Pattern Recognition 132 (2022) 108923

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Robust convolutional neural networks against adversarial attacks on medical images

Xiaoshuang Shi[a], Yifan Peng[a,b], Qingyu Chen[a], Tiarnan Keenan[c], Alisa T. Thavikulwat[c], Sungwon Lee[d], Yuxing Tang[d], Emily Y. Chew[c], Ronald M. Summers[d], Zhiyong Lu[a,*]

[a] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health (NIH), Bethesda, MD 20894, USA
[b] Department of Population Health Sciences, Weill Cornell Medicine, New York, NY 10065, USA
[c] Division of Epidemiology and Clinical Applications, National Eye Institute, National Institutes of Health (NIH), Bethesda, MD 20892, USA
[d] Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences Department, National Institutes of Health (NIH) Clinical Center, Bethesda, MD 20892, USA

ARTICLE INFO

Article history:
Received 12 September 2021
Revised 29 June 2022
Accepted 21 July 2022
Available online 22 July 2022

Keywords:
CNNs
Adversarial examples
Sparsity denoising

ABSTRACT

Convolutional neural networks (CNNs) have been widely applied to medical images. However, medical images are vulnerable to adversarial attacks by perturbations that are undetectable to human experts. This poses significant security risks and challenges to CNN-based applications in clinic practice. In this work, we quantify the scale of adversarial perturbation imperceptible to clinical practitioners and investigate the cause of the vulnerability in CNNs. Specifically, we discover that noise (i.e., irrelevant or corrupted discriminative information) in medical images might be a key contributor to performance deterioration of CNNs against adversarial perturbations, as noisy features are learned unconsciously by CNNs in feature representations and magnified by adversarial perturbations. In response, we propose a novel defense method by embedding sparsity denoising operators in CNNs for improved robustness. Tested with various state-of-the-art attacking methods on two distinct medical image modalities, we demonstrate that the proposed method can successfully defend against those unnoticeable adversarial attacks by retaining as much as over 90% of its original performance. We believe our findings are critical for improving and deploying CNN-based medical applications in real-world scenarios.

1. Introduction

Medical images, such as chest x-rays (CXRs) and color fundus photographs (CFPs), are common diagnostic and prognostic imaging modalities in patient care [1,2]. In clinical practice, manual examination of medical images by human specialists, like radiologists and ophthalmologists, is labor-intensive and prone to error. Automated machine learning systems have therefore been proposed for assisting human diagnosis and supporting clinical decision-making [3–5]. Recently, thanks to advances in deep neural network (DNN) technology, such as convolutional neural networks (CNNs) [6–8], a number of deep learning (DL) systems have shown performance at or above the human specialist level on a series of medical image diagnosis tasks [9,10].

However, CNNs are vulnerable to adversarial attacks (Fig. 1) [11,12]. Images can be attacked by adding a small adversarial perturbation to the original images; the perturbation is imperceptible to humans but misleads a standard CNN model into producing incorrect outputs, with a substantial decline in its predictive performance [13,14]. Adversarial attacks can be categorized by the level of information to which the "attacker" has access. In the white-box settings, the attacker has direct access to the target model parameters [15,16]; whereas, in the black-box settings, the attacker has no access to the model parameters. Therefore, black-box attacks are more applicable in many realistic scenarios. CNNs under adversarial attacks would fail to assist and might even mislead human clinicians. Importantly, such a vulnerability also poses severe security risks and represents a barrier to the deployment of automated CNN-based systems in real-world use, especially in the medical domain where accurate diagnostic results are of paramount importance in patient care [17,18].

While a body of literature on adversarial machine learning exists, most previous studies have focused on natural images [19]. Studies on medical images have been limited in scope: some confirmed the susceptibility of medical DL systems to adversarial at-

* Corresponding author.
E-mail addresses: xsshi2013@gmail.com (X. Shi), zhiyong.lu@nih.gov (Z. Lu).

https://doi.org/10.1016/j.patcog.2022.108923

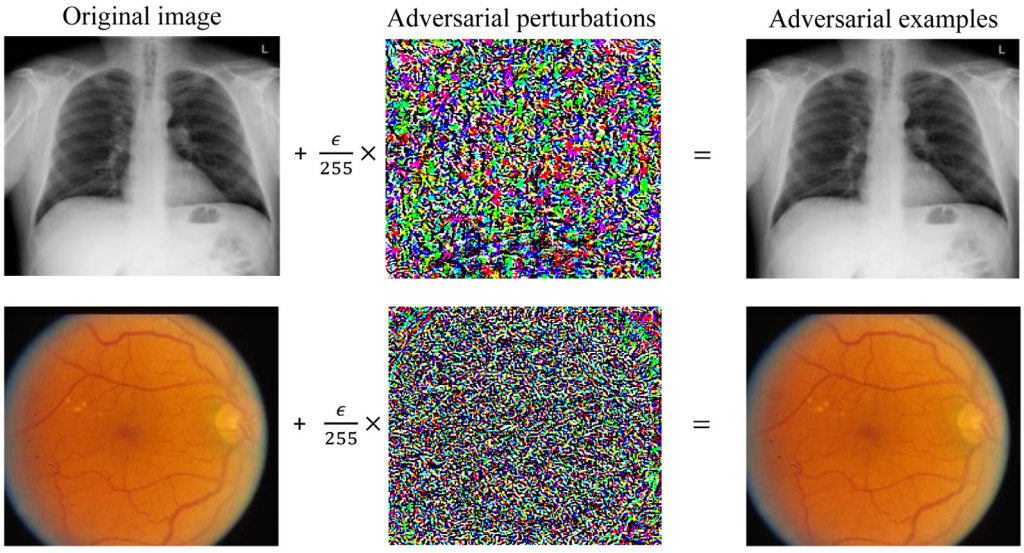


Fig. 1. Adversarial attack examples. The first row shows an original CXR and its adversarial example; the second row displays an original CFP and its adversarial example. Adversarial perturbations are generated by projected gradient descent (PGD) [19] with 20 steps, the maximal pixel perturbation $\epsilon = 1$ and the attack step size $\alpha = \frac{\epsilon}{4}$.

tacks by traditional or image-specific attacking techniques across medical specalities [18,20,21]; others proposed methods to automatically detect adversarial attacks in white-box settings only [15,16].

Despite these efforts, designing a generic detection method that can accurately detect various small adversarial perturbations, especially under black-box attacks, remains inherently challenging for the machine learning research community. Moreover, we are not aware of any previous studies investigating the deep causes of the vulnerability of CNNs to adversarial attacks, towards improved robustness of (medical) CNNs and consequently superior diagnostic performance.

Much of the previous work views adversarial attack susceptibility as related to the high dimensionality of training data [22], where each image usually consists of thousands or even tens of thousands of pixels. Based on this viewpoint, the current state-of-the-art defense strategy, adversarial training (AT) [19,23], first utilizes CNNs to generate adversarial attack examples in each training step and then employs these examples as training data to enhance model robustness. However, AT is reported to significantly decrease CNNs' accuracy on original images, and its performance relies entirely on the adversarial attack examples generated during training [22,24].

From a different perspective, we hypothesize and discover that noise (irrelevant or corrupted discriminative information) learned during CNN training is of considerable importance to CNNs robustness against adversarial attacks. More specifically, images usually contain many noisy features that are irrelevant to human classification. CNNs are likely to learn these noisy features unconsciously during training, resulting in noise in feature representations for decision-making. As such, adversarial perturbations might manipulate the noise in feature representations to degrade model accuracy [25].

To alleviate the effect of noisy features learned during training, we propose a robust CNN framework (Fig. 2), which consists of a novel denoising operator embedded into each convolutional layer to reduce the noise in its outputs, thereby combatting the effect of adversarial perturbations. The denoising operator contains two layers: an inter-sample denoising layer and an intra-sample denoising layer. The former utilizes the entire batch of data to decrease the noise, which might otherwise be mistaken under adversarial attacks as discriminative features. The latter reduces the noise in each medical image, to further lower the noise in feature representations.

Compared to existing methods, the contributions of our work are threefold. First, this work attempts to examine the causes of vulnerability in medical DL-systems and subsequently to propose a new defense method for improving their robustness against adversarial attacks. Second, we recruited both radiologists and retinal specialists to identify and qualify the scale of adversarial perturbation imperceptible to humans. Third, we validate the generalizability of the proposed method on two distinct imaging modalities: CXRs [26] and CFPs [27], which are widely used in radiology and ophthalmology, respectively.

2. Methods

In this section, we first illustrate that reducing the noise in medical images can improve CNNs' robustness against adversarial attacks, and then introduce the proposed sparsity denoising operators to improve the denoising capability of CNNs.

2.1. Noise reduction improves CNNs' robustness against adversarial attacks

Gaussian filtering is a linear smoothing filter to remove high frequency or Gaussian noise, which is widely existing in images. Thus, for simplicity, we employ Gaussian filters to preprocess natural and medical images (where nature images are from one popular natural image dataset, CIFAR-10 [28], and medical images are from CXRs) for noise reduction, and then utilize original and preprocessed images to train models using a popular network, ResNet50 [29], on natural and medical images, respectively. After that, we attack well-trained models by a state-of-the-art adversarial attack method, PGD with 20 steps (PGD-20) [19]. Note that we only select two classes from CIFAR-10, because medical images in CXRs are categorized into two classes. Additionally, we resize natural and medical images to the same size of $224 \times 224 \times 3$ and then train the models on them to achieve almost the same accuracy on normal testing samples.

