, and can be widely used in various medical image tasks to improve the robustness of medical models.

KEYWORDS
adversarial examples, attention mechanism, defense, medical image model, prune

1 | INTRODUCTION

Deep learning has been proven to reach or even exceed the average level of humans in some fields such as natural language processing, speech recognition, or computer vision. Driven by the good performance of deep neural networks (DNNs) on natural images (e.g., CIFAR-10 and ImageNet), DNNs are widely used in medical image processing tasks,[1] such as disease screening and lesion localization.[2–5] However, it has been shown that the DNNs are vulnerable to adversarial perturbations that are deliberately constructed to make it imperceptible and make the model misclassify.[6] Due to the safety-critical nature of some tasks such as medical diagnosis,[7] it is necessary to ensure that the deployed model is robust and well generalized various changes that may occur in the input. Thus, how to improve the robustness of the model on some sensitive tasks has attracted widespread attention. Currently, the defenses against the adversarial attacks are being developed along two main directions: modified training/input and modifying the network.[8] (1) **Input-based defense methods**. This type of defense method directly acts on the input and does not require model modification

and additional training. It is easy to implement and has a certain degree of defensive effect on various attack methods. Typical of these defense methods are: ①Image preprocessing:[9–11] image rotation, filter, contrast, brightness, and noise will all affect the robustness of adversarial examples. Therefore, when resisting adversarial attacks, the above steps can be appropriately added in the image preprocessing link. ②Adversarial training: adversarial training can improve the robustness of neural networks against adversarial examples. It has been a consensus in the related literature. By adding the adversarial images generated from different attack methods to the training image dataset, adversarial training makes the image classification model easier to simulate the distribution of the entire image space. (2) **Model-based defense methods**. This approaches that modify the neural networks for defense against the adversarial attacks, and needs to adjust the model structure or add a detection block to the original model to detect whether the input is against the examples. Compared with the input-based defense method, the training of the new model in the model-based defense method is more complicated and time-consuming, but its defense effect will be better. (1) Autoencoders denoising:[12] autoencoder can filter out the irregular noise superimposed on the original data to a certain extent by learning the characteristics of the dataset itself, so the autoencoder is also usually used for denoising. Therefore, the denoising feature of the autoencoder can be used to eliminate the noise that misleads the model recognition in the adversarial example. (2) Gradient regularization/masking:[13] this method trains differentiable models while penalizing the degree of variation resulting in the output with respect to change in the input. Implying a small adversarial perturbation becomes unlikely to change the output of the trained model drastically. (3) Generative adversarial networks (GANs): Defense-GAN[14] uses a GAN to model the image manifold. Adversarial perturbations are removed by projecting the examples onto the learned manifold. InvGAN[15] involves training an encoder network capable of inverting a pretrained generator network without access to any training data. And it can be used to implement reparameterization white-box attacks on projection-based defense mechanisms.

However, most of these existing defense works on adversarial machine learning research have focused on natural images. Although the adversarial defense methods proposed for natural images can also be used to improve the robustness of medical image models, medical images have unique biological textures and greater detection range that make it difficult for adversarial defense methods designed for natural images to be fully utilized in medical images. A recent work has confirmed that adversarial attacks on medical images can succeed more easily than those on natural images and less perturbation is required to craft a successful attack.[16] Moreover, several healthcare start-ups such as Zebra Medical Vision and Aidoc announced that The United States Food and Drug Administration (FDA) has clearance for their AI medical image systems that suggest that deep-learning-based medical image systems will be potentially applicable for clinical diagnosis in the near future. Therefore, we believe that before deep learning models and medical image techniques become increasingly used in the process of medical diagnostics, decision support, and pharmaceutical approvals, it is necessary to ensure that the medical image systems are safe and robust enough to resist malicious disruption by imperceptibly manipulating images that may cause misdiagnosis.

As for the reasons why medical images are vulnerable to attacks from adversarial examples, it can be explained from the following two perspectives, medical images, and DNNs models. (1) Medical Image Viewpoint: Unlike natural images, first, medical images are acquired in a specific range based on the location of the disease. Therefore, the DNNs model needs to detect a larger and wider image range. Second, the differences between the various categories that medical image needs to learn are smaller; in the suspected disease image, there may be only slight differences between different categories of diseases in the same area. In the face of medical images with rich biological texture information, medical DNNs models must focus on larger image areas and a huge amount of texture information. These reasons make it easier for a well-trained medical DNNs model to generate aggressive adversarial examples. The attacker can easily deceive the medical DNNs model by adding elaborate noise to other biological texture information in the medical image. (2) DNNs Model Viewpoint: When solving tasks related to medical image, the choice of models mostly refers to existing models that have been proven to perform well on natural images. Due to the small amount of medical image data and the severely uneven categories, models that perform well on natural image datasets are often difficult to achieve the same excellent results on medical image datasets. Even if the dataset is increased through data enhancement, it is difficult for a model with a complex structure to get enough information from a small amount of texture information with a low signal-to-noise ratio to fit the model. State-of-the-art DNNs designed for natural image processing can be overparameterized for medical image tasks, resulting in a sharp loss landscape and high vulnerability to adversarial attacks. Therefore, in this paper, we tried to improve the medical image DNNs models by adding the attention mechanism to make the models' focus on the key features among the complex biological textures. Meanwhile, we attempted to adopt the method of model pruning to solve the problem of overparameterization, and the results have shown that it is useful to defense adversarial examples.

Here, we propose a novel model-based defense framework equipped with pruning and attention

mechanism module for medical image DNNs models. And we conducted experiments on three common medical image datasets (chest X-ray,[17] fundoscopy,[18] and dermoscopy[19]). We showed that compared to the existing natural image model-based defense methods and our own ablation experiments, our method could effectively improve the robustness of the medical image DNNs model. Moreover, this model-based defense method only needs to add modules to the original medical model to improve the medical image DNNs model's ability to resist adversarial examples. We believe that with the development and popularization of artificial intelligence-assisted medical diagnosis, our method will be used as a general strategy to improve the safety of medical image DNNs models.

2 | METHOD

2.1 | Datasets, DNNs models, and classification tasks

To prove that our defense method with attention mechanism and pruning designed for medical image in this article is universally applicable to medical images. We validated our defense method on three very successful applications of DNNs in medical image classification: (1) classifying diabetic retinopathy (DR) (a type of eye disease) from retinal fundoscopy; (2) classifying thorax diseases from chest X-rays; and (3) classifying melanoma (a type of skin cancer) from dermoscopic photographs.

2.1.1 | Datasets

For the DR classification task, we use the dataset of the DR detection task on Kaggle.[19] DNNs need to identify DR by the presence of lesions associated with the vascular abnormalities caused by the disease. In this dataset, a clinician has rated DR in each image on a scale of 0–4 (No DR/Mild/Moderate/Severe/Proliferative). Retinal images were provided by EyePACS, a free platform for retinopathy screening.

For the thorax disease classification task, we use National Institutes of Health Chest X-Ray Dataset (NIH chest X-rays).[17] This NIH chest X-ray dataset comprises 112 120 X-ray images with disease labels from 30 805 unique patients. There are 15 classes (14 diseases, and one for "No findings"). Images can be classified as "No findings" or one or more disease classes.

For the melanoma classification task, we use the Skin Cancer MNIST[18]: HAM10000 as our dataset. The dataset consists of 10 015 dermatoscopic images that can serve as a training set for academic machine learning purposes. Cases include a representative collection of all-important diagnostic categories in the realm of pigmented lesions: Actinic keratoses and intraepithelial carcinoma/Bowen's disease, basal cell carcinoma, benign keratosis-like lesions, dermatofibroma, melanoma, melanocytic nevi, and vascular lesions.

2.1.2 | DNNs models

For the above three datasets, we use the ImageNet pretrained ResNet-34[20] as base network and adjust its internal parameters according to different tasks. We add attention layers and pruning to the base network separately or at the same time. The networks are trained for 300–500 epochs using a stochastic gradient descent (SGD) optimizer with an initial learning rate 10^{-4} and momentum 0.9. All images are center-cropped to the size $224 \times 224 \times 3$. Simple data augmentations, including random rotations, width/height shift, and horizontal flip, are used. When the training is completed, the networks are fixed in subsequent adversarial experiments.

