Defense Against Adversarial Attacks based on Stochastic Descent Sign Activation Networks on Medical Images

Yanan Yang, Frank Y. Shih, and Usman Roshan

Department of Computer Science

New Jersey Institute of Technology, Newark, NJ 07102

Contact: shih@njit.edu

Abstract

Machine learning techniques in medical imaging systems are accurate, but minor perturbations in the data known as adversarial attacks can fool them. These attacks make the systems vulnerable to fraud and deception, and thus a significant challenge has been posed in practice. We present the gradient-free trained sign activation networks to detect and deter adversarial attacks on medical imaging AI systems. Experimental results show that a higher distortion value is required to attack our proposed model than other existing state-of-the-art models on MRI, chest X-ray, and histopathology image datasets, where our model outperforms the best and even twice superior. The average accuracy of our model in classifying the adversarial examples is 88.89%, whereas MLP and LeNet are 81.48%, and ResNet18 is 38.89%. It is concluded that the sign network is a solution to defend adversarial attacks due to high distortion and high accuracy on transferability. Our work is a significant step towards safe and secure medical AI systems.

Keywords: Robust machine learning; adversarial attack; medical AI imaging system; medical image classification

1. Introduction

Medical images, such as magnetic resonance imaging (MRI), computational tomography (CT), and histopathology, provide detailed information for diagnosing various diseases. With more accurate and efficient classification systems on medical images being deployed, the demands on robust medical machine learning systems have increased. The systems can help the experts diagnose diseases and accelerate treatment processes.

The classification approaches of human experts and machine learning systems are different. The human experts search for abnormal areas which are distinguishable from normal portions using their knowledge. While the machine learning techniques use a function to map the healthy and unhealthy as the specific labels. They learn the information in a supervised way. Therefore, the quality and resolution of medical image datasets could impact the performance of machine learning.

Machine learning algorithms have been proven to achieve high accuracy in the classification tasks and more new modules have been proposed to enhance accuracy[1]; however, they could misclassify by minor perturbations in such data known as adversarial attacks [2-5]. Adversarial examples have been shown to transfer across models, making it possible to perform transfer-based (substitute model) black-box attacks. Transfer adversarial attacks and boundary attacks are the most lethal as they can be performed effectively without access to the model's parameters [6].

The attackers can fool machine learning systems with adversarial images, which are often imperceptible to human eyes. In other words, the models could make mistakes by these adversarial inputs, which are intentionally crafted. As a result, machine learning systems would generate false results, misdiagnosis, or even insurance fraud.

Researchers have investigated adversarial attacks on medical images and mainly focused on testing the robustness of deep learning models designed for medical image analysis [7, 8]. Paschali et al. [9] showed that classification accuracy drops from above 87% on the regular medical images to almost 0% on the adversarial examples. Hokuto et al. [10] demonstrated UAPs achieved over 80% success rates on DNNs model. Many defense methods have been proposed to defend against adversarial attacks, in which adversarial training is most prevalent [11, 12]. However, this tends to lower accuracy on clean test data. To overcome this problem, the transfer-based methods were developed [13, 14], but they are still vulnerable. Thus, adversarial robustness is still an open problem in machine learning.

Gradient-free trained sign activation networks have been proven to be able to defend against adversarial attacks with a higher possibility [15, 16]. These networks are trained with a stochastic coordinate descent algorithm [17, 18], and their higher minimum distortions indicate that an image must comply with a more distinct modification to fool a model. In this paper we adopt the gradient-free stochastic coordinate descent algorithm for training sign activation networks on medical image datasets, including MRI, chest X-ray, and histopathology. The rest of this paper is organized as follows. Section 2 presents the proposed sign activation networks. Section 3 describes the datasets and experimental results. Conclusions are drawn in Section 4.

2. The Proposed Sign Activation Networks

We propose to train the sign activation networks with a gradient-free stochastic coordinate descent algorithm, which is named as the *Stochastic Coordinate Descent*, abbreviated by SCD.

2.1 The Stochastic Coordinate Descent (SCD)

We denote a given binary class data $x_i \in R^d$ and $y_i \in \{-1, +1\}$, for i = 0, 1, ..., n−1. A linear classifier $w \in R^d$, $w_0 \in R$ minimizes the empirical risk for a given loss function defined as

$$L_{scd} = \sum_i L(w, w_0, x_i, y_i) \quad (1)$$

We start with a random solution $w_i \in N(0, 1)$, $w_0 \in N(0, 1)$, for i = 0, 1, ..., d − 1 and iteratively make incremental changes that improve the risk. In each iteration, we select a random set of features (coordinates) from w called F. For each feature $w_i \in F$, we add/subtract a learning rate η and then determine w_0 that optimizes the risk. We compute all possible values of w_0 defined as

$$w_0 = \frac{w_i^T x_i + w_{i+1}^T x_{i+1}}{2} \quad (2)$$

for i = 0, 1, ..., n − 2 and select the one that minimizes the loss L_{scd}. A random sample of the training data in each iteration is generated to avoid local minima. To train a single hidden layer network, we apply SCD to the final node and then a randomly selected hidden node in each iteration of the algorithm. We apply parallelism and several heuristics in practice to speed up the run time.