Fig. 3 a shows testing accuracy of ResNet50 on original and preprocessed images under a white-box attack by PGD-20, and Fig. 3c display their accuracy under a black-box attack by PGD-20. As shown, ResNet50 has more robust performance on preprocessed

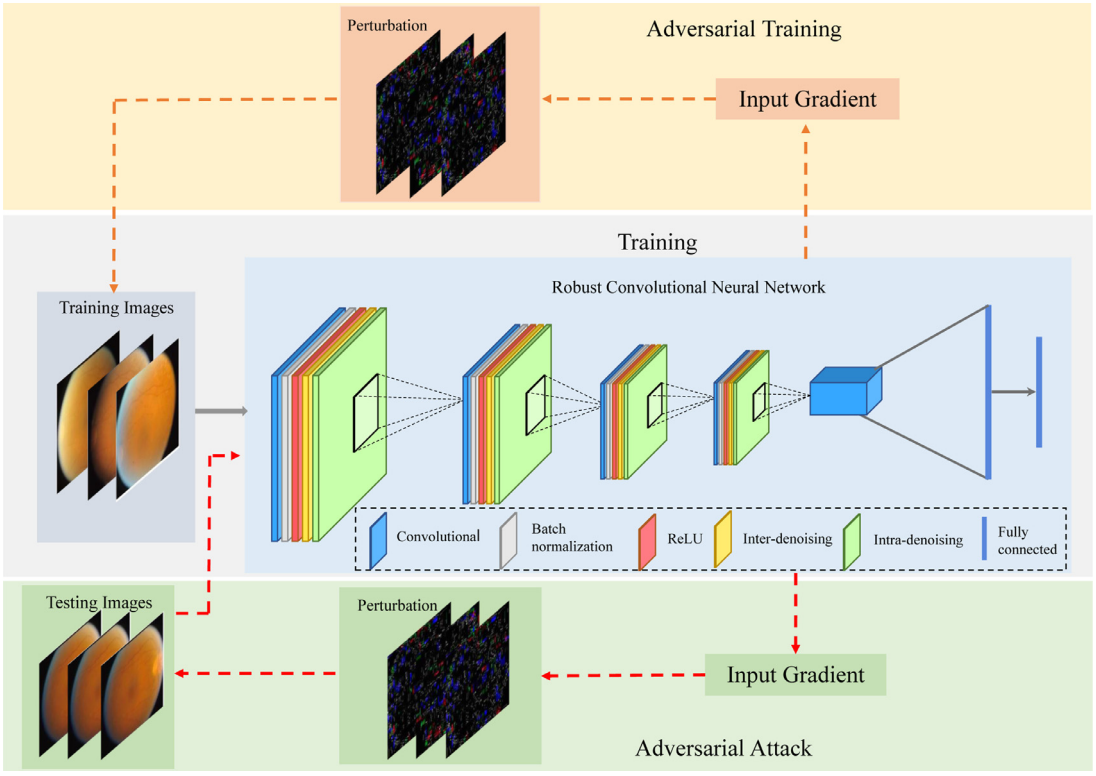


Fig. 2. The frameworks flowchart for adversarial training and testing adversarial attack examples. Inter-denoising and intra-denoising layers follow the ReLU layer. Inter-denoising refers to the inter-sample denoising layer, which leverages the whole batch of data to decrease noise contained in feature representations (Eq. (3)). Intra-denoising denotes the intra-sample denoising layer, which reduces the noise in the image itself (Eq. (5)).

images than on their original images. This might because original images are noisy, leading to a high degree of noise in the learned feature representations, and that the noise is magnified by adversarial perturbations, thereby decreasing model accuracy substantially. By contrast, noisy features are substantially decreased in the preprocessed images, leading to less noise in the learned feature representations, thereby alleviating the effect of adversarial perturbations and increasing model accuracy.

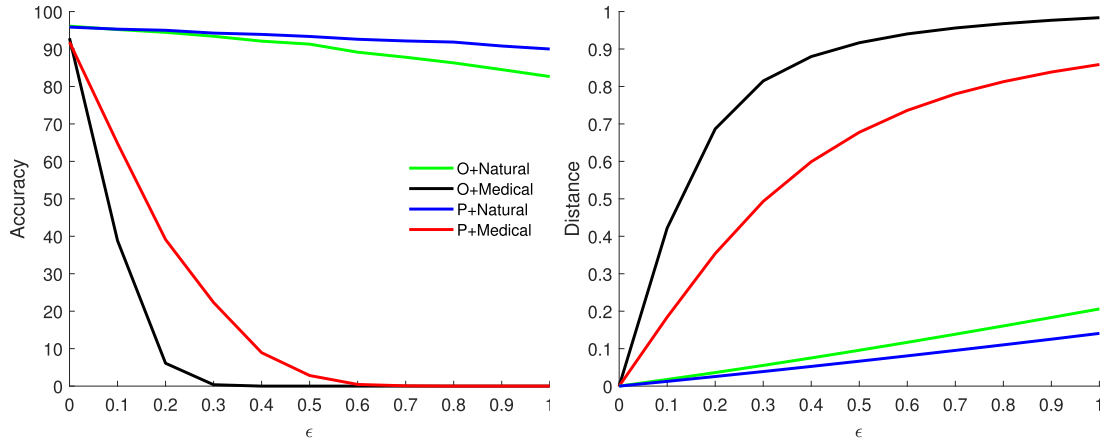
To verify that adversarial perturbations can magnify the noise in feature representations, thereby enlarging the distances between feature representations of the inputs (original and preprocessed images) and their adversarial examples, we present their feature distances in Fig. 3b, where feature representations were learned by ResNet50. In addition, Fig. 3d presents their feature distances between inputs and their transferable adversarial examples. These two figures illustrate that the feature distances between the preprocessed images and their adversarial examples (or transferable adversarial examples) were smaller than those between the original images and their adversarial samples. Therefore, they suggest that reducing noise can improve the robustness of CNNs against adversarial attacks.

Fig. 3 also shows that ResNet50 achieves lower accuracy on adversarial medical images than on adversarial natural images, but with a higher feature distance between the original images and their adversarial examples. Similar findings are also observed for CFPs. Taken together, these results suggest that CNNs are more vulnerable on adversairal medical images, and reducing noise in medical images might be more important compared to natural images.

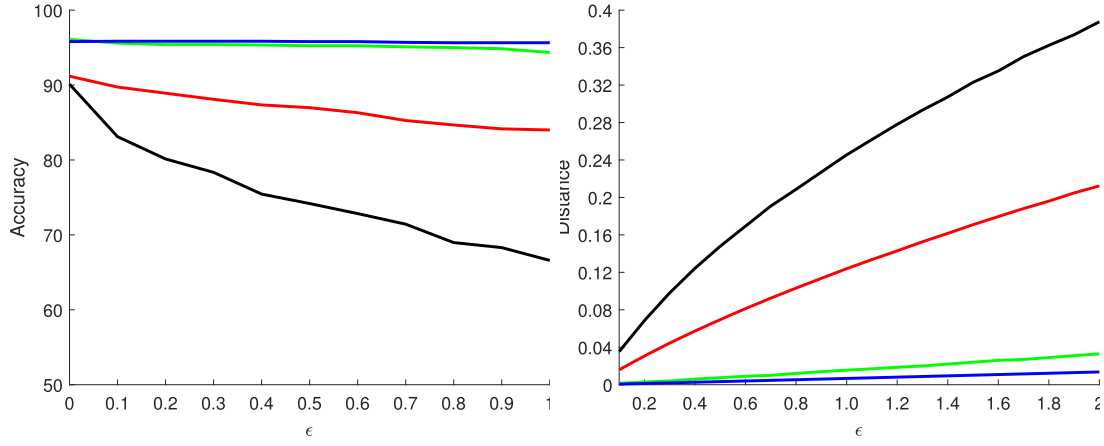
To further demonstrate processed images containing less noise, we present the visualization of original, processed images and their feature maps extracted from ResNet50 under different adversarial perturbations in Figs. 4 and 5, which suggest that the processed chest image is more clear than the original one, thereby enhancing the signifcant pixels in the image. Additionally, Fig. 4 (b) and (d) display that the weight in feature maps of the processed chest image has smaller changes than that in the original image with the change of perturbations, while Fig. 5 suggests the weight in feature maps of both original and processed natural images changes slightly under different adversarial attacks. This might because CNNs under attacks are more vulnerable on medical images than that on natural ones.

2.2. Sparsity denoising

Given a batch of data $\mathbf{X} \in \mathbb{R}^{n\times c\times h\times w}$ and an L-layer convolutional neural network $f_\theta(\cdot)$ with ReLU as its activation function, let $\mathbf{Z}^m \in \mathbb{R}_{0+}^{n\times c^m\times h^m\times w^m}$ $(1 \le m \le M)$ be the output of the m^{th} layer for $\mathbf{X}$ after convolution, batch normalization and ReLU, and $\mathbf{Z}_j^m \in \mathbb{R}_{0+}^{n\times h^m\times w^m}$ denote the j^{th} $(1 \le j \le c^m)$ channel of $\mathbf{Z}^m$, where $\mathbb{R}_{0+}$ denotes the set of positive real numbers, n is the number of samples in one batch, c, h and w are the channel number, height and width of $\mathbf{X}$, respectively, c^m, h^m and w^m denote the channel number, height and width of the output of the m^{th} layer, respectively, and θ represents model parameters. Assume that the noise in $\mathbf{Z}_j^m$ contains small inter-sample and intra-sample noise (The noise that requires the other samples in each batch to remove is defined as the inter-sample noise, and the noise that can be removed by the sample itself is regarded as the intra-sample noise), $\mathbf{D}_j^m$ is a clean output matrix without inter-sample and intra-sample noise, $\tilde{\mathbf{E}}_j^m \in \mathbb{R}_{0^+}^{h^m\times w^m}$ denotes an additive error matrix to represent inter-sample noise, and $\mathbf{e}_j^m \in \mathbb{R}_{0+}^n$ represents an error column vector to represent intra-sample noise, we have $\mathbf{Z}_j^m = \mathbf{D}_j^m + \mathbf{e}_j^m \otimes \mathbf{1}_{h^m} \otimes \mathbf{1}_{w^m} + \mathbf{1}_n \otimes \tilde{\mathbf{E}}_j^m$, where $\mathbf{1}_{h^m} \in \mathbb{R}^{h^m}$, $\mathbf{1}_{w^m} \in \mathbb{R}^{w^m}$ and $\mathbf{1}_n \in \mathbb{R}^n$ represent column vectors with all entries being ones, respectively, and $\otimes$ denotes the outer product. In order to attain the clean output $\mathbf{D}_j^m$, we divide