2.2 | Attack method

Adversarial examples are images crafted to purposely fool machine learning models, whereas the added perturbations are imperceptible to human eyes.

Given a trained model F, an original input X with output label Y, we generate an adversarial example X by solving a box-constrained optimization problem $X = \max_r L(F, X, Y)$ subject to $F(X) = Y, F(X) = Y, Y \neq Y$. Such an optimization minimizes the added perturbation, say r ($i.e. \hat{X} = X + r$), while simultaneously fooling the model F. By imposing an additional constraint such as $\|r\| \leq \varepsilon$, we can restrict the perturbation to be small enough to be imperceptible to humans. Image pixels are normalized in the [0,1]. So, we also ensure that introduced perturbations result in valid images by adding the constraint that $X \in [0, 1]$.

Adversarial examples can be divided into white-box attacks, black-box attacks, and physical attacks according to the attack cost. A white box attack requires a complete acquisition of the model, an understanding of the structure of the model and the specific parameters of each layer, complete control of the input of the model, and even bit-level modifications to the input data. Compared with black box attacks and physical attacks, the research on white box attacks is more mature, and various algorithms are emerging one after another. The basic framework of adversarial generation is illustrated in Figure 1. It consists of three parts: (1) building a predictive model that maps medical images to clinical labels such as diagnoses, (2) generating adversarial medical perturbation based on the output of the predictive model, and (3) using adversarial medical examples to defraud

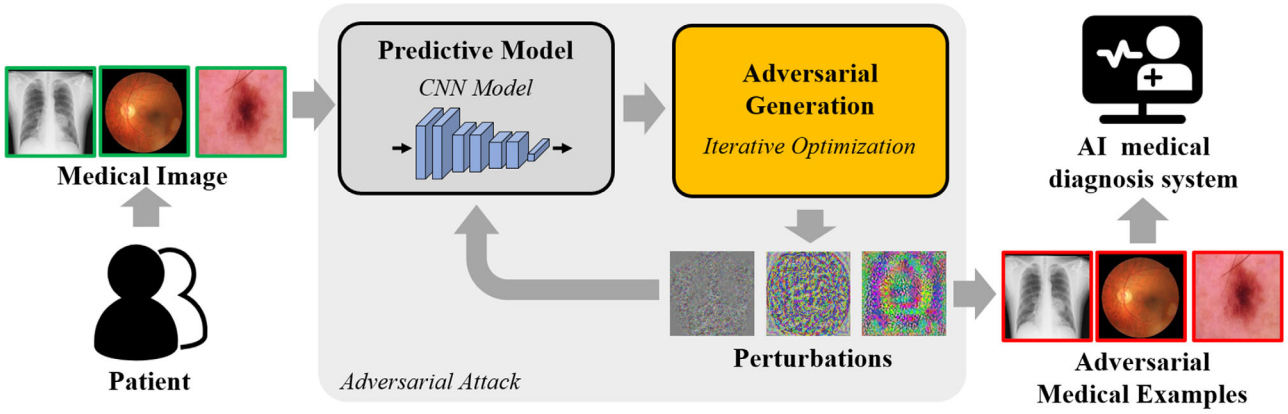

FIGURE 1 Illustration of the proposed framework of generating adversarial examples in electronic medical diagnosis system. Adversarial medical examples are generated by an adversarial perturbation procedure, which are then used to compute a susceptibility medical image over the medical diagnosis system. The adversarial examples are then used to cheat AI medical diagnosis system that will cause high damage if not accurately recorded or measured

the AI medical diagnosis system. The different attack methods are mainly different in the second step. And then we will introduce five representative attacks algorithms below.

Fast gradient sign method (FGSM). FGSM[21] is developed to efficiently compute an adversarial perturbation for a given image by directly increasing the loss of the model. *J*is the loss function. The FGSM attack expression is as follows:

$$\hat{X} = X + \varepsilon \cdot sign(\nabla_X J(X, Y)).$$

Projected gradient descent (PGD). PGD[22] is an iterative attack, which can be regarded as an iterative version of FGSM—K-FGSM (K represents the number of iterations). The general idea is that FGSM only does one iteration and takes a big step. And PGD is to do multiple iterations. The disturbance will be cliped to the specified range in a small step of each iteration. The attack effect of PGD is better than that of FGSM, because when faced with a nonlinear model, just do one iteration, the direction is not necessarily completely correct. *S* is a set of allowed perturbations that formalizes the manipulative power of the adversary The PGD attack expression is as follows:

$$X_{t+1} = \prod_{X+S}(X_t + \alpha \cdot sign(\nabla_X J(X_t, Y))).$$

DeepFool. DeepFool,[23] a kind of attack method based on hyperplane classification, is a simple and accurate method to fool DNNs. As we all know, in the binary classification problem, the hyperplane is the basis for classification. Then to change the classification of a sample *x*, the smallest disturbance is to move *x* to the hyperplane. The distance from this sample to the hyperplane is where the cost is the least. The problem of multiclassification is similar. In this way, we can know the extent to which the adversarial attack is just right. The DeepFool attack simple expression is as follows:

$$r_i \leftarrow -\frac{F(X)}{\|\nabla F(X)\|_2^2}\nabla F(X).$$

Jacobian-based saliency matrix attack (JSMA). JSMA[24] is a typical L0-norm white box targeted attack algorithm, which seeks to minimize the number of pixels modified. The characteristic of JSMA is that it introduces the concept of Saliency Map during the attack to characterize the influence of input features on the prediction results.

C&W attack (CW). CW[25] is an optimization-based attack. The algorithm treats the adversarial example as a variable. If the attack is to be successful, two conditions must be met: (1) The gap between the adversarial example and the corresponding clean sample should be as small as possible. (2) Adversarial examples should make the model classification wrong, and the higher the probability of the wrong category, the better.

$$r = \frac{1}{2}(\tanh(\omega_n + 1) - X)$$

$$\min_{\omega_n} \|r\| + c \cdot f\left(\frac{1}{2}(\tanh(\omega_n) + 1\right).$$

2.3 | Proposed model-based defense method

In this part, we proposed a defense method with attention mechanism and pruning designed for medical image. The basic block of our proposed defense framework for medical image model equipped with

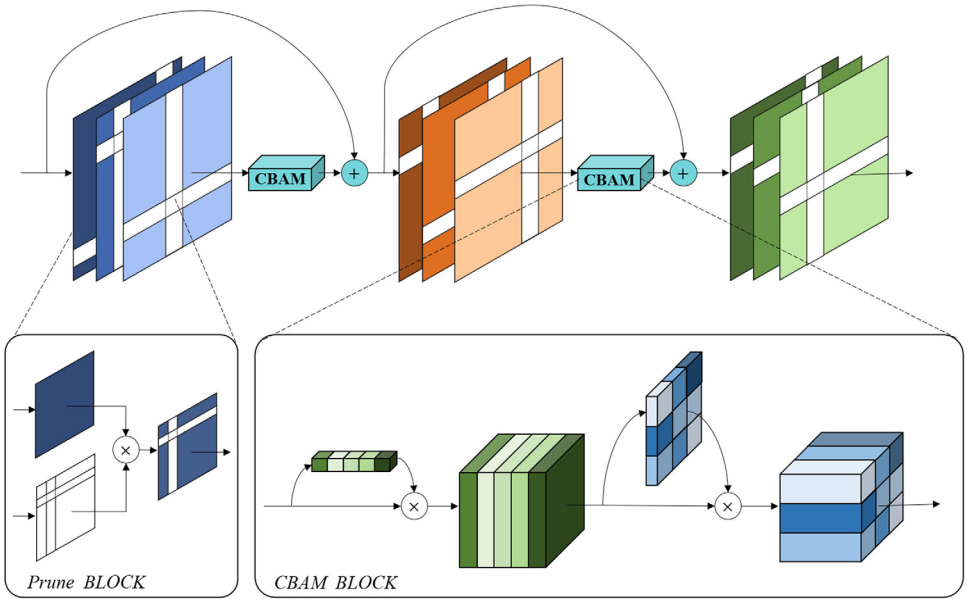

FIGURE 2 The basic block of our proposed defense framework (Prune-CBAM-ResNet34) for medical image model equipped with pruning and attention mechanism module

pruning and attention mechanism module is shown in Figure 2. It comprises two parts: (1) they show pruning module as the white stripes in the convolutional layer in the figure. (2) Attention mechanism module is the convolutional block attention module (CBAM) in the figure. It is an attention mechanism module that combines spatial and channel to improve the robustness of the medical image DNNs model. They show the principles of the two modules below.