2.2 Network Implementation

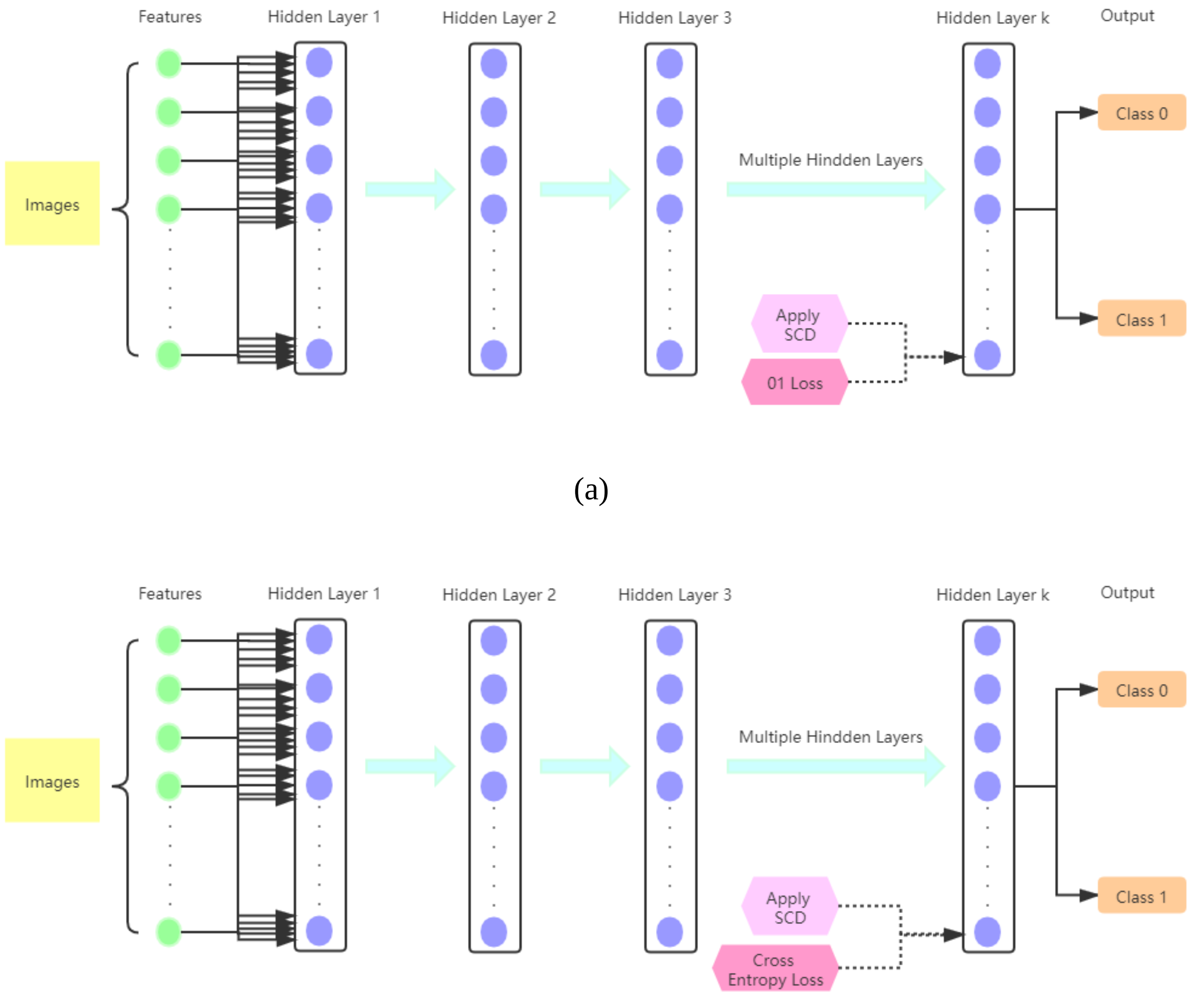
We train the following three types of sign activation networks using the proposed algorithm:

(1) SCD01: 01-loss in the final node

(2) SCDCE: Cross-entropy loss in the final node

(3) SCDCEBNN: Cross-entropy in the final node with binary weights throughout the model

The basic architecture of SCD models is shown in Fig. 1. The training procedure is implemented in Python, Numpy, and Pytorch [19]. Since sign activation is non-convex, our training process starts from a different random initialization. We run it 100 times and output the majority vote.