(a) White-box@Accuracy vs. ϵ (b) White-box@Distance vs. ϵ



(c) Black-box@Accuracy vs. ϵ (d) Black-box@ Distance vs. ϵ

Fig. 3. Testing accuracy and feature distance of ResNet50 on natural and medical images under white-box and black-box attacks by PGD-20. (a) Testing accuracy of ResNet50 under the white-box attack; (b) Feature distances of ResNet50 under the white-box attack between the inputs (original and preprocessed images) and their adversarial examples. (c) Testing accuracy of ResNet50 on transferable adversarial examples; (d) Feature distances of ResNet50 between the inputs (original and preprocessed images) and their transferable adversarial examples. The natural and medical images are from CIFAR10 and CXRs, respectively. Transferable adversarial examples are generated by ResNet101 under the attack PGD-20. Feature distance is the Euclidean distance between original images and their adversarial examples under different perturbations, where features are extracted from the last convolutional layer of ResNet50. ϵ denotes the maximal pixels modified by adversarial perturbations and the attack step size $\alpha = \frac{\epsilon}{4}$. 'O and 'P respectively denote original and preprocessed images obtained by using Gaussian filters to preprocess original training and testing data.

the denoising process into two individual steps, inter-sample and intra-sample denoising.

Firstly, we remove the inter-sample noise $\tilde{\mathbf{E}}_j^m \in \mathbb{R}_{0+}^{h^m \times w^m}$. Let $\tilde{\mathbf{D}}_j^m = \mathbf{D}_j^m + \mathbf{e}_j^m \otimes \mathbf{1}_{h^m} \otimes \mathbf{1}_{w^m}$, and thus $\mathbf{Z}_j^m = \tilde{\mathbf{D}}_j^m + \mathbf{1}_n \otimes \tilde{\mathbf{E}}_j^m$. Due to the usage of ReLU, $\mathbf{Z}_j^m$ is sparsity so that $\tilde{\mathbf{D}}_j^m$ is sparse. Meanwhile, $\tilde{\mathbf{D}}_j^m$ should maximally preserve the most information of $\mathbf{Z}_j^m$ because of the small values in $\tilde{\mathbf{E}}_j^m$. To attain $\tilde{\mathbf{D}}_j^m$, we aim to solve the following model:

$$\min_{\tilde{\mathbf{D}}_j^m} \frac{1}{2} \left\| \mathbf{Z}_j^m - \tilde{\mathbf{D}}_j^m \right\|_F^2 + \left\| \mathbf{1}_n \otimes \Lambda_j \cdot \tilde{\mathbf{D}}_j^m \right\|_1, \tag{1}$$

where $\Lambda_j \in \mathbb{R}_{0+}^{h^m \times w^m}$ is a regularization matrix for removing the noise $\tilde{\mathbf{E}}_j^m$, and $\cdot$ denotes the dot product. Eq. (1) has a closed-form solution and it can be easily calculated as:

$$\tilde{\mathbf{D}}_j^m = \left\| \mathbf{Z}_j^m - \mathbf{1}_n \otimes \Lambda_j \right\|_+. \tag{2}$$

There are many choices for Λ_j, however, it is empirically infeasible to set a constant for Λ_j, because $\mathbf{Z}_j^m$ is changing during training. Additionally, Λ_j aims to remove the inter-sample noise, which means it should depend on different samples to remove their general noise. Hence, we set $\Lambda_j = \frac{\gamma}{n} \mathbf{1}_n^T \mathbf{Z}_j^m$ so that it is changing with $\mathbf{Z}_j^m$, and meanwhile $\mathbf{1}_n^T$ makes Λ_j depend on n samples to represent their general noise, where γ is a positive constant. Substituting $\Lambda_j = \frac{\gamma}{n} \mathbf{1}_n^T \mathbf{Z}_j^m$ into Eq. (2), it becomes:

$$\tilde{\mathbf{D}}_j^m = \left\| \mathbf{Z}_j^m - \frac{\gamma}{n} \mathbf{1}_n \otimes (\mathbf{1}_n^T \mathbf{Z}_j^m) \right\|_+. \tag{3}$$

After obtaining $\tilde{\mathbf{D}}_j^m$, we aim to eliminate the intra-sample noise $\mathbf{e}_j^m$ to attain the clean output $\mathbf{D}_j^m$. Because $\tilde{\mathbf{D}}_j^m$ is sparsity and $\mathbf{e}_j^m$ is with small positive values, $\mathbf{D}_j^m$ should be sparse and maximally maintain the significant information in $\tilde{\mathbf{D}}_j^m$. Similar to Eq. (1), we

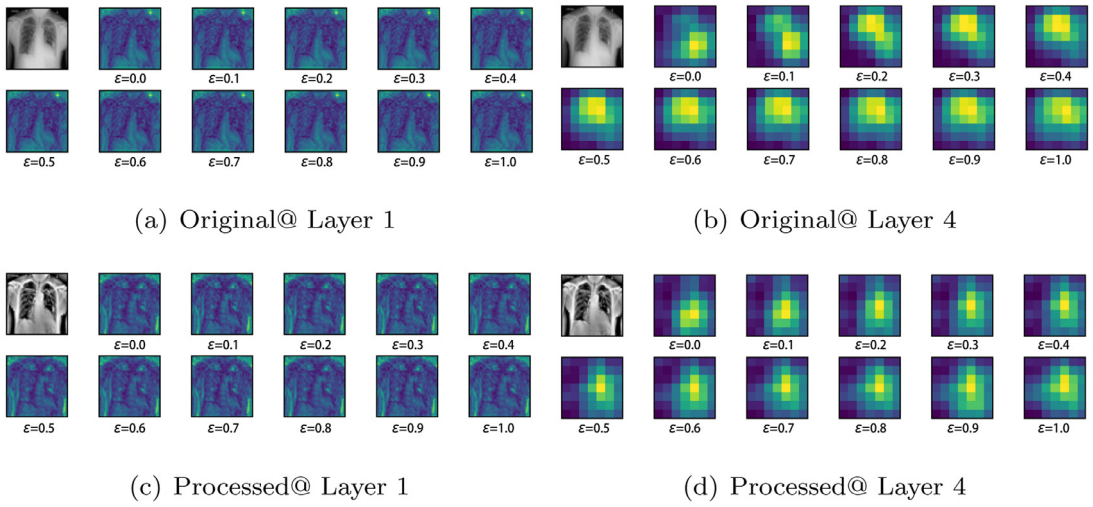


Fig. 4. Visualization of original, processed chest images and their feature maps extracted from Layer 1 and Layer 4 of ResNet50 under different adversarial perturbations. Note that in the feature map, the higher the brightness, the higher the weight.

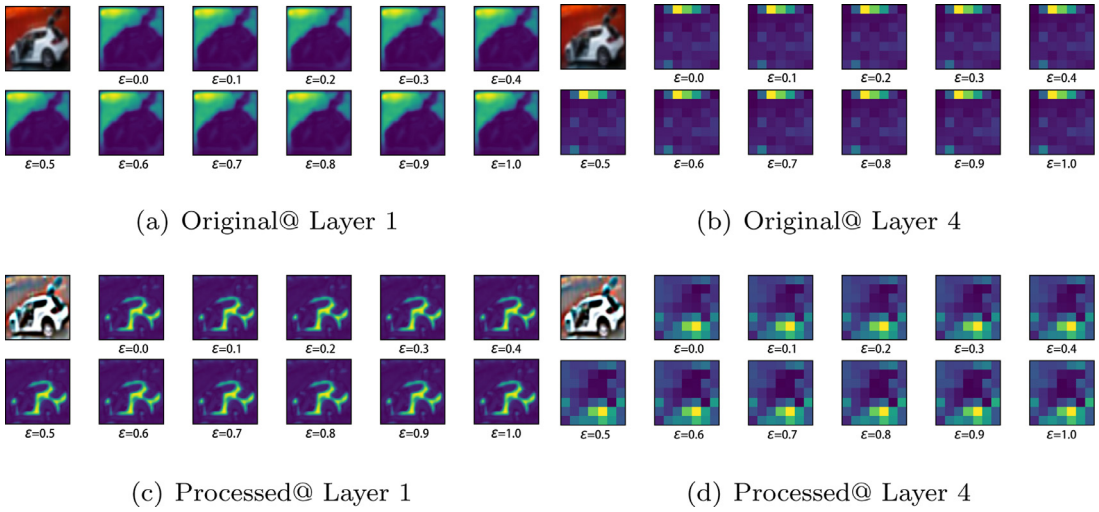


Fig. 5. Visualization of original, processed CIFAR10 images and their feature maps extracted from Layer 1 and Layer 4 of ResNet50 under different adversarial perturbations. Note that in the feature map, the higher the brightness, the higher the weight.

employ the following model to obtain $\tilde{\mathbf{D}}_j^m$:

$$\min_{\mathbf{D}_j^m} \frac{1}{2} \left\| \tilde{\mathbf{D}}_j^m - \mathbf{D}_j^m \right\|_F^2 + \left\| \lambda_j \otimes \mathbf{1}_{h^m} \otimes \mathbf{1}_{w^m} \cdot \mathbf{D}_j^m \right\|_1, \tag{4}$$

where $\lambda_j \in \mathbb{R}_{0+}^n$ is a regularization vector for removing the noise $\mathbf{e}_j^m$. The closed-form of Eq. (1) is $\mathbf{D}_j^m = \left\| \tilde{\mathbf{D}}_j^m - \lambda_j \otimes \mathbf{1}_{h^m} \otimes \mathbf{1}_{w^m} \right\|_+$. Similar to Eqs. (2) and (3), we also make λ_j dependent on $\tilde{\mathbf{D}}_j^m$, e.g. $\lambda_j = \frac{\gamma}{h^m w^m} \tilde{\mathbf{D}}_j^m \mathbf{1}_{w^m} \mathbf{1}_{h^m}$. Hence, we attain $\mathbf{D}_j^m$ by:

$$\mathbf{D}_j^m = \left\| \tilde{\mathbf{D}}_j^m - \frac{\gamma}{h^m w^m} (\tilde{\mathbf{D}}_j^m \mathbf{1}_{w^m} \mathbf{1}_{h^m}) \otimes \mathbf{1}_{h^m} \otimes \mathbf{1}_{w^m} \right\|_+. \tag{5}$$

Based on Eqs. (3) and (5), we can eliminate the inter-sample and intra-sample noise to attain the output matrix $\mathbf{D}_j^m$ $(1 \le j \le c^m)$, and then feed it into the following layer.