2.3.1 | Attention mechanism

For viewpoint of medical image described above point of view, we believe that the robustness of the model can be improved by adding an attention layer to the infrastructure of the original model. The following will reason the robustness of the model added to the attention layer to adversarial examples from two aspects. On the one hand, from the most intuitive point of view, the medical DNNs model trained after adding the attention mechanism[26,27] will have a smaller range of attention when processing medical images. This will make the model pay more attention to the key areas that directly affect disease identification, and will not be attracted by other irrelevant areas. Therefore, even if the attacker adds noise in other unrelated areas to mislead the model to misclassify, the model with the attention mechanism will focus on the information of the key area, reduce the weight of other areas, and resist most of the image from the noncritical area interference. On the other hand, the defense method of applying attention mechanism to medical DNNs model is like other method named random block shuffle (RBS),[11] which bases on adversarial training for natural image by using robust local features to train DNNs models. Adversarial training is to improve the model's ability to capture global features, but normal training is more inclined to improve the model's ability to capture local features. The global structure features are robust against disturbances, but it is difficult to generalize to examples that have not been seen before. Local features have good generalization ability for invisible examples, but poor generalization ability for adversarial examples. Like RBS using RBS to make models focus on local features, adding the attention mechanism will cause medical DNNs models to pay more attention to local features rather than global structure, which is easily attacked by adversarial. But there are also differences between the two methods. RBS destroys the global features of the image, but the method of adding the attention mechanism only emphasizes the local features, but still allows the model to maintain the learning of the overall features of the image. This also ensures the model's generalization ability to unseen images.

CBAM[26] is a simple but effective attention module. Given an intermediate feature map, the attention weight is inferred in sequence along the two dimensions of space and channel, and then the original feature map is multiplied to adjust the feature adaptively. The channel attention module focuses on what features the model should learn, and the spatial attention module focuses on what features should be learned, and adopts two methods of average pooling and maximum pooling to use different information. It can be integrated into all current network architectures as a plug-and-play module.

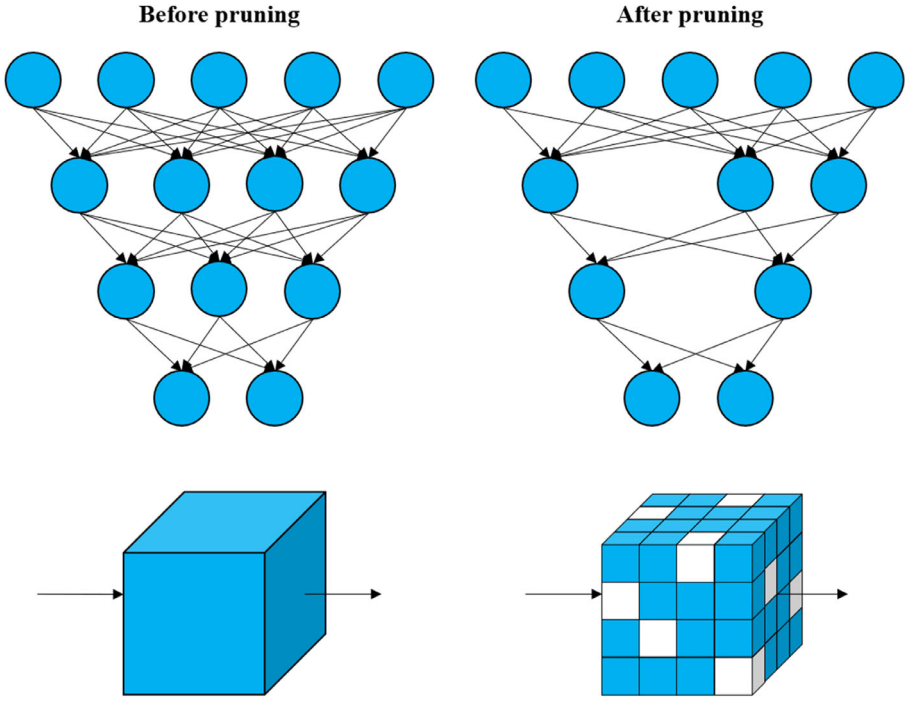

FIGURE 3 Unstructured pruning block in our proposed defense framework (Prune-CBAM-ResNet34). The model before pruning is shown in left and pruned model is shown in right

2.3.2 | Model simplification using prune

To reduce model size and facilitate implementing DNNs for customer applications, the DNN pruning method that reduces the number of weights while preserving the accuracy of the compressed DNNs models is proposed. The pruning process starts from learning the connectivity through normal network training, followed by pruning the connections whose weights are below a given threshold. After making it a sparser network, the DNNs are retrained to finalize weights of the remaining connections. This pruning-and-retraining process is performed iteratively until the network is pruned to the largest extent without accuracy loss. Existing research[28–30] has shown that pruning the model can improve the robustness of the model on natural imaging tasks. According to our analysis of the reasons why the medical image DNNs model is vulnerable to attacks, we believe that the method of using pruning to enhance the robustness of the natural imaging model is also applicable to the medical image model. Even theoretically, the method of using pruning to improve the robustness of the model will be more suitable for medical image DNNs models. By pruning the model, the phenomenon of excessive parameterization of the medical image DNNs model can be alleviated. The amount of training parameters is reduced, the model is simplified, and the overfitting phenomenon caused by too little data of the medical image model is also prevented. Therefore, the distance between the correct example and the adversarial example and the decision boundary in the input space will increase, which increases the difficulty of adversarial attacks, thereby enhancing the robustness of the model. At the same time, the method of pruning the medical image DNNs model also solves the redundancy caused by the use of complex and advanced models designed for natural images on medical images, making the medical DNNs model pay more attention to the low latitude of diagnostic information of the input image.

As shown in Figure 3, we used an unstructured pruning model,[31] which is a method of pruning network connections in a way that preserves the accuracy of the original network. The weight of the neuron during training is used as the criterion for judging whether the neuron is important for pruning and random strategy.

In fact, besides weight-based unstructured pruning, channel-based structural pruning[32,33] is also widely used to compress the model. However, based on the following two reasons, we believe that nonstructural pruning has more advantages in improving the robustness of the model. First of all, the structural pruning of one layer of the model will affect other layers, so it will have a greater impact on the performance of the model. Second, as the pruning of unstructured pruning at each layer is relatively independent, a certain degree of pruning on each layer is more conducive to alleviating model overparameterization.