(a)

(b)

Fig. 1. The Architecture of SCD models. (a) The sign activation networks with our algorithm and 01-loss in the final node, (b) the sign activation networks with our algorithm and cross-entropy loss in the final node.

To illustrate the run time and clean test accuracies, we compare our models with the convolutional networks LeNet [22], ResNet18 [23], and a single hidden layer of 20 nodes to the equivalent network with sigmoid activation and logistic loss function (denoted as MLP). The MLP classifier in scikit-learn is used to implement MLP and the Larq library with the approximation to the sign activation. In addition, we use the HopSkipJump implementation in the IBM Adversarial Robustness Toolkit [14]. It is a family of algorithms and includes both untargeted and targeted attacks optimized for L2 and L∞ similarity metrics respectively. The model is developed based on a novel estimate of the gradient direction using binary information at the decision boundary. Theoretical analysis and experiments show HopSkipJump requires significantly fewer model parameters than several state-of-the-art decision-based adversarial attacks. It also achieves competitive performance in attacking several widely-used defense mechanisms.

In Fig. 2, a predictive model is attacked by HopSkipJump to generate an adversarial image, which would fool the model. To obtain as accurate of estimation as possible, we run HopSkipJump 10 times. In each time, we use an initial pool size of 1,000 random data points and maximum iterations of 100 to report the minimum value. For a single data point, this typically takes several hours to finish. Thus, we can report the distortion of only five random points.

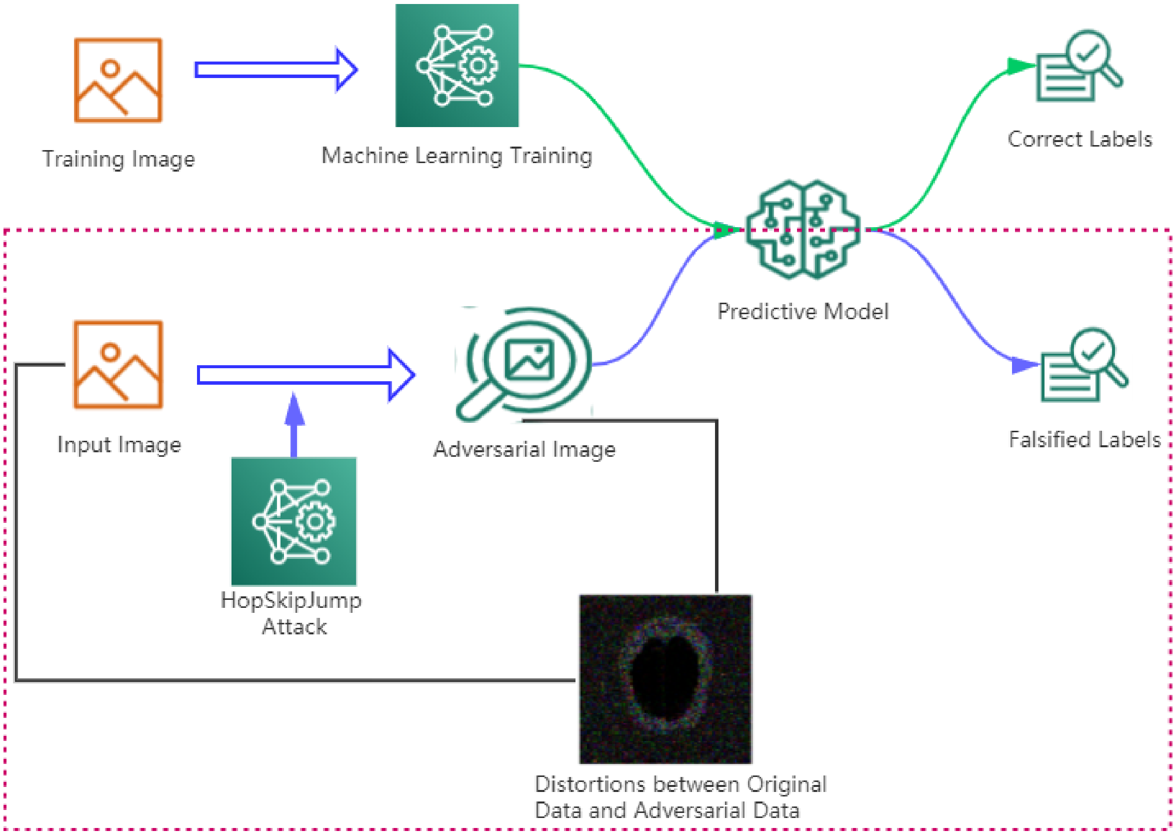


Fig. 2. The procedure of attacking the models with HopSkipJump.