3. Experimental results and analysis

In this section, we first investigate the maximum adversarial perturbation that human specialists or the computer algorithm can check. Then, based on the investigation, we evaluate the performance of the proposed sparsity denoising method on model robustness against white-box and black-box attacks. We adopt the following two databases:

NIH chest X-ray normal/abnormal dataset [26]: This dataset is a subset of the National Institutes of Health (NIH) "Chest X-ray 14" dataset obtained retrospectively from the clinical PACS database at the NIH Clinical Center [30]. It contains 11,624 images, among which 8574 were used for training, 1706 for validation, and 1344 for testing. The labels were obtained by applying the Natural Language Processing tool on the radiology reports associated with the CXRs [31]. Specifically, we text-mined the radiological reports with 14 abnormal findings and binned them into the "abnormal" category. Cardiothoracic and pulmonary abnormalities included 14 abnormal findings: cardiomegaly, lung infiltration, mass, nodule, pneumonia, pneumothorax, atelectasis, consolidation, edema, emphysema, fibrosis, hernia, pleural effusion, and thickening. We binned the other images including the negative studies into the "normal" category. The testing set labels were obtained by taking the consensus of three US board-certified radiologists who read the text reports individually.

AREDS dataset [27]:We extracted 65,705 color fundus images from 4549 participants, and then divided them into two subsets:

(i) a training set of 52,539 images from 4099 participants captured at multiple study visits; and (ii) a testing set of bilateral images (i.e., one image from each eye) captured from the remaining 450 participants. We categorized these images into two classes: age-related macular degeneration (AMD) and no AMD. The training set had 34,682 images with AMD and 17,857 images with no AMD, and there were 8852 images with AMD and 4314 with no AMD in the testing set. In our experiments, we randomly selected 10% images from the training set for model validation and utilized the remaining ones for training.

For system development and evaluation, on CXRs, we present the performance of eight models: (1) ResNet50 [29], a widely used DNN as our frameworks backbone, (2) ResNet50-M [23], embedding median filters into ResNet50, (3) ResNet50-N [23], embedding nonlocal means and Gaussian filters into ResNet50, (4) ResNet50-D, our proposed denoising network, (5) ResNet50-A, ResNet50 with the popular adversarial training strategy [19], (6) ResNet50-A-M, ResNet50-M with the adversarial training strategy, (7) ResNet50-A-N, ResNet50-N with the adversarial training strategy, (8) ResNet50-A-D, ResNet50-D with the adversarial training strategy. On CFPs, we show the performance of four major models: ResNet50, ResNet50-D, ResNet50-A and ResNet50-A-D.

3.1. Implementation details

We implemented ResNet50, ResNet50-M, ResNet50-N, ResNet50-D, ResNet50-A, ResNet50-A-M, ResNet50-A-N, and ResNet50-A-D by using the PyTorch platform, where the backbone network ResNet50 is pretrained by the ImageNet database [32]. Additionally, we employed the optimizer, stochastic gradient descent (SGD), to optimize model parameters, and set the maximum learning rate $\eta = 0.005$, decayed by multiplying 0.1 at the 25th and 40th epoch, respectively. We trained the model 50 epochs using two GPUs and a batch size of 64. We resized all images to be $224 \times 224 \times 3$, and normalize their pixels to be within $[0, 1]$. For ResNet50-D, we chose $\gamma = 1$ on CXRs and CFPs. For ResNet50-A, we utilized PGD to produce adversarial samples, with the maximal pixel perturbation $\epsilon = 0.5$ on CXRs and $\epsilon = 0.25$ on CFPs, the attack step size $\alpha = \frac{\epsilon}{10}$, and the number of attack steps $s = 20$. For ResNet50-A-D, we set $\gamma = 0.1$ for CXRs and $\gamma = 0.2$ for CFPs, respectively. ResNet50-A-M, ResNet50-A-N and ResNet50-A-D employed the same setting as ResNet50-A on the two databases.

Before evaluating the robustness of aforementioned models, we employed a popular classifier, support vector machine (SVM) [33], to investigate the maximum adversarial perturbation that one computer algorithm can detect. For white-box robustness, we evaluated the performance of different models on adversarial examples with the perturbation $\epsilon \in [0, 1]$. For black-box robustness, we first employed a popular larger network, ResNet101 [29], to train a model with original training samples. Then we adopted PGD-20 to attack the well-trained model of ResNet101 to generate adversarial examples, which are called transferable adversarial examples for the models using ResNet50 as the backbone network. Finally, we utilized the well-trained models to test transferable adversarial examples. We evaluated the different models on transferable adversarial examples from CXRs and CFPs with the perturbation $\epsilon \in [0, 2]$. We repeat all evaluations five times and report their average results, and their standard deviations are usually within 1%.

Metrics for evaluation. To evaluate the performance of the methods under white-box and black-box attacks, we employed two metrics: Accuracy and F_1-score. They were calculated based on true positive (TP), false positive (FP), true negative (TN) and false negative (FN). Specifically, Accuracy $= \frac{TP+TN}{TP+FP+TN+FN}$, and F_1-score $= \frac{2TP}{2TP+FP+FN}$.

Table 1
Average detection accuracy (%) of specialists on CXRs and CFPs.

	CXRs					
	$\epsilon = 1$			$\epsilon = 2$		
	Normal	Abnormal	Avg.	Normal	Abnormal	Avg.
Same	100	100	100	100	100	100
Different	4	4	4	96	88	92
	CFPs					
	AMD	no AMD	Avg.	AMD	no AMD	Avg.
Same	98	100	99	98	98	98
Different	6	32	19	86	98	92

3.2. Detection of medical images under adversarial attacks

Before evaluating the performance of the proposed method on model robustness, we first examined the degree of pixel perturbations in CXRs and CFPs required for human specialists to detect adversarial attacks. The rationale is that human clinicians would likely ignore the output from CNNs when medical images are identified as having been attacked. We select 100 pairs (original image vs. original image, or original image vs. adversarial image) and divide them into two groups: 50 pairs consist of the same images (original image vs. original image), and 50 of them are composed of different ones (original image vs. adversarial image). For CXRs, each group has 25 pairs with normal images and the other 25 pairs are with abnormal images; For CFPs, each group contains 25 pairs with (age-related macular degeneration) AMD images and the remaining 25 pairs have the images without AMD. The adversarial examples generated by PGD-20 with the attack step size $\alpha = \frac{\epsilon}{4}$ at $\epsilon = 1$ and $\epsilon = 2$, respectively.

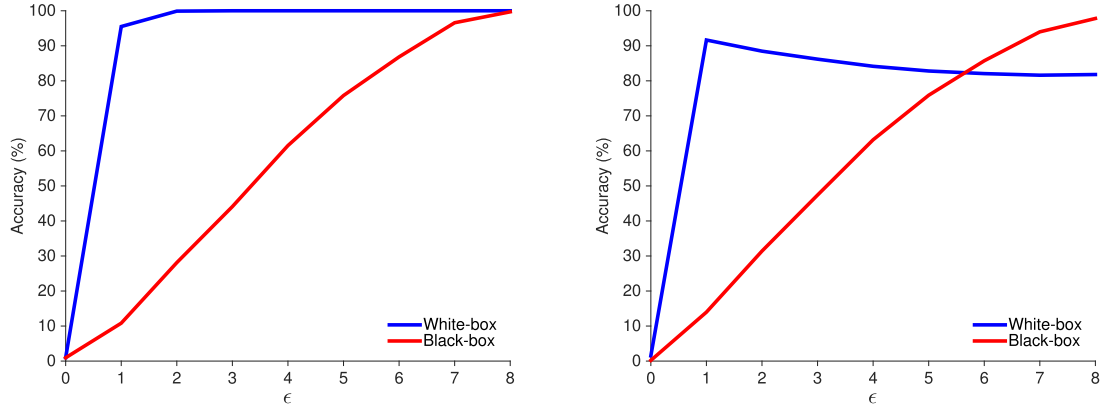
To investigate the maximum adversarial perturbation that human specialists can detect, Table 1 presents their detection accuracy on CXRs and CFPs, respectively. As we can see, with medical images under adversarial attacks by PGD-20, human specialists fail to detect images with a maximum adversarial perturbation of $\epsilon = 1$, but succeeded in detecting images (in 92% of cases) at $\epsilon = 2$.

We further found that, for white-box attacks by PGD-20 at $\epsilon = 1$, over 80% of adversarial examples of CXRs and CFPs were successfully detected (Fig. 6) by a computer algorithm, SVM. Due to the nature of the attacks, black-box ones are much more challenging to be detected automatically, even when $\epsilon = 2$. Hence, in our investigations of our denoising method, we set the threshold $\epsilon \in [0, 1]$ under white-box attacks and $\epsilon \in [0, 2]$ for black-box attacks, as perturbations larger than these thresholds can usually be detected by either human experts or computer algorithms.