We verified the effectiveness of the pruning model for adversarial example defense in natural image and

TABLE 1 The attack success rate (%) of various model against four types of attacks crafted on the MNIST datasets

Pruning rate	Parameters (M)	Accuracy	Attack			
			FGSM	CW	PGD	JSMA
0%	11.18	99.21%	100.00%	100.00%	100.00%	100.00%
20%	8.95	99.12%	66.02%	10.50%	93.51%	8.24%
40%	5.37	99.10%	77.76%	16.20%	95.28%	9.51%
60%	2.16	98.62%	58.31%	11.20%	82.86%	4.03%
80%	0.44	97.83%	40.60%	58.90%	62.79%	2.06%

TABLE 2 The attack success rate (%) of various model against four types of attacks crafted on the CIFAR10 datasets

Pruning rate	Parameters (M)	Accuracy	Attack			
			FGSM	CW	PGD	JSMA
0%	11.18	92.80%	100.00%	100.00%	100.00%	100.00%
20%	8.95	93.10%	91.44%	45.50%	96.93%	20.13%
40%	5.37	93.03%	93.01%	45.50%	83.33%	20.49%
60%	2.16	92.71%	92.63%	43.90%	90.01%	20.38%
80%	0.44	93.20%	94.24%	44.20%	96.84%	20.91%

medical image datasets, MNIST, CIFATR10, chest X-ray, fundoscopy, and dermoscopy. The results are shown in Tables 1–5. Different degrees of unstructured pruning were performed, and the adversarial examples generated by the four adversarial methods of FGSM, CW, PGD, and JSMA were selected, and other pruned models were attacked. Our pruning is carried out in all conv modules in the model, and we adopt the L1-norm unstructured pruning method that removes the specified amount of units with the lowest L1-norm. For natural images, we can find that the accuracy of the model has not been greatly affected, and the models that have undergone unstructured pruning have varying degrees of defense effects on the four adversarial examples. For medical images, the accuracy of the model decreases significantly as the pruning rate increases. The model learns more in three complex medical image datasets, so it has fewer redundant parameters. Whether in natural images or medical images, a certain degree of pruning is beneficial to defend against adversarial examples. And the pruning used in this article is all unstructured pruning with a parameter of 0.2 used on the convolutional layer.

TABLE 3 The attack success rate (%) of various model against four types of attacks crafted on the chest X-ray datasets

Pruning rate	Parameters (M)	Accuracy	Attack			
			FGSM	CW	PGD	JSMA
0%	21.28	88.71%	100.00%	100.00%	100.00%	100.00%
20%	17.03	81.74%	81.22%	43.48%	89.03%	65.12%
40%	10.22	76.90%	75.39%	19.70%	93.51%	8.51%
60%	4.07	26.84%	26.08%	20.90%	88.69%	10.28%
80%	0.83	23.62%	17.71%	27.00%	79.73%	2.90%

TABLE 4 The attack success rate (%) of various model against four types of attacks crafted on the fundoscopy datasets

Pruning rate	Parameters (M)	Accuracy	Attack			
			FGSM	CW	PGD	JSMA
0%	21.28	97.71%	100.00%	100.00%	100.00%	100.00%
20%	17.03	97.49%	82.35%	76.30%	67.86%	85.06%
40%	10.22	94.87%	51.54%	40.50%	97.87%	61.48%
60%	4.07	10.31%	26.20%	41.10%	98.35%	11.57%
80%	0.83	10.04%	15.41%	30.30%	90.01%	6.16%

TABLE 5 The attack success rate (%) of various model against four types of attacks crafted on the dermoscopy datasets

Pruning rate	Parameters (M)	Accuracy	Attack			
			FGSM	CW	PGD	JSMA
0%	21.28	86.12%	100.00%	100.00%	100.00%	100.00%
20%	17.03	82.38%	79.03%	52.90%	97.22%	50.04%
40%	10.22	13.97%	47.71%	34.40%	82.82%	41.59%
60%	4.07	9.55%	51.88%	47.80%	89.86%	56.22%
80%	0.83	9.13%	55.41%	32.90%	96.84%	39.68%

3 | RESULTS AND DISCUSSIONS

3.1 | Attack results and robustness of models

To verify that adding an attention layer and pruning the network is beneficial to improve the robustness of the medical image model. We have trained four models for each of the above three aspects of medical image tasks. Respectively, (1) ResNet34, (2) Prune-ResNet34 (20% pruning rate), (3) SelfAttn-ResNet34, (4) BAM-ResNet34, (5) CBAM-ResNet34, and (6) Prune-CBAM-ResNet34. The above six models are ablation experiments designed in this paper, and the attention module and pruning module are added separately and at the same time to verify the effect of the two on improving the robustness of the medical model. We also apply the model-based defense method (kWTA,[13] Auto Encoder,[12] InvGAN[15]) that shows excellent robustness in natural images to the above three datasets, and test its defense effect to compare with the method proposed in this article. The nine models are effectively and fully trained, so that they can achieve effective classification in each task.

Then perform five types of attacks against examples on the well-trained ResNet34, which are FGSM, CW, PGD, DeepFool, and JSMA. And record the adversarial examples that successfully attacked the model and correct labels. Use the adversarial examples that have successfully attacked ResNet34 to attack the other three networks separately. And evaluate the robustness of the model by observing the attack success rate. The specific settings of the above five types of attacks are as follows: ϵ of FGSM is 0.01; JSMA is adopted with max iterations 2000, theta 0.1 and max perturbations per pixel 7; max iteration of CW attack is set as 10; DeepFool is set as iterations 100 and overshoot 9; step size of PGD attack is 1*2/40.

We report the success rate of five attack methods (FGSM, CW, PGD, DeepFool, and JSMA attacks) on nine defense models (ResNet34, Prune-ResNet34, SelfAttn-ResNet34, BAM-ResNet34, CBAM-ResNet34, Prune-CBAM-ResNet34, kWTA, Auto Encoder, and InvGAN) across the three datasets in Tables 6–8. Transferable attack is the most common attack method in black box attacks, which uses adversarial examples generated by alternative models to attack the target model. The closer the structure of the alternative model and the target model are, the higher the attack success rate will be. Therefore, we choose ResNet34 as an alternative model of Prune-ResNet34, CBAM-ResNet34, and Prune-CBAM-ResNet34 models to test the robustness of the model under stronger adversarial examples. The medical images used to generate adversarial examples will not participate in any model training, and only the adversarial examples that successfully attacked in ResNet34 are selected for migration attacks. Of course, the attack success rate is also related to the correct rate

TABLE 6 The attack success rate (%) of various model against five types of attacks crafted on the chest X-ray datasets. The lower the attack success rate, the better the robustness of the model. The best results are highlighted by *

Chest X-ray	Acc	Attack				
		FGSM	CW	PGD	DeepFool	JSMA
ResNet34	88.71	100.00	100.00	100.00	100.00	100.00
Auto Encoder	87.69	44.46	64.24	47.47	42.77	63.19
kWTA	85.37	39.27	46.81	30.71	40.51	47.02
InvGAN	81.45	27.47	36.27	23.30	30.53	38.53*
Prune-ResNet34	81.74	81.22	43.48	89.03	54.11	65.12
BAM-ResNet34	87.08	70.26	65.89	46.75	67.26	70.56
SelfAttn-Resnet34	89.65	44.21	42.87	37.64	47.23	56.23
CBAM-ResNet34	91.93	27.14*	37.80	25.46	31.25	40.36
Prune-CBAM-ResNet34 (Ours)	86.19	27.88	35.61*	22.82*	30.51*	40.72

TABLE 7 The attack success rate (%) of various model against five types of attacks crafted on the fundoscopy datasets. The lower the attack success rate, the better the robustness of the model. The best results are highlighted by *

Fundoscopy	Acc	Attack				
		FGSM	CW	PGD	DeepFool	JSMA
ResNet34	97.71	100.00	100.00	100.00	100.00	100.00
Auto Encoder	96.15	86.26	68.74	61.46	65.18	62.73
kWTA	93.27	57.76	51.67	52.33	63.88	68.16
InvGAN	89.54	48.73	49.57	37.04	54.92	68.68
Prune-ResNet34	97.49	82.35	76.30	67.86	84.32	85.06
BAM-ResNet34	96.39	80.78	88.43	81.55	70.19	74.73
SelfAttn-Resnet34	99.52	65.66	59.40	46.30	69.96	71.82
CBAM-ResNet34	98.03	46.35	49.44	37.24	55.95	66.08*
Prune-CBAM-ResNet34 (Ours)	97.64	45.39*	48.34*	36.04*	54.83*	66.25

of the target model. We will show the accuracy of each target model to show that they achieve good classification results on corresponding medical image classification task. And use ResNet34 as an alternative model to make effective adversarial examples, which are used to migrate and attack other models. Therefore, the success rate of the ResNet34 line in each attack is 100%.