3. Experimental Results

3.1 Datasets

We use three popular medical imaging datasets: BraTs18, Chest X-rays, and Colorectal Histopathology, to evaluate the classification accuracy.

3.1.1 BraTs18

The BratS18 dataset is 210 high-grade glioma (HGG) and 75 low-grade glioma (LGG)

MRI with binary masks for the tumor. Each 3D MRI contains 155 slices of size 240 × 240. We use the FLAIR modality images for all the experiments because the entire tumor is represented well by this modality. In total, we have 17,100 abnormal and 18,500 normal images for training. For testing, we have 1,800 abnormal and 1,900 normal images. We show more experimental results on other modalities, where ANT-GAN presents impressive synthesis quality. A more detailed medical description of the data can be referred to [20]. We down-sample two classes to be a balanced dataset, and each class contains 1,462 images, which are resized to 96 × 96. We split training and testing datasets by a ratio of 80 : 20.

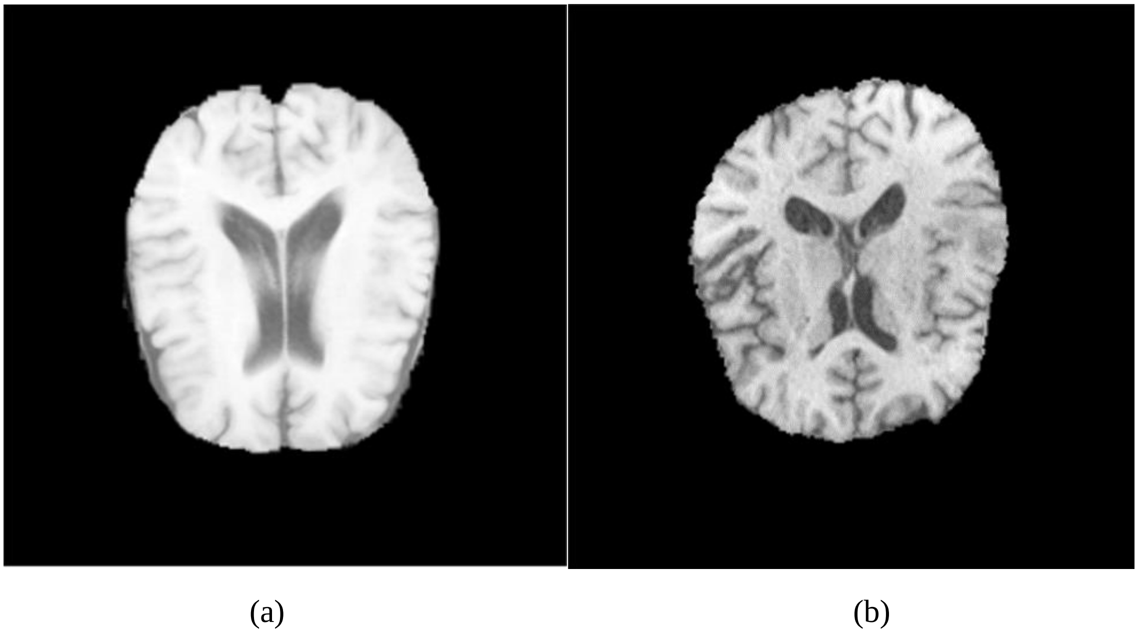
3.1.2 Chest X-rays

The Chest X-ray images (anterior-posterior) are selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou, China. All chest X-ray imaging was performed as part of patients' routine clinical care. The dataset is organized into two folders (train and test) and contains subfolders for each image category (pneumonia/normal). There are 5,863 X-ray images and 2 categories (pneumonia/normal). All chest radiographs are initially screened for quality control by removing all low quality or unreadable scans. Two expert physicians then grade the diagnoses for the images before being cleared for training the AI system. To account for any grading errors, the evaluation set is checked by a third expert. We resize the images to 96 × 96 and down-sample to 1,584 for each category as the balanced dataset. We have 3,168 images in total, which are split into training and testing sets by a ratio of 80 : 20.