3.3. Experiments on CXRs

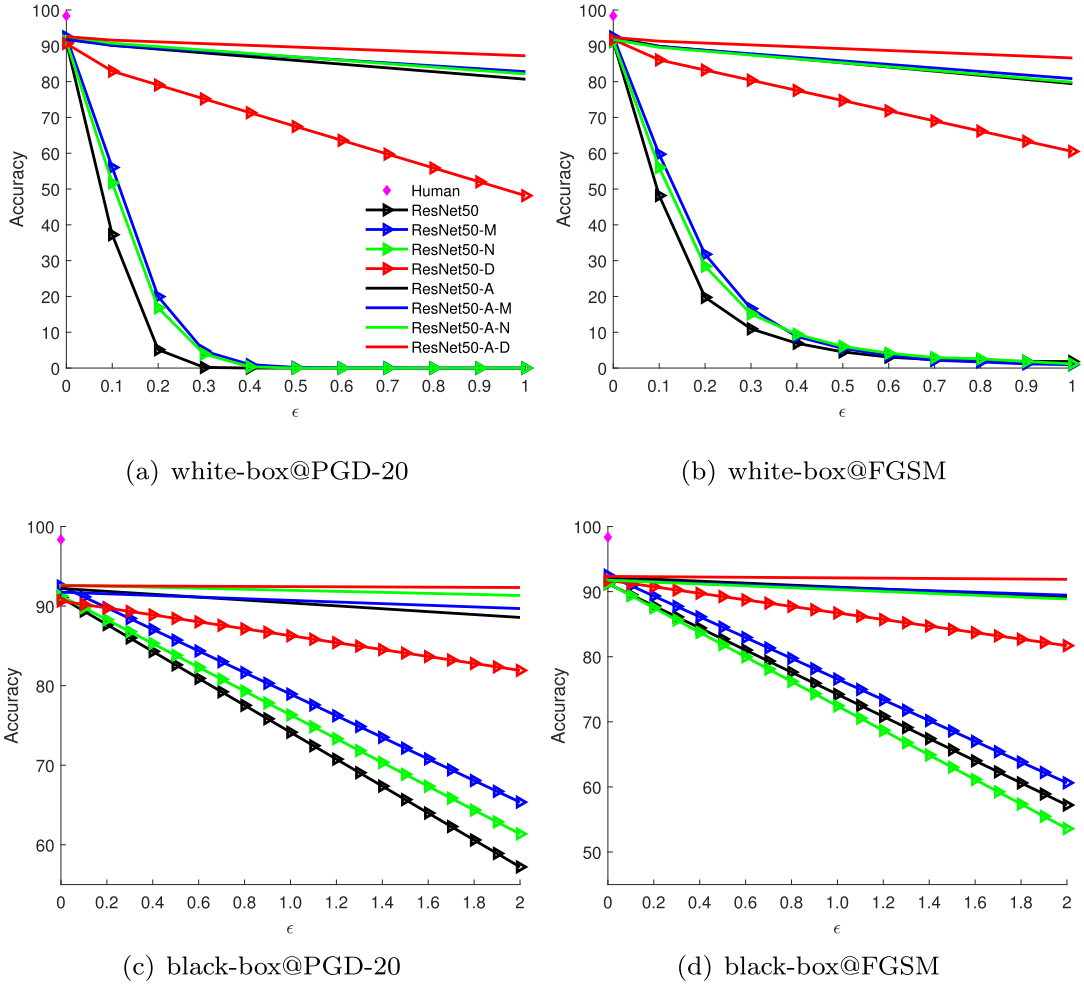
Robustness to white-box adversarial attacks Based on the detection accuracy of specialists in Table 1, we evaluate the performance of the eight methods on model robustness against white-box attacks with the maximum adversarial perturbation $\epsilon \in [0, 1]$. Specifically, we utilized PGD-20 to generate adversarial examples by directly attacking eight well-trained models of ResNet50, ResNet50-M, ResNet50-N, ResNet50-D, ResNet50-A, ResNet50-A-M, ResNet50-A-N, and ResNet50-A-D. Fig. 7a-b display their classification accuracy on CXRs under PGD-20 and a fast gradient sign method (FGSM) [11] at different adversarial perturbations, respectively, which modify images with the maximal pixels during $[0, 1]$. Tables 2 shows their testing accuracy on original ($\epsilon = 0$) and adversarial examples ($\epsilon = 1$) from CXRs under PGD-20.

For CXRs under PGD-20, without any adversarial attacks ($\epsilon = 0$), ResNet50 (93.76%) obtains the best accuracy than the others. How-



(a) CXRs (b) CFPs

Fig. 6. Detection accuracy on CXRs and CFPs for ResNet50. We first extracted features of normal and adversarial training examples from ResNet50, where adversarial training examples were generated by using FGSM to attack the well-trained model of ResNet50 with $\epsilon = 1$. Then we utilized SVM to train a classifier for detecting testing adversarial and transferable adversarial samples generated by PGD-20.



(a) white-box@PGD-20 (b) white-box@FGSM

(c) black-box@PGD-20 (d) black-box@FGSM

Fig. 7. Testing accuracy on CXRs under attacks by PGD-20 and FGSM. ϵ denotes the maximal pixels modified by adversarial perturbations and the attack step size $\alpha = \frac{\epsilon}{4}$ for PGD-20. The accuracy obtained by Human is on normal medical images.

ever, under adversarial attacks ($\epsilon = 1$), the accuracy of ResNet50 was substantially inferior to ResNet50-D and the other models with adversarial training. Specifically, the accuracy of ResNet50 decreased to zero, while ResNet50-D and ResNet50-A-D, were 45.68% and 87.20%, respectively. They illustrate that ResNet50-D obtains superior performance over ResNet50, ResNet50-M and ResNet50-N, and ResNet50-A-D attains better results than ResNet50-A, ResNet-A-M and ResNet-A-N. In particular, over 90% of ResNet50-A-D's original performance can be retained at $\epsilon = 1$.

Besides accuracy, we evaluated the models on F_1-score and observed the similar tendency (Table 2). In addition, similar results

Table 2
Testing results (%) on CXRs under an adversarial attack PGD-20.

Method	Original		Adversarial		Transferable	
	Accuracy	F_1-score	Accuracy	F_1-score	Accuracy	F_1-score
Radiologists	**98.36**	**98.33**	–	–	–	-
ResNet50	93.76	90.56	0	0	55.02	66.90
ResNet50-M	91.28	90.89	0	0	65.35	66.82
ResNet50-N	92.21	90.53	0	0	61.37	64.53
ResNet50-D	91.94	92.03	45.68	48.25	82.36	82.95
ResNet50-A	92.12	92.24	80.90	82.51	91.40	91.54
ResNet50-A-N	91.78	90.52	82.14	82.82	91.33	90.59
ResNet50-A-D	92.96	92.66	**87.20**	**87.05**	**92.54**	**92.02**

Table 3
Testing results (%) on CFPs under an adversarial attack PGD-20.

Method	Original		Adversarial		Transferable	
	Accuracy	F_1-score	Accuracy	F_1-score	Accuracy	F_1-score
Ophthalmologist	81.20	85.13	–	–	–	–
ResNet50	**85.20**	**88.48**	0	0	40.71	57.81
ResNet50-D	84.84	88.25	28.92	44.78	46.57	62.61
ResNet50-A	82.19	86.21	37.44	53.61	71.98	78.61
ResNet50-A-D	81.93	86.11	**48.66**	**58.82**	**74.97**	**80.70**

were obtained under FGSM. All these results demonstrate that our proposed model is robust to white-box adversarial attacks.

Robustness to black-box adversarial attacks Based on the detection accuracy of SVM in Fig. 6, we evaluate the performance of the proposed sparsity denoising on model robustness against black-box attacks with the maximum adversarial perturbation $\epsilon \in [0, 2]$. Specifically, we utilized PGD-20 to attack a well-trained model of ResNet101 and generate transferable adversarial examples. This scenario is more challenging to detect black-box attacks. Fig. 7c-d present the accuracy of the eight models on the transferable adversarial examples for CXRs under PGD-20 and FGSM, respectively. We observed that ResNet50-D was more robust than the ResNet50, ResNet50-M and ResNet50-N against transferable adversarial examples on CXRs. Additionally, ResNet50-A-D achieved a more robust performance than the others.

Tables 2 shows the accuracy of the four models on transferable adversarial examples generated by PGD-20 at $\epsilon = 2$, the accuracy of ResNet50 was only 55.02% on CXRs, while ResNet50-D and ResNet50-A-D had accuracies of 82.36% and 92.54%, respectively. As shown, ResNet50-A-D achieved similar accuracy on transferable adversarial examples of CXRs to ResNet50 (93.76%) on normal examples of CXRs. These results demonstrate that our proposed model is robust to black-box adversarial attacks, and similar findings can be observed under FGSM.

Moreover, we present the model performance of aforementioned methods on a different backbone network, VGG16 [34], in Fig. 11 (shown in the Appendix section). Similar observations can be attained.

3.4. Experiments on CFPs

Robustness to white-box adversarial attacks For CFPs under white-box attacks, we utilized PGD-20 to generate adversarial examples by directly attacking well-trained models of ResNet50, ResNet50-D, ResNet50-A, and ResNet50-A-D. Fig. 8a-b display their classification accuracy under PGD-20 and a fast gradient sign method (FGSM) [11] at different adversarial perturbations during [0,1], respectively. Tables 3 presents their testing accuracy on original ($\epsilon = 0$) and adversarial examples ($\epsilon = 1$) from CFPs under PGD-20.

For CFPs under PGD-20, without any adversarial attacks, the accuracies of ResNet50-D (84.84%), ResNet50-A (82.19%), and ResNet50-A-D (81.93%) were only slightly lower than that of ResNet50 (85.20%). However, two ophthalmologists only achieved a mean diagnostic accuracy of 81.20%. On adversarial attacks (ϵ=1), the accuracy of ResNet50 decreased to zero. By contrast, ResNet50-D, ResNet50-A, and ResNet50-A-D had accuracies of 28.92%, 37.44%, and 48.66%, respectively. Similarly, almost 60% of ResNet50-A-D original performance can be retained at $\epsilon = 1$. In addition, similar results were obtained under FGSM.