Chest X-ray. As is demonstrated in Table 6, InvGAN shows better defense effects on medical imaging models than Auto Encoder and kWTA on most attacks; and even achieved the best defensive effect on JSMA attacks in chest X-ray. This is because the model-based defense method using GAN can more easily resist adversarial examples that deviate from the original image distribution. By comparing the defense effects of Prune-ResNet34 and ResNet34 models, it can be known that the addition of the pruning module can indeed improve the defense power of the model. But simple pruning cannot achieve the defensive effect of the existing model-based defense methods (kWTA, Auto Encoder, InvGAN). In contrast, the CBAM-ResNet34 model shows a better defense effect than kWTA, Auto Encoder, Prune-ResNet34, and other attention model such as BAM-ResNet34 and SelfAttn-ResNet34. This shows that the attention module is a very critical module to improve the robustness of the model and CBAM block seems to be the best attention block for model-base defense method. The ablation experiments of the Prune-ResNet34 model and the CBAM-ResNet34 model show that the addition of the pruning module and the attention mechanism module are indeed effective for the robustness of the medical imaging model. Similarly, Prune-CBAM-ResNet34 is also better than other defense methods, such as CBAM-ResNet34 in CW ($p = 0.0093$), PGD ($p = 0.0006$), and DeepFool ($p = 0.0058$) attacks. But we also found that when defending against individual attacks in some datasets, such as FGSM in chest X-ray, the model (CBAM-ResNet34) with only the attention module is more robust than the model (Prune-CBAM-ResNet34) with both pruning and attention modules in very few cases. This shows that the attention module is a very critical module to improve the robustness of the model. And by comparing the model-based defense methods (kWTA, Auto Encoder, InvGAN) that perform well on natural images, the experiments in the table prove that the method (Prune-CBAM-ResNet34) we propose in this article performs better on medical images. This means that the defense methods proposed for natural image DNNs models may not always perform well on medical image DNNs models because of the difference between natural images and medical images. And it can effectively improve the robustness of medical image models.

Fundoscopy. Success rate data of five different attacks in the fundoscopy datasets are higher than chest X-ray datasets in Table 7. The design of the attack parameters applied to the three datasets is the same. The adversarial examples with the same amount of disturbance are easier to attack successfully in the fundoscopy dataset. This is because the key to whether the medical imaging model is easier to be attacked is the training of the medical dataset of the model. Uneven distribution of medical image examples, small amount of data, large focus on the field of view, and multiple and complex biological textures will make the trained medical image model more vulnerable to attack. Like the results of the chest X-ray dataset, the method (Prune-CBAM-ResNet34) we propose shows better defense effect than other models in the four attacks like FGSM ($p = 0.0035$), JSMA ($p = 0.0061$), PGD ($p = 0.0465$), and DeepFool ($p = 0.0267$).

Dermoscopy. Similar to the results of the chest X-ray dataset and fundoscopy datasets in Table 8, the method (Prune-CBAM-ResNet34) we propose shows better defense effect than other models in the four attacks like FGSM ($p = 0.0345$), DeepFool ($p = 0.0411$), PGD ($p = 0.0236$), and JSMA ($p = 0.0357$). But it is worth noting that CBAM-ResNet34 shows better defense effect than Prune-CBAM-ResNet34 under CW attack. We believe that this situation is not common. Only

TABLE 8 The attack success rate (%) of various model against five types of attacks crafted on the dermoscopic datasets. The lower the attack success rate, the better the robustness of the model. The best results are highlighted by*

		Attack				
Dermoscopy	**Acc**	**FGSM**	**CW**	**PGD**	**DeepFool**	**JSMA**
ResNet34	86.12	100.00	100.00	100.00	100.00	100.00
Auto Encoder	86.05	86.81	84.75	83.28	79.01	87.64
kWTA	84.44	66.05	55.74	52.43	57.97	58.68
InvGAN	79.54	66.85	52.75	50.36	57.21	52.10
Prune-ResNet34	82.38	79.03	52.90	97.22	65.22	50.04
BAM-ResNet34	84.62	86.87	73.45	69.01	68.63	72.92
SelfAttn-Resnet34	91.96	77.76	62.49	72.11	62.25	69.80
CBAM-ResNet34	84.57	62.32	51.67*	48.55	53.42	50.36
Prune-CBAM-ResNet34 (Ours)	84.43	62.27*	51.88	47.58*	53.02*	49.05*

when the noncritical network parameters are pruned, pruning did not play a role in alleviating the overparameters of the model.

In general, the model with the pruning module and the CBAM module shows better robustness in adversarial example defense. Other model-based defense methods, such as InvGAN, performed better than the model with the pruning module and the CBAM module in some attacks in some tasks. However, we found that the defense model using GAN is difficult to train and is not suitable for the more complex tasks of general application in medical image detection. Therefore, GAN-based defense methods are difficult to be widely used in medical imaging tasks. Compared with using the CBAM module alone (CBAM-ResNet34), we do find that using the CBAM module and the pruning module at the same time (Prune-CBAM-ResNet34) improves the robustness of the model relatively little. We will explain the role of the pruning module in the next section.

3.2 | Understand the prune and CBAM module to defense

To qualitatively and intuitively show the changes of the medical image model that has improved robustness after adding pruning and attention modules, we exploit the Gradient-weighted Class Activation Mapping (Grad-CAM)[34] technique to find the critical regions in the input image that mostly activate the network output. Grad-CAM uses the gradients of a target class, flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the class. As shown in Figures 4–6, we found that the pruning module does not affect the focus area of the model. Compared with the BAM and self-attention modules, CBAM module has a smaller and more accurate field of attention, which is the intuitive reason for making the model more robust. A smaller field of attention means higher requirements for adversarial examples. For adversarial examples, only when the image adds noise to the high attention area, it is more likely to achieve an effective attack on the model. A smaller area of noise production means that the degree of perturbation of the adversarial examples must be greater, which increases the possibility of the adversarial examples being detected. We also show attention map for CBAM from both spatial attention and channel attention in ResNet34 for three datasets in Figure 7.

To explore the role of the pruning module in improving the robustness of the model, we conducted a statistical analysis of the successfully defended adversarial examples of the pruned module and the CBAM module. It is calculated that the number of adversarial examples that only successfully defended by the CBAM module, the number of adversarial examples that only successfully defended by the pruning module, and the number of examples successfully defended by both modules. As demonstrated in Figure 8, we found that compared to the pruning module, the CBAM module can detect relatively more adversarial examples. Although the adversarial examples detected by the pruning module and the adversarial examples detected by the CBAM module have a large number of repetitions, the pruning module still provides a part of the ability to resist adversarial examples, which the CBAM module cannot provide. This also shows that the pruning module has its role in improving the defense performance of the model.

Also, we do need to make some trade-offs between the robustness and accuracy of the model. Robustness requires the model to pay more attention to the global characteristics of the data, so that the model can learn deeper features that can be generalized. The accuracy requires the model to pay more attention to the local features of the data, so that the model can learn the detailed features to distinguish the pictures to the greatest extent, but this outstanding performance is often limited by the training dataset. In real life, medical datasets are relatively difficult to establish and obtain. A well-trained model on limited data is often difficult to