3.1.3 Colorectal Histopathology

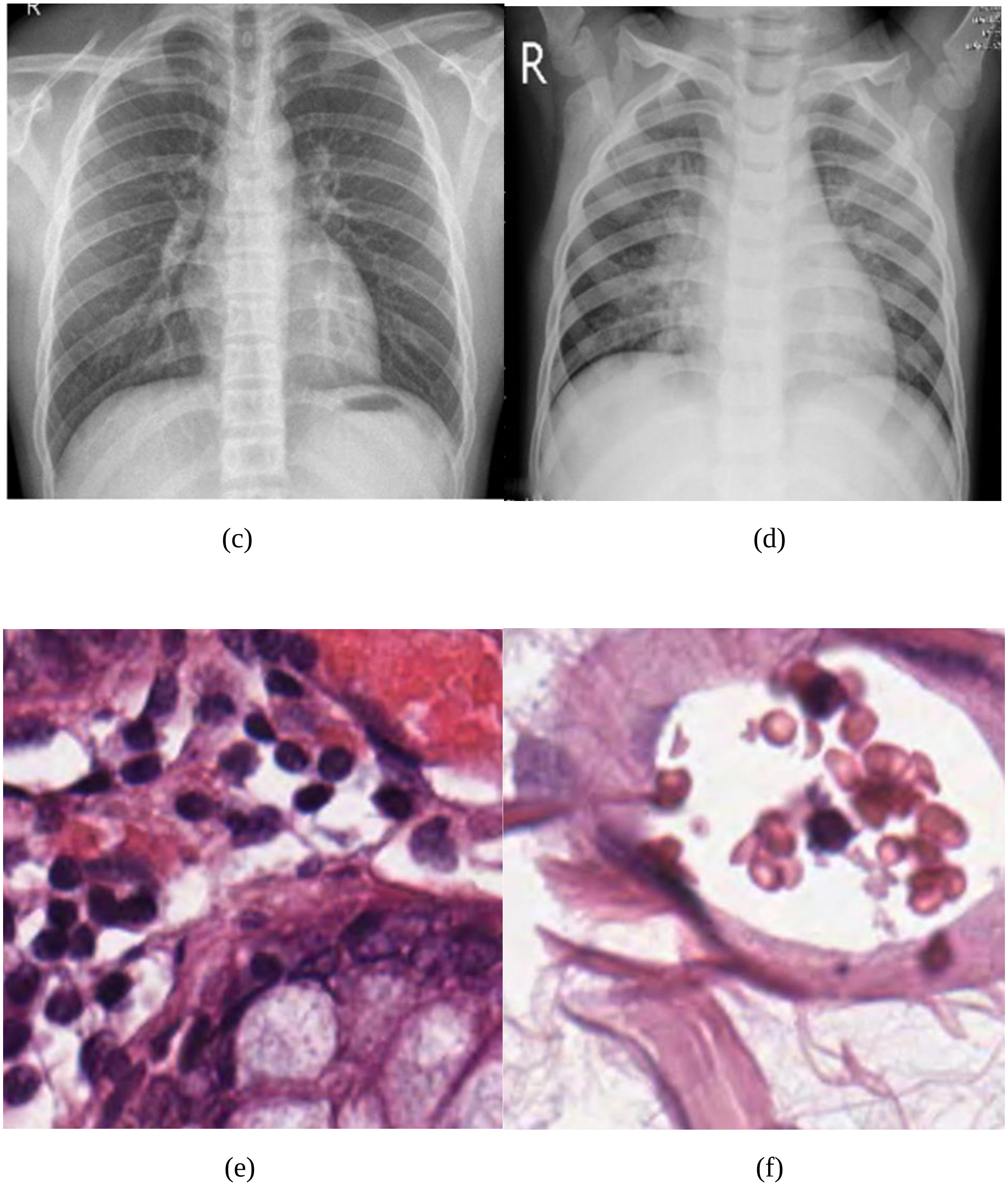
This dataset represents a collection of textures in histological images of human colorectal cancer [21]. Ten anonymized H&E stained CRC tissue slides are obtained from the pathology archive at the University Medical Center Mannheim, Heidelberg University, Mannheim, Germany. The low-grade and high-grade tumors are included in this set, and no further selection is applied. The slides are first digitized, and then the contiguous tissue areas are manually annotated and tessellated to create 625 non-overlapping tissue tiles of size 150 × 150 (74 μm × 74 μm). Thus, the texture features of different scales are included, ranging from individual cells (approximate 10 μm) to larger structures such as mucosal glands (>50 μm).

The following eight types of tissue are selected for analysis: tumor epithelium, simple stroma, complex stroma, immune cells, debris, normal mucosal gland, adipose tissue, and background (no tissue). Together, the resulting 5,000 images represent the training and testing sets. We randomly pick 2 classes, immune cells and normal mucosal glands, resize them to 96 × 96, and split train and test sets with a ratio of 80 : 20. Aside from the difference in imaging tissue and modality of these three data sets, the images are shown in Fig. 3.



(a)

(b)



(c) (d)

(e) (f)

Fig. 3. The sample images from three datasets. (a) Normal brain MRI, (b) abnormal brain MRI, (c) health chest X-ray, (d) pneumonia chest X-ray, (e) and (f) two different classes of human colorectal cancer, normal mucosal glands and immune cells.

3.2 Qualitative Analysis

3.2.1 Evaluation of the Test Accuracy

We first conduct experiments to compare the clean test accuracies of all seven models on chest X-ray, histopathology, and BraTs18. The results are listed in Table 1. On the Chest X-ray dataset, the convolutional networks LeNet [22] and ResNet18 [23] have higher accuracies since they have the advantage of convolutions. On the histopathology, the MLP and random forest [24] have higher accuracies. On BraTs18, the ResNet18, LeNet and random forest have higher accuracies, but other models are not too far behind.