Robustness to black-box adversarial attacks For CFPs under black-box attacks, we utilized PGD-20 to attack a well-trained model of ResNet101 and generate transferable adversarial examples. Fig. 8c-d present the accuracy of ResNet50, ResNet50-D, ResNet50-A and ResNet50-A-D on the transferable adversarial examples for CFPs under PGD-20 and FGSM, respectively. Tables 3 displays the accuracy of the four models on transferable adversarial examples generated by PGD-20 at $\epsilon = 2$. As we can see, ResNet50 obtained an accuracy of 40.71% on CFPs, while the other models had accuracies of 46.57%, 71.98%, and 74.97%, respectively. In particular, the accuracy of ResNet50-A-D was only 6.23% lower than that of ophthalmologists (81.2%). Similar findings can be observed under FGSM.

3.5. Why sparsity denoising can improve CNNs' robustness to adversarial attacks?

Figs. 7 and 8 illustrate that ResNet50-D (or ResNet50-A-D) is more robust than ResNet50 (or ResNet50-A) on small perturbations generated by adversarial attacks. The underlying possible reason is that models with our methodology usually obtained a smaller feature distance between the (original and preprocessed) inputs and their adversarial examples than models without, because sparsity denoising reduces the noise in outputs of each convolutional layer. For clarity, Fig. 9 shows testing accuracy of ResNet50 and ResNet50-D on original and preprocessed CFPs, and the feature distance between inputs and their adversarial examples. As shown, ResNet50-D can achieve smaller feature distances than ResNet50 on both original and preprocessed images, thereby leading to better testing accuracy. Additionally, ResNet50-D using preprocessed images can further improve model accuracy that using original images. Similar observations can be found on CXRs.

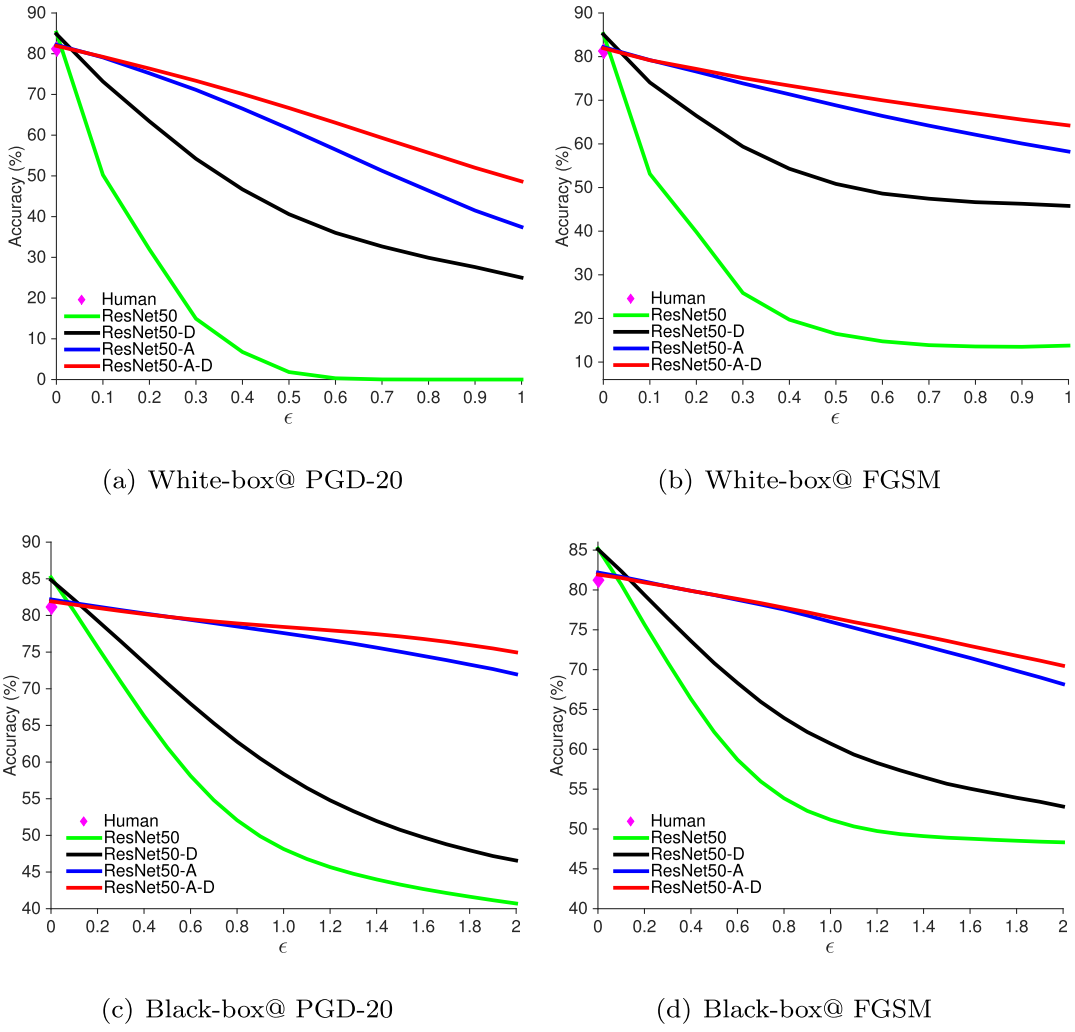


Fig. 8. Testing accuracy on CFPs under the attacks by PGD-20 and FGSM. ϵ denotes the maximal pixels modified by adversarial perturbations and the attack step size $\alpha = \frac{\epsilon}{4}$ for PGD-20. The accuracy obtained by Human is on normal medical images.

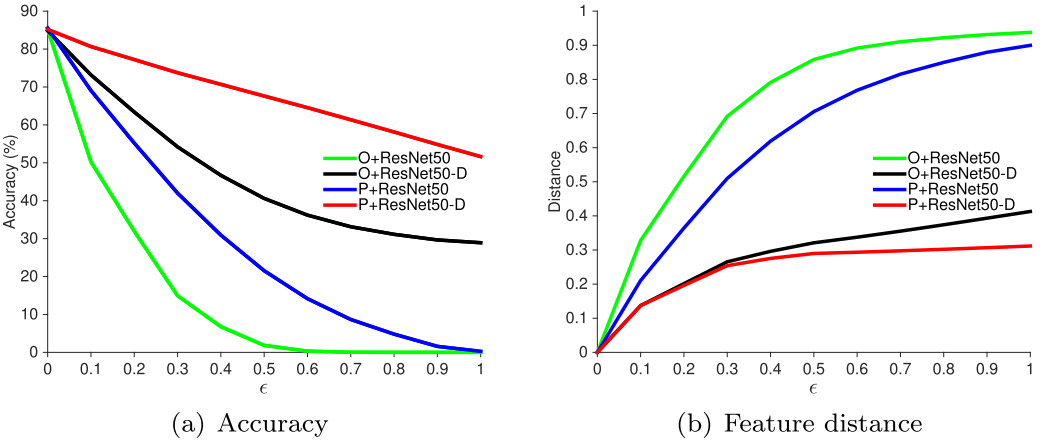
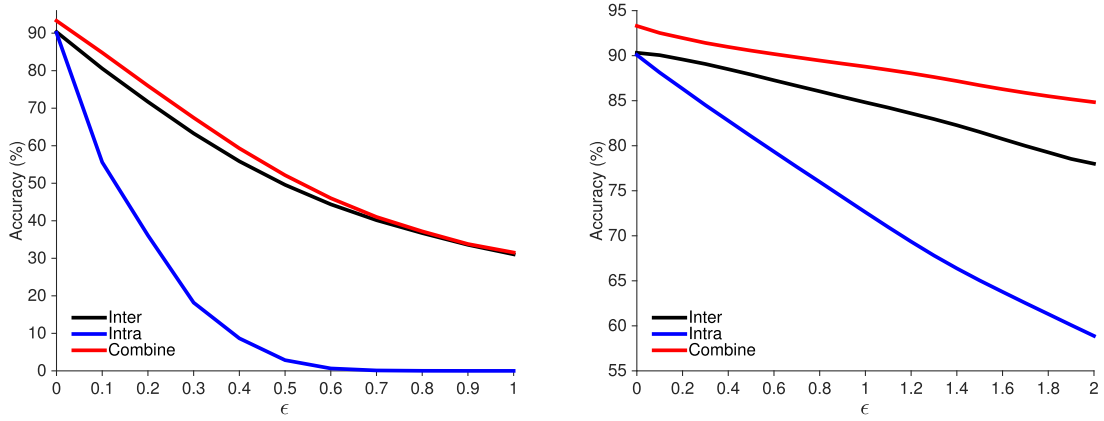
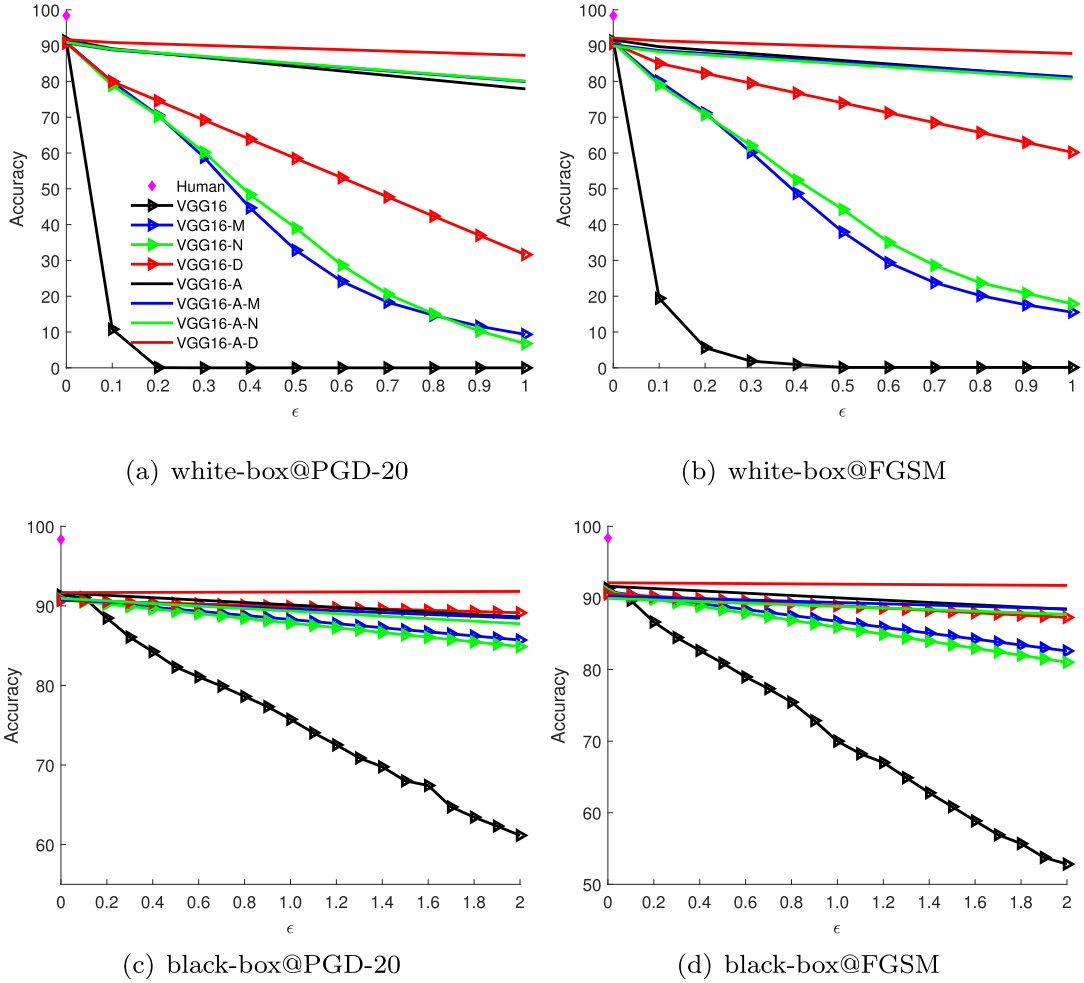