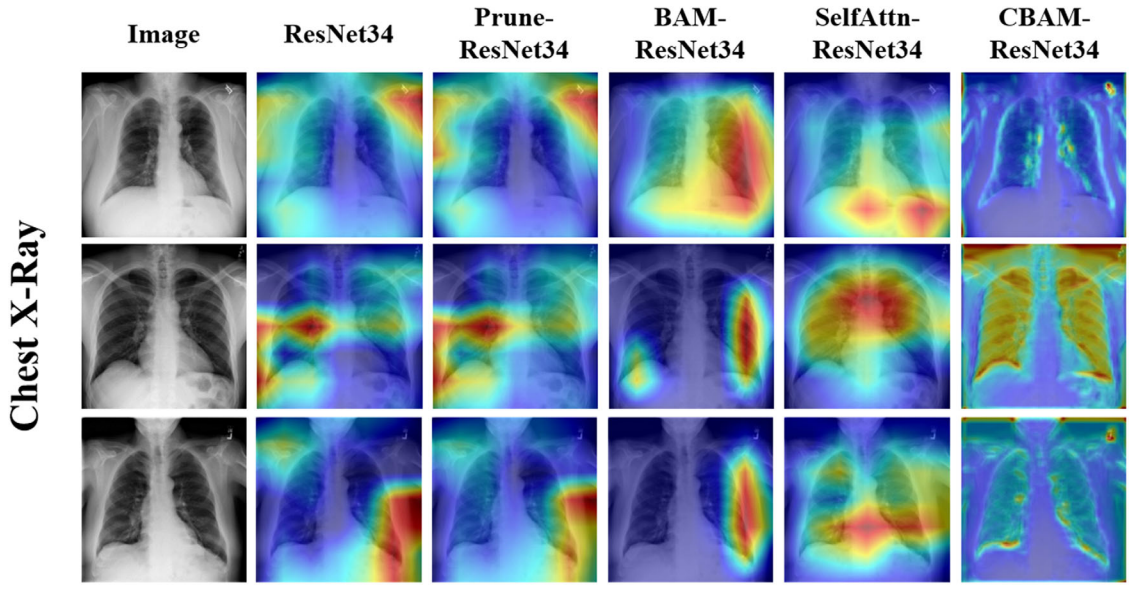

FIGURE 4 The normal images (left one), the saliency maps of the images learned at the ResNet34 (left two), Prune-ResNet34 (left three), BAM-ResNet34 (right three), SelfAttn-ResNet34 (right two), and the CBAM-ResNet34 (right one) in chest X-ray

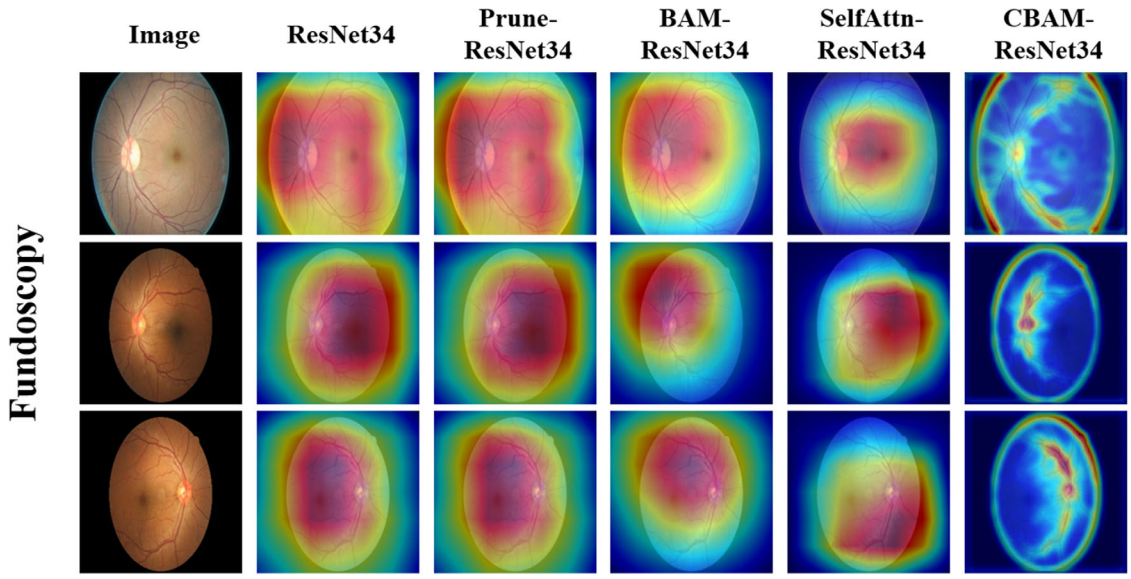

FIGURE 5 The normal images (left one), the saliency maps of the images learned at the ResNet34 (left two), Prune-ResNet34 (left three), BAM-ResNet34 (right three), SelfAttn-ResNet34 (right two), and the CBAM-ResNet34 (right one) in fundoscopy

generalize in most data. Therefore, it is necessary to balance the robustness and accuracy of the model according to the actual application scenario of the model.

4 | CONCLUSION

With the wide application of deep learning networks in various fields, the black box nature of DNNs models brings that the models are easily attacked by adversarial examples. In medical diagnosis, the diagnosis of medical images is mainly performed by doctors who are difficult to be interfered with by adversarial examples. Although, the harm of adversarial examples in the medical field has not yet been obvious, with the rapid development of the field of artificial intelligence, it is believed that in the foreseeable future, artificial intelligence-assisted medical diagnosis will be used on a large scale, whereas security and accuracy are very important features of it. Therefore, it is necessary to propose corresponding adversarial defense methods for the medical DNNs model.

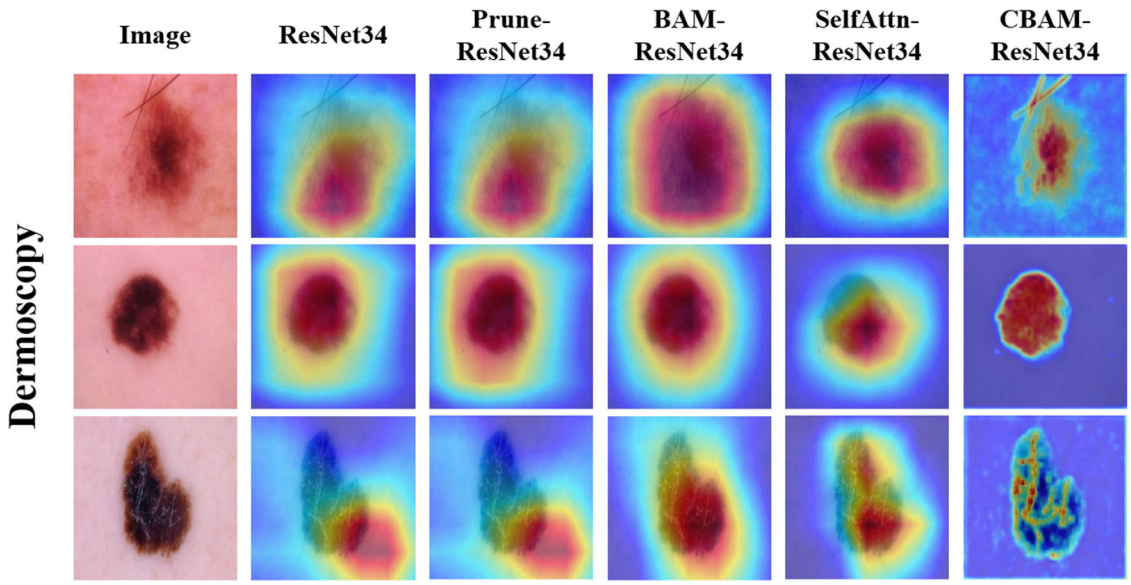

FIGURE 6 The normal images (left one), the saliency maps of the images learned at the ResNet34 (left two), Prune-ResNet34 (left three), BAM-ResNet34 (right three), SelfAttn-ResNet34 (right two), and the CBAM-ResNet34 (right one) in dermoscopy