Table 1: Average Accuracy of Validation Data on BraTs18, Chest X-ray and Histopathology Image Datasets

	SCD01	SCDCE	SCDCE BNN	MLP	LeNet	ResNet18	Random Forest
BraTs18	98.38%	98.92%	95.31%	98.76%	99.1%	99.64%	99.07%
Chest X-ray	90.69%	91.32%	89.12%	88.72%	92.59%	94.32%	89.12%
Histopathology	99.2%	99.6%	99.6%	100%	99.6%	99.6%	100%

3.2.2 Evaluation of the Defense Ability by L2 Distance

We compare the minimum distortion required to make an adversarial image on different models to evaluate the defense ability of adversarial attacks. The larger the value, the more robust the model since a significant distortion is likely to be detected in advance. Finding the exact minimum distortion is an NP-hard problem evaluated in ReLu activated neural networks [25, 26] and tree ensemble classifiers [27]. Even the approximation of the minimum distortion in ReLu activated neural networks is NP-hard [28].

The distortions reported by HopSkipJump have been shown to be lower (i.e., tighter and more accurate) than other boundary attack methods [29]. Therefore, we run the HopSkipJump boundary-based black-box attack [14] to determine the adversarial distortion of randomly selected images from the BraTs18, chest X-ray, and the colorectal cancer histopathology validation datasets. The HopSkipJump is run ten times on each image to report the minimum value.

As shown in Fig. 4, we observe that after 90 iterations, the distortions are minimum and become stable. Therefore, considering the best results and the computational ability, we pick 100 as the maximum iteration.

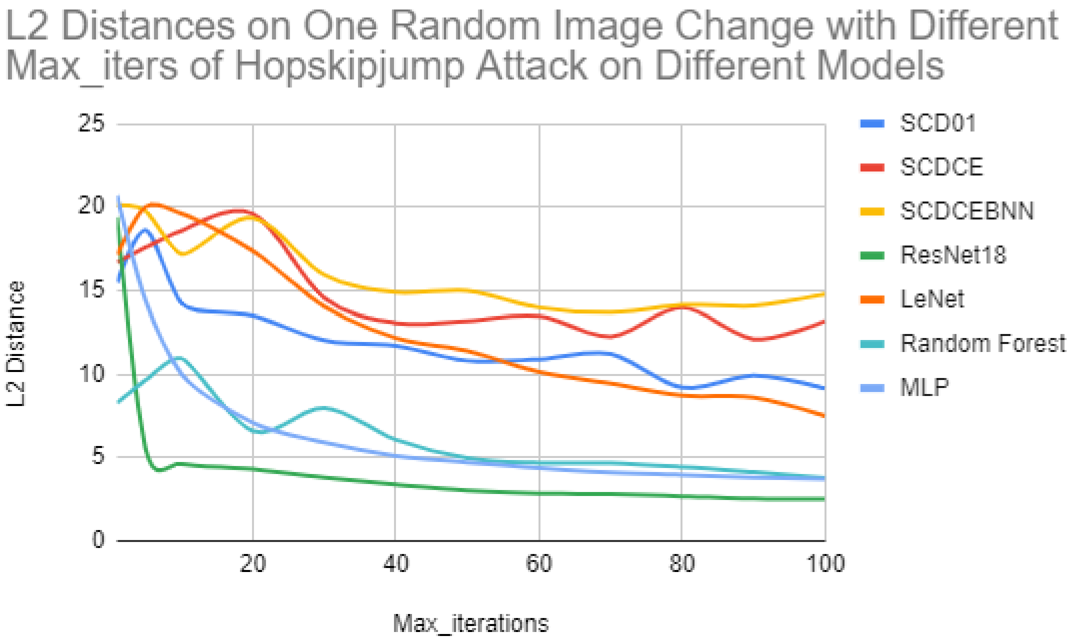


Fig. 4. L2 distances on one image change with different max iterations when Hopskipjump attack on different models.

We quantitatively measure the robustness of defense by measuring the distance between normal and abnormal samples under the L2 metric as most attacks did [30]. The Lp distance is the difference between original examples and adversarial examples, defined as $\|d\|_p$=$(\sum_{i=0}^{n}|vi|^{p})^{1/p}$. Common choices of p include: L0, a measure of the number of pixels changed; L2, the standard Euclidean norm; or L∞, a measure of the maximum absolute value change to any pixel. If the distortion under any of these three distance metrics is small, the images will likely appear visually similar.