Fig. 9. Testing accuracy and feature distance of ResNet50 and ResNet50-D under the white-box attack by PGD-20. Feature distance is the Euclidean distance between original images and their adversarial examples under different perturbations, where features are extracted from the last convolutional layer of ResNet50 or ResNet50-D. ϵ denotes the maximal pixels modified by adversarial perturbations and the attack step size $\alpha = \frac{\epsilon}{4}$. 'O' and 'P' respectively denote original and preprocessed images obtained by using Gaussian filters to preprocess original training and testing data.



(a) White-box attacks (b) Black-box attacks

Fig. 10. Testing accuracy of ResNet50-D with only inter-sample denoising, intra-sample denoising and their combinations on CXRs under the attack PGD-20 with the attack step size $\alpha = \frac{\epsilon}{4}$. ϵ denotes the maximal pixels modified by adversarial perturbations. 'Inter' denotes that ResNet50-D only leverages the inter-sample denoising layer, 'Intra' represents only employing the intra-sample denoising layer, and 'Combine' means utilizing both inter-sample and intra-sample denoising layers.



(a) white-box@PGD-20 (b) white-box@FGSM

(c) black-box@PGD-20 (d) black-box@FGSM

Fig. 11. Testing accuracy of eight methods using VGG16 as the backbone on CXRs under attacks by PGD-20 and FGSM. ϵ denotes the maximal pixels modified by adversarial perturbations and the attack step size $\alpha = \frac{\epsilon}{4}$ for PGD-20. The accuracy obtained by Human is on normal medical images.

3.6. Ablation study

We further examine the roles of two denoising layers in our proposed framework. Fig. 10 presents testing accuracy of ResNet50-D on three cases: (1) only using inter-sample denoising operator (Eq. (3)); (2) only using intra-sample denoising operator (Eq. (5)); (3) using both inter-sample and intra-sample denosing operators (Eq. (3)+Eq. (5)). Fig. 10 indicates that inter-sample denoising is the key to defend against both adversarial and transferable adversarial examples, probably because it utilizes the other

samples to reduce the noise in feature representations of one sample in each batch. Meanwhile, intra-sample denoising can further enhance model robustness against both white-box and black-box attacks, especially for black-box attacks.

3.7. Discussion

We developed, trained, and validated a CNN framework against the adversarial attacks in medical images. Our approach utilized inter-sample and intra-sample denoising layers to reduce the noise in visual representations, resulting in enhanced robustness to adversarial perturbations in both black-box and white-box attack settings. Additionally, the success of combining the popular AT and our method suggests that AT might be complementary to the denoising layers in our method by further decreasing the noise manipulated by adversarial perturbations. Note that when the norm for adversarial examples becomes larger than a threshold, the model accuracy might rise, probably because a larger norm leading to a larger optimization iteration step might cause the attack methods, including FGSM and PGD, fail to find the optima.

Previously proposed feature denoising methods [23] treat adversarial perturbations as the primary cause of the noise in feature representations and aim to eliminate the perturbations using nonlocal means and filters. Alternatively, our method hypothesizes that inherent noisy features in images lead to noise in learned feature representations of CNNs. Hence, our goal is to remove the noise in feature representations during CNN training based on sparsity denoising. Experiments on CXRs illustrate that sparsity denoising can obtain superior performance over previous feature denoising methods [23], probably because the proposed method can better reduce the noise in feature representations.

Additionally, our proposed method differs fundamentally from input denoising methods [35], which only consider transforming input images before feeding them into CNN networks, by making direct enhancements to the CNNs architecture for improved robustness. Although such input denoising methods may directly reduce noisy features in high-dimensional inputs for defending against adversarial perturbations, they have been demonstrated to be ineffective in general and are usually vulnerable to black-box attacks [36].

4. Conclusion and future work

In this paper, we propose a novel robust CNN framework, which embeds sparsity denoising operators into each layer of the network, to reduce the effect of noisy features learned during training, based on one hypothesis and discovery that noise learned during CNN training is one significant cause of CNNs' vulnerability to adversarial attacks. The proposed sparsity denoising operators consist of inter-sample and intra-sample denoising, where inter-sample denoising employs the entire batch of data to decrease the noise in outputs of each layer and intra-sample denoising reduces the noise in each data. Experimental results on two medical databases demonstrate the effectiveness of sparsity denoising on boosting CNNs' robustness against adversarial attacks. Additionally, they illustrate that the inter-sample denoising is a key contributor of CNNs' robustness. Moreover, they suggest the correctness of our hypothesis and discovery.

Our experimental results might suggest that, if CNNs could eliminate all noise in their learned feature representations, they would be more robust against adversarial perturbations. As such, in the future, we will apply the proposed denoising method to other different types of images, and further improve the denoising capability of CNNs to reduce the noise in their learned feature representations. For example, attention mechanisms might improve model robustness against adversarial perturbations, because they can simultaneously enhance significant features and lower trivial features. Additionally, the proposed denoising method might have the capability to boost the robustness and generalizability of CNNs in one common scenario in practice. CNNs' performance is often reported to be unstable on medical images from multiple diverse sources. Given promising results in this work, we would like to investigate such possibilities in the future. Although sparsity denoising can obtain promising performance, when images are tested one by one, inter-sample denoising requires additional images (or a dictionary) to reduce the noise in feature representations of each testing image, thereby requiring more testing time than the model without. In the future, we will further investigate the reason for the success of inter-sample denoising, and then reduce the testing time.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This work is supported by the NIH Intramural Research Program, National Library of Medicine.

We present the performance of eight models using VGG16 as the backbone network: (1) VGG16 [34], (2) VGG16-M [23], embedding median filters into VGG16, (3) VGG16-N [23], embedding nonlocal means and Gaussian filters into VGG16, (4) VGG16-D, embdding sparsity denoising into VGG16, (5) VGG16-A, VGG16 with the popular adversarial training strategy, (6) VGG16-A-M, VGG16-M with the adversarial training strategy, (7) VGG16-A-N, VGG16-N with the adversarial training strategy, (8) VGG16-A-D, VGG16-D with the adversarial training strategy. Their performance on CXRs is shown in Fig. 11.

References

[1] G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian, J.A. Van Der Laak, B. Van Ginneken, C.I. Sánchez, A survey on deep learning in medical image analysis, Med. Image Anal. 42 (2017) 60–88.
[2] C. Barata, M.E. Celebi, J.S. Marques, Explainable skin lesion diagnosis using taxonomies, Pattern Recognit. 110 (2021) 107413.
[3] A.N. Basavanhally, S. Ganesan, S. Agner, J.P. Monaco, M.D. Feldman, J.E. Tomaszewski, G. Bhanot, A. Madabhushi, Computerized image-based detection and grading of lymphocytic infiltration in her2+ breast cancer histopathology, IEEE Trans. Biomed. Eng. 57 (3) (2009) 642–653.
[4] M.M. Dundar, S. Badve, G. Bilgin, V. Raykar, R. Jain, O. Sertel, M.N. Gurcan, Computerized classification of intraductal breast lesions using histopathological images, IEEE Trans. Biomed. Eng. 58 (7) (2011) 1977–1984.
[5] Q. Zhang, C. Lin, F. Li, Application of binocular disparity and receptive field dynamics: a biologically-inspired model for contour detection, Pattern Recognit. 110 (2021) 107657.
[6] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[7] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.
[8] M. Xu, J. Yao, Z. Zhang, R. Li, B. Yang, C. Li, J. Li, J. Zhang, Learning eeg topographical representation for classification via convolutional neural network, Pattern Recognit. 105 (2020) 107390.
[9] D.S. Kermany, M. Goldbaum, W. Cai, C.C. Valentim, H. Liang, S.L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al., Identifying medical diagnoses and treatable diseases by image-based deep learning, Cell 172 (5) (2018) 1122–1131.
[10] S.M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G.C. Corrado, A. Darzi, et al., International evaluation of an ai system for breast cancer screening, Nature 577 (7788) (2020) 89–94.
[11] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, in: International Conference on Learning Representations, 2015.
[12] X. Bai, X. Wang, X. Liu, Q. Liu, J. Song, N. Sebe, B. Kim, Explainable deep learning for efficient and robust pattern recognition: a survey of recent developments, Pattern Recognit. 120 (2021) 108102.