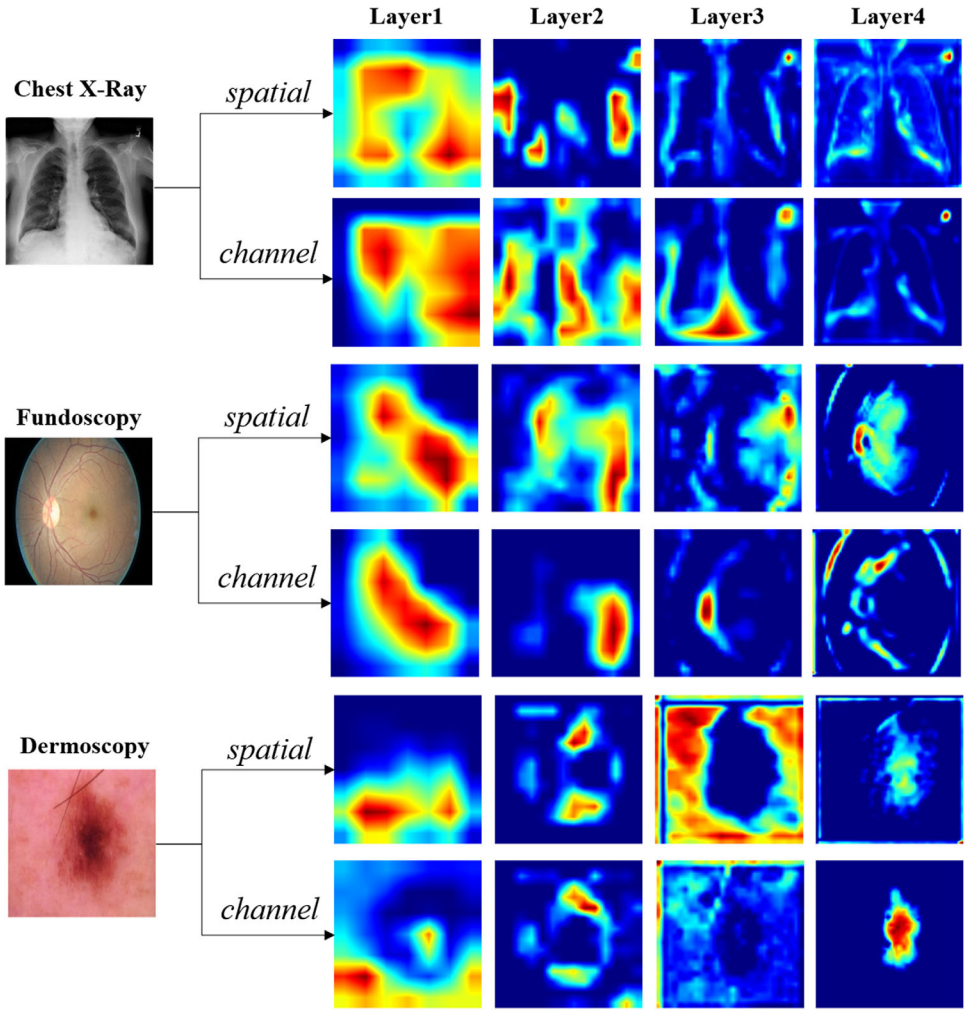

FIGURE 7 Attention map for CBAM from both spatial attention and channel attention in ResNet34 for three datasets

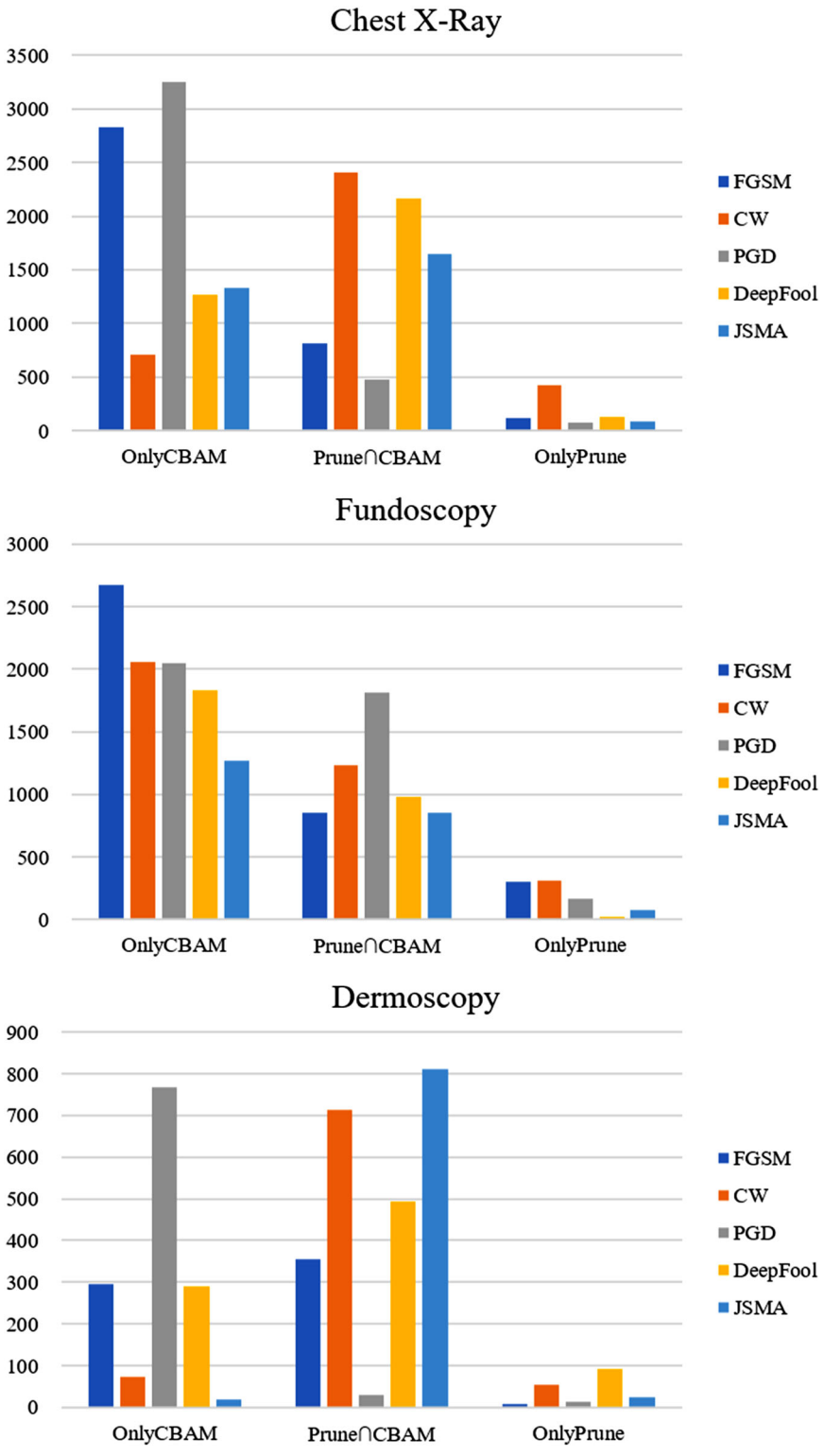

FIGURE 8 The number of adversarial examples that only successfully defended by the CBAM module, pruning module, and both modules on five attacks in three different medical datasets

In this paper, we have investigated the problem of adversarial attacks on deep-learning-based medical image analysis. The main contributions of this study are summarized as follows: (1) We refer to the reasons why the medical image DNNs model is vulnerable to the attack of adversarial examples in previous studies, and propose a model-based defense framework equipped with pruning and attention mechanism module for medical image DNNs models. (2) We conducted a series of experiments with five types of attack on three common medical image datasets. Compared with other model-based defense methods proposed for natural images, our defending method can even achieve up 77.18% defense rate for PGD attack and 69.49%

defense rate for DeepFool attack in chest X-ray dataset. This identified that adding pruning and attention modules to the medical model can effectively improve the robustness of the medical image DNNs model. Moreover, this defense method is better than the defense method designed for natural images, and can more effectively defend against the interference of a wide area outside the pathological area, and improve the robustness of the medical image DNNs model. Thus, we provide a novel general strategy for improving medical image models. And it can be widely used in various medical image tasks to improve the robustness of medical models.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 62176272), Guangzhou Science and Technology Fund (Grant No. 201803010072), Science, Technology & Innovation Commission of Shenzhen Municipality (JCYL 20170818165305521) and China Medical University Hospital (DMR-107-067, DMR-108-132, DMR-110-097). We also acknowledge the start-up funding from SYSU "Hundred Talent Program".

COMPLIANCE WITH ETHICAL STANDARDS

Conflict of interest: The author reports no conflicts of interest in this work.

DATA AVAILABILITY STATEMENT

All data included in this study are available upon request by contact with the corresponding author.

CODE AVAILABILITY

All code for data cleaning and analysis associated with the current submission will be available at https://github.com/alanchancl/Prune-CBAM-ResNet.git. Any updates will also be published on GitHub.