Table 2 shows the average adversarial distortions of random test images from the BraTs18. The gradient free trained sign networks have the higher distortions than other state-of-the-art models, and the SCDCEBNN has the highest distortion.

Table 2: Average Minimum Estimated L2 Adversarial Distortion of on BraTs18 Datasets as Given by HopSkipJump When Attacking Different Models

	SCD01	SCDCE	SCDCEBNN	MLP	LeNet	ResNet18	Random Forest
Image 1	14.61	19.13	23.47	8.95	12.28	2.00	3.44
Image 2	10.55	13.44	16.18	4.32	9.06	1.95	4.03
Image 3	8.17	12.05	15.13	2.75	7.47	1.82	2.12
Image 4	7.49	13.00	3.33	3.67	7.50	2.50	3.33
Image 5	8.75	11.66	2.87	3.99	8.75	2.12	3.99
Average	9.06	12.23	13.7	4.38	8.27	2	2.78

We plot the original and adversarial images of “Image 1” from BtaTs18 dataset in Fig. 5 to get a visual feel for the distortions. The first six adversarial images have a high distortion, where (b) - (d) are the adversarial images from SCD models. Note that they have higher distortions than the currently available state-of-the-art models. Clearly, there are more noises than the original, while the other images are hard to observe the difference by human eyes.

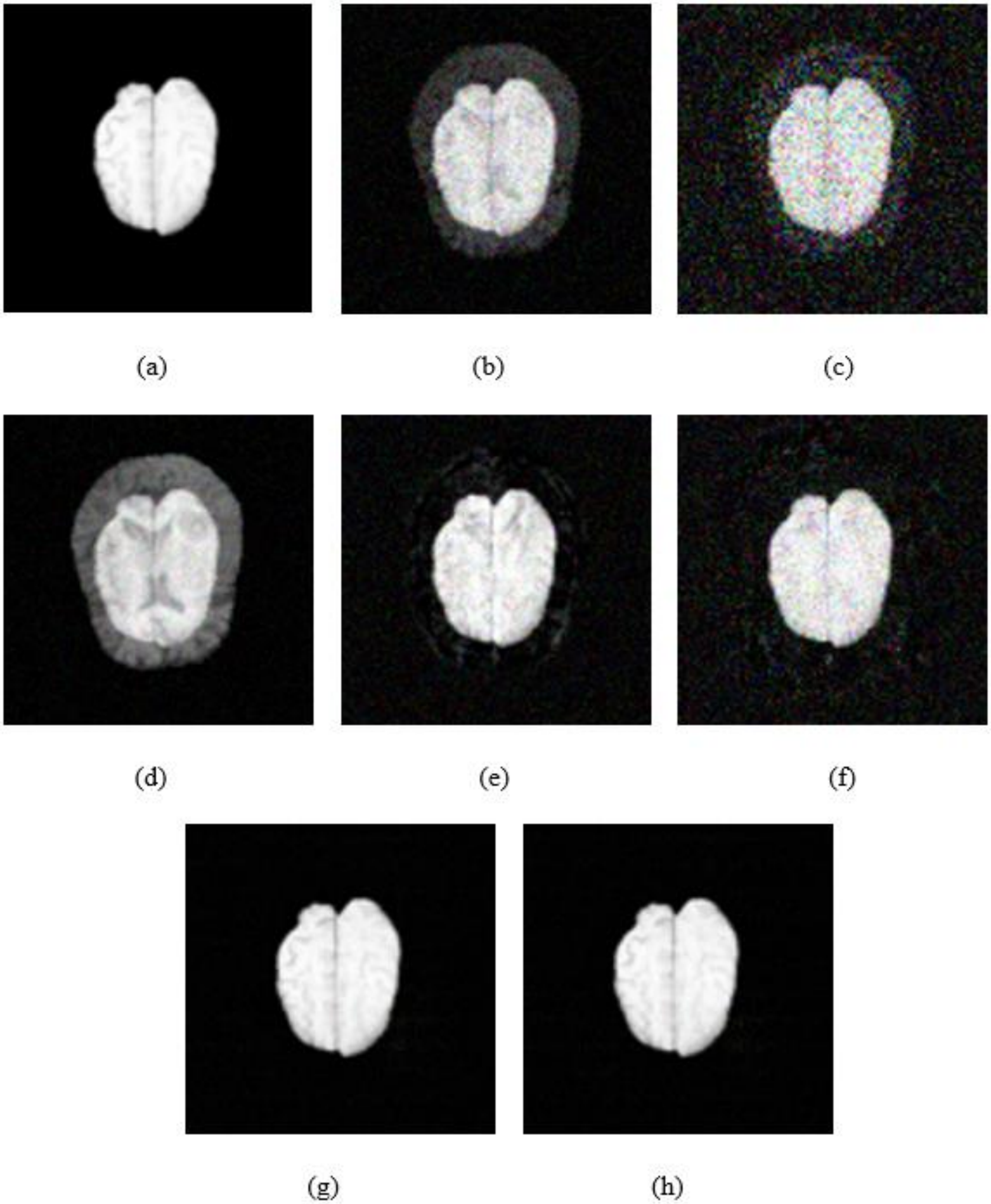


Fig. 5 Visualization of original images and adversarial images among different networks from BraTs18 dataset. (a) The original image, (b) the adversarial example which will fool SCD01, (c) the adversarial example which will fool SCDCE, (d) the adversarial example which will fool SCDCEBNN, (e) the adversarial example which will fool MLP, (f) the adversarial example which