[13] K. Xu, H. Chen, S. Liu, P.-Y. Chen, T.-W. Weng, M. Hong, X. Lin, Topology attack and defense for graph neural networks: an optimization perspective, in: International Joint Conference on Artificial Intelligence, 2019, pp. 3961–3967.
[14] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, X. Lin, Adversarial t-shirt! evading person detectors in a physical world, in: European Conference on Computer Vision, 2020, pp. 665–681.
[15] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, F. Lu, Understanding adversarial attacks on deep learning based medical image analysis systems, Pattern Recognit. (2020) 107332.
[16] X. Li, D. Zhu, Robust detection of adversarial attacks on medical images, in: International Symposium on Biomedical Imaging (ISBI), 2020, pp. 1154–1158.
[17] M. Paschali, S. Conjeti, F. Navarro, N. Navab, Generalizability vs. robustness: investigating medical imaging networks using adversarial examples, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2018, pp. 493–501.
[18] S.G. Finlayson, J.D. Bowers, J. Ito, J.L. Zittrain, A.L. Beam, I.S. Kohane, Adversarial attacks on medical machine learning, Science 363 (6433) (2019) 1287–1289.
[19] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, in: In International Conference on Learning Representations, 2018.
[20] X. Han, Y. Hu, L. Foschini, L. Chinitz, L. Jankelson, R. Ranganath, Deep learning models for electrocardiograms are susceptible to adversarial attack, Nat. Med. (2020) 1–4.
[21] T.K. Yoo, J.Y. Choi, Outcomes of adversarial attacks on deep learning models for ophthalmology imaging domains, JAMA Ophthalmol. 138 (11) (2020) 1213–1215.
[22] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, D. Mukhopadhyay, Adversarial attacks and defences: a survey, arXiv preprint arXiv:1810.00069(2018).
[23] C. Xie, Y. Wu, L.v.d. Maaten, A.L. Yuille, K. He, Feature denoising for improving adversarial robustness, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 501–509.
[24] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, D. Mukhopadhyay, A survey on adversarial attacks and defences, CAAI Trans. Intell. Technol. 6 (1) (2021) 25–45.
[25] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, A. Madry, Adversarial examples are not bugs, they are features, in: Advances in Neural Information Processing Systems, 2019, pp. 125–136.
[26] Y.-X. Tang, Y.-B. Tang, Y. Peng, K. Yan, M. Bagheri, B.A. Redd, C.J. Brandon, Z. Lu, M. Han, J. Xiao, et al., Automated abnormality classification of chest radiographs using deep convolutional neural networks, NPJ Digit. Med. 3 (1) (2020) 1–8.
[27] Y. Peng, S. Dharssi, Q. Chen, T.D. Keenan, E. Agrón, W.T. Wong, E.Y. Chew, Z. Lu, Deepseenet: a deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs, Ophthalmology 126 (4) (2019) 565–575.
[28] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images (2009).
[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[30] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2097–2106.
[31] Y. Peng, X. Wang, L. Lu, M. Bagheri, R. Summers, Z. Lu, Negbio: a high-performance tool for negation and uncertainty detection in radiology reports, AMIA Summits Transl. Sci. Proc. 2018 (2018) 188.
[32] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[33] R.-E. Fan, K.-W. Chang, C.-j. Hsieh, X. Wang, C.-J. Lin, Liblinear: a library for large linear classification, J. Mach. Learn. Res. 9 (2008) 1871–1874.
[34] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition (2015).
[35] W. Xu, D. Evans, Y. Qi, Feature squeezing: Detecting adversarial examples in deep neural networks, in: Network and Distributed Systems Security Symposium, 2018, pp. 2080–2088.
[36] A. Athalye, N. Carlini, D. Wagner, Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples, in: International Conference on Machine Learning, 2018.

Xiaoshuang Shi is an assistant professor in the Department of Computer Science and Engineering at the University of Electronic Science and Technology of China (UESTC). He obtained his PhD degree (2019) from University of Florida, Master degree (2013) from Tsinghua University, and Bachelor degree (2009) from Northwestern Polytechnical University. Before joining UESTC, he worked as a Postdoctoral fellow at the National Institutes of Health (NIH) (2020.01–2021.04), and as a research assistant at Tsinghua University (2013.09–2015.04). His major research interests include large-scale data retrieval, deep learning, medical image analysis.

Yifan Peng received the PhD degree. He is currently an assistant professor with Weill Cornell Medicine. He was a research fellow with the National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH). His main research interests include biomedical and clinical natural language processing and medical image analysis. He has published many papers in top journals and conferences, including the Nucleic Acids Research, npj Digital Medicine, Journal of the American Medical Informatics Association, CVPR, and MICCAI. He is also an academic editor of the PLoS ONE.

Qingyu Chen is a Postdoctoral fellow at the National Institutes of Health (NIH). He obtained his PhD degree (2018) from University of Melbourne, and Bachelor degree (2013) from RMIT University. His research interests include biomedical text mining and information retrieval, medical image analysis, healthcare applications, and Biocuration.

Tiarnan Keenan is staff clinician in retinal disease in the Division of Epidemiology and Clinical Applications at the National Eye Institute (NEI). He received his undergraduate and medical degrees as a scholar at the University of Oxford. Funded by the UK National Institute of Health Research, he completed integrated academic-clinical training in ophthalmology and biomedical research predominantly at the Manchester Royal Eye Hospital and the Oxford Eye Hospital. He was awarded a Ph.D. at the University of Manchester for biochemical analyses of human macular tissue in relation to age-related macular degeneration (AMD). As a Fulbright scholar at the Moran Eye Center (University of Utah), he conducted post-graduate research into the genetics and mechanisms of AMD. He was admitted as a fellow (FRCOphth) into the UK Royal College of Ophthalmologists and, as a Bayer Global Ophthalmology Awards Program awardee, completed fellowship training in medical retinal disease at NEI. His research is focused on adult retinal disease, particularly age-related macular degeneration (AMD), the leading cause of legal blindness in all developed countries.

Alisa T. Thavikulwat is working at National Eye Institute (NEI), NIH. She obtained Bachelor and PhD degree from the University of Rochester. Her research interests include ophthalmology and age-related macular degeneration.

Sungwon Lee received the MD and PhD degrees. She is currently a radiologist and research fellow with the National Institutes of Health (NIH). Her research interests include segmentation and clas- sification of medical imaging, especially chest, body, and musculoskeletal images of CT and MRI.

Yuxing Tang received the BS and MS degrees from the Department of Information and Telecom-munication Engineering, Beijing Jiaotong University, Beijing, China, in 2009 and 2011, respectively, and the PhD degree in computer science from the Department of Mathematics and Computer Science, Ecole Centrale de Lyon, Ecully, France, in 2016. He is a postdoctoral fellow with the Imaging Biomarkers and Computer-Aided Diagnosis (CAD) Laboratory, National Institutes of Health (NIH) Clinical Center. His main research interests include computer vision and machine learning, in particular, deep learning techni- ques for visual category recognition, object detection, image segmentation and their application in medical imaging.

Emily Y. Chew is the Director of the Division of Epidemiology and Clinical Applications (DECA), at the National Eye Institute, the National Institutes of Health in Bethesda, Maryland. She is also the Chief of the Clinical Trials Branch in the division. Emily received her medical degree and her ophthalmology training at the University of Toronto, Canada. She completed her fellowship in medical retina at the Wilmer Eye Institute, the Johns Hopkins Medical Institutes and the University of Nijmegen, the Netherlands. Her research interest includes phase I/II clinical trials and epidemiologic studies in retinovascular diseases such as age-related macular degeneration, diabetic retinopathy, and other ocular diseases. She has worked extensively in large multi-centered trials headed by the staff from her division, including the Early Treatment Diabetic Retinopathy Study (ETDRS), the Age-Related Eye Disease Study (AREDS) and the Age-Related Eye Disease Study 2 (AREDS2), which she chairs. She works on other clinical trials in collaboration with other institutes within NIH such as the Actions to Control Cardiovascular Risk in Diabetes (ACCORD) Trial and she chairs the ACCORD Eye Study. She directs the clinical portion of the international study, Macular Telangiectasia Project.

Ronald M. Summers received the MD and PhD degrees. He is currently a senior investigator with the NIH. He joined the Diagnostic Radiology Department, NIH Clinical Center, in 1994. He directs the Imaging Biomarkers and Computer-Aided Diagnosis (CAD) Laboratory. His research interests include virtual colonoscopy, CAD, multi-organ multi-atlas registration, and development of large radiologic image databases. His clinical areas of specialty are thoracic and gastrointestinal radiology and body cross-sectional imaging. His current research focuses on developing fully-automated interpretation of abdominal CT scans.

Zhiyong Lu received the PhD degree. He is currently a deputy director for Literature Search at the National Center for Biotechnology (NCBI), leading its overall efforts of improving literature search and information access in NCBIs production resources. He is also an NIH senior investigator (early tenure) and directs the Text Mining / Natural Language Processing (NLP) Research Program, NCBI/NLM, where they are developing computational methods and software tools for analyzing and making sense of unstructured text data in biomedical literature and clinical notes towards accelerated discovery and better health.