REFERENCES

1. Wang H, Wang L, Lee EH, et al. Decoding COVID-19 pneumonia: comparison of deep learning and radiomics CT image signatures. *Eur J Nucl Med Mol Imaging*. 2020. Published online.
2. Reda I, Ayinde B, Elmogy M, et al. A new CNN-based system for early diagnosis of prostate cancer. In: *Proceedings of the International Symposium on Biomedical Imaging*; 2018. https://doi.org/10.1109/ISBI.2018.8363556
3. Shaffie A, Soliman A, Abu Khalifeh H, et al. Radiomic-based framework for early diagnosis of lung cancer. In: *Proceedings of the International Symposium on Biomedical Imaging*; 2019:1293-1297. https://doi.org/10.1109/ISBI.2019.8759540
4. Daniels Z, Metaxas D. Exploiting visual and report-based information for chest X-ray analysis by jointly learning visual classifiers and topic models. In: *Proceedings of the International Symposium on Biomedical Imaging*; 2019:1270-1274. https://doi.org/10.1109/ISBI.2019.8759548
5. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42:60-88.
6. Barucci A, Neri E. Adversarial radiomics: the rising of potential risks in medical imaging from adversarial learning. *Eur J Nucl Med Mol Imaging*. 2020;47(13):2941-2943.
7. Mustafa A, Khan SH, Hayat M, Goecke R, Shen J, Shao L. Deeply supervised discriminative learning for adversarial defense. *IEEE Trans Pattern Anal Mach Intell*. 2020:1. Published online.
8. Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: a survey. *ArXiv180100553 Cs*. 2018. http://arxiv.org/abs/1801.00553. Published online February 26,. Accessed November 11, 2019.
9. Fu C, Chen H, Ruan N, Jia W. Label smoothing and adversarial robustness. *ArXiv200908233 Cs*. 2020. http://arxiv.org/abs/2009.08233. Published online September 17. Accessed October 7, 2020.
10. Mustafa A, Khan SH, Hayat M, Shen J, Shao L. Image super-resolution as a defense against adversarial attacks. *IEEE Trans Image Process*. 2020;29:1711-1724.
11. Song C, He K, Lin J, Wang L, Hopcroft JE. Robust local features for improving the generalization of adversarial training. *ArXiv190910147 Cs*. 2020. http://arxiv.org/abs/1909.10147. Published online February 2. Accessed July 21, 2020..
12. Kokalj-Filipovic S, Miller R, Chang N, Lau CL. Mitigation of adversarial examples in RF deep classifiers utilizing autoencoder pre-training. *ArXiv190208034 Cs Eess Stat*. 2019. http://arxiv.org/abs/1902.08034. Published online February 16. Accessed December 3, 2020.
13. Xiao C, Zhong P, Zheng C. Enhancing adversarial defense by k-winners-take-all. *ArXiv190510510 Cs Stat*. 2019. http://arxiv.org/abs/1905.10510. Published online October 28,. Accessed July 17, 2020.
14. Samangouei P, Kabkab M, Chellappa R. Defense-GAN: protecting classifiers against adversarial attacks using generative models. Published online 2018:17.
15. Lin W-A, Balaji Y, Samangouei P, Chellappa R. Invert and defend: model-based approximate inversion of generative adversarial networks for secure inference. *ArXiv191110291 Cs Stat*. 2019. http://arxiv.org/abs/1911.10291. Published online November 22. Accessed June 1, 2021.
16. Ma X, Niu Y, Gu L, et al. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognit*. 2021;110:107332.
17. NIH Chest X-rays. Accessed December 3, 2020. https://kaggle.com/nih-chest-xrays/data
18. Skin Cancer MNIST: HAM10000. Accessed December 3, 2020. https://kaggle.com/kmader/skin-cancer-mnist-ham10000
19. Diabetic Retinopathy Detection | Kaggle. Accessed December 3, 2020. https://www.kaggle.com/c/diabetic-retinopathy-detection
20. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *ArXiv151203385 Cs*. 2015. http://arxiv.org/abs/1512.03385. Published online December 10. Accessed December 3, 2020.
21. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. *ArXiv14126572 Cs Stat*. 2015. http://arxiv.org/abs/1412.6572. Published online March 20. Accessed December 3, 2020.
22. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. *ArXiv170606083 Cs Stat*. 2019. http://arxiv.org/abs/1706.06083. Published online September 4. Accessed December 3, 2020.
23. DeepFool: A, Simple and accurate method to fool deep neural networks. IEEE Conference Publication. Accessed December 3, 2020. https://ieeexplore.ieee.org/document/7780651
24. Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: *2016 IEEE European Symposium on Security and Privacy (EuroS P)*; 2016:372-387. https://doi.org/10.1109/EuroSP.2016.36

25. Carlini N, Wagner D. Towards evaluating the robustness of neural networks. *ArXiv160804644 Cs*. 2017. http://arxiv.org/abs/1608.04644. Published online March 22. Accessed December 14, 2020.
26. Woo S, Park J, Lee J-Y, Kweon IS. CBAM: convolutional block attention module. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, eds. *Computer Vision – ECCV 2018*. Lecture Notes in Computer Science. Springer International Publishing; 2018: 3-19.
27. Park J, Woo S, Lee J-Y, Kweon IS. BAM: bottleneck attention module. *ArXiv180706514 Cs*. July 18, 2018. http://arxiv.org/abs/1807.06514. Published online. Accessed November 5, 2020.
28. Cai X, Yi J, Zhang F, Rajasekaran S. Adversarial structured neural network pruning. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. CIKM '19. Association for Computing Machinery; 2019:2433-2436. https://doi.org/10.1145/3357384.3358150
29. Dhillon GS, Azizzadenesheli K, Lipton ZC, et al. Stochastic activation pruning for robust adversarial defense. *ArXiv180301442 Cs Stat*. 2018. http://arxiv.org/abs/1803.01442. Published online March 4. Accessed September 10, 2020.
30. Wang S, Wang X, Ye S, Zhao P, Lin X. Defending DNN adversarial attacks with pruning and logits augmentation. In: *IEEE Global Conference on Signal and Information Processing*; 2018:1144-1148. https://doi.org/10.1109/GlobalSIP.2018.8646578
31. Han S, Pool J, Tran J, Dally WJ. Learning both weights and connections for efficient neural networks. *ArXiv150602626 Cs*. 2015. http://arxiv.org/abs/1506.02626. Published online October 30. Accessed January 21, 2021.
32. Li H, Kadav A, Durdanovic I, Samet H, Graf HP. Pruning filters for efficient ConvNets. *ArXiv160808710 Cs*. 2017. http://arxiv.org/abs/1608.08710. Published online March 10. Accessed June 2, 2021.
33. Luo J-H, Wu J, Lin W, ThiNet: a filter level pruning method for deep neural network compression. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE; 2017:5068-5076. https://doi.org/10.1109/ICCV.2017.541
34. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis*. 2020;128(2):336-359.

AUTHOR BIOGRAPHIES

Lun Chen is a postgraduate student at Artificial Intelligence Medical Center, School of Intelligent Systems Engineering, Sun Yat-sen University. His research interests include computer vision and medical image analysis.

Lu Zhao received the Ph.D. degrees from the Department of Pathogen Biology, Sun Yat-sen University in 2018. She is currently technician of Department of Clinical Laboratory, The Sixth Affiliated Hospital, Sun Yat-sen University.

Prof. Calvin Yu-Chian Chen now is the Director of Intelligent Medical Center and a professor of school of intelligent systems engineering, Sun Yat-sen University. He also had been served as an Advisor or guest Professor in China Medical University, Massachusetts Institute of Technology (MIT), Peking University, University of Pittsburgh, and adjunct professor in Zhejiang University. His research interests include the computer vision, natural language processing and deep learning.

How to cite this article: Chen L, Zhao L, Chen CY-C. Enhancing adversarial defense for medical image analysis systems with pruning and attention mechanism. *Med. Phys.* 2021;48:6198–6212. https://doi.org/10.1002/mp.15208