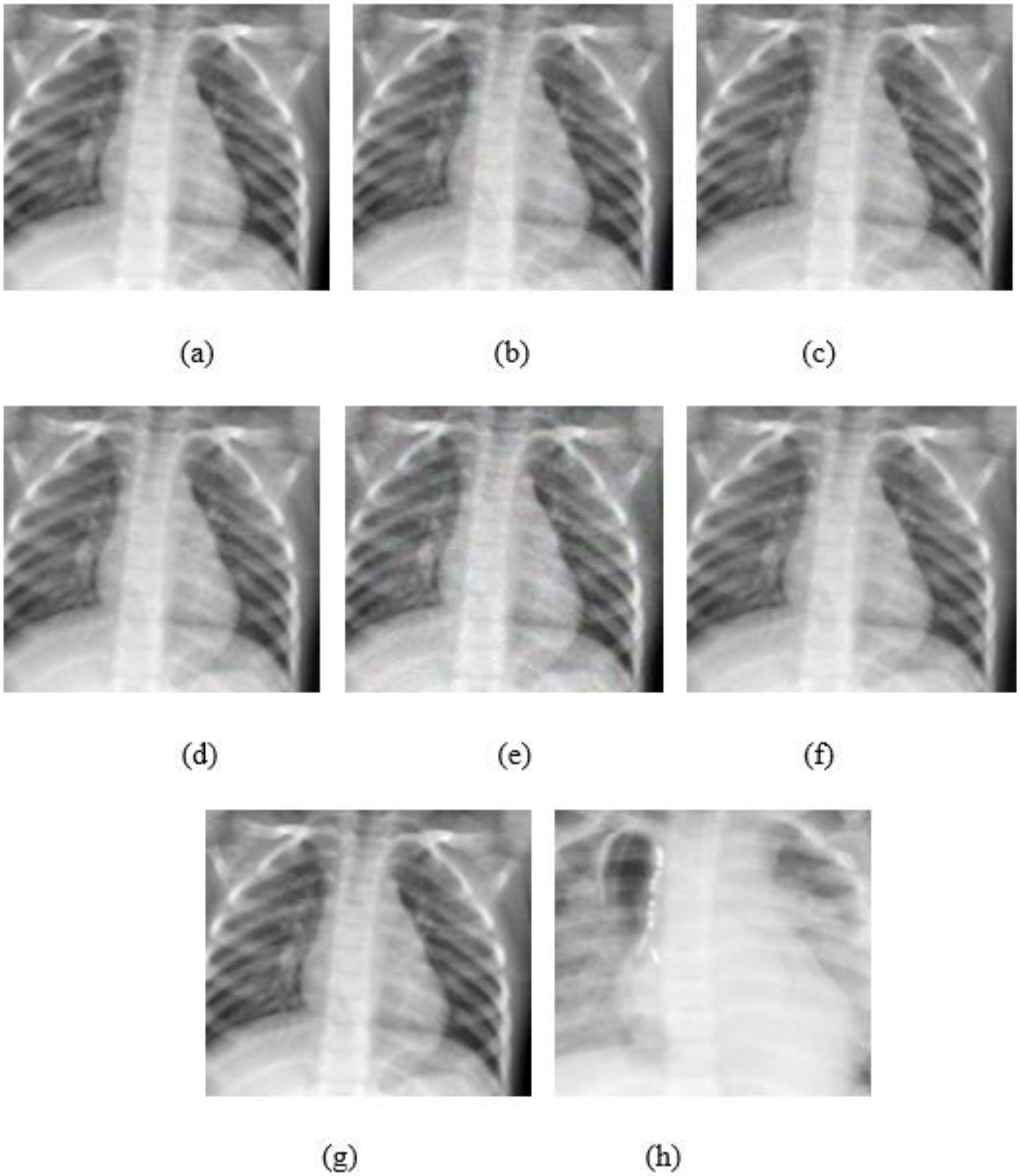
will fool LeNet, (g) the adversarial example which will fool Resnet18, and (h) the adversarial example which will fool Random Forest.

Table 3 lists the average adversarial distortions of random test images from the Chest X-ray dataset, where MLP is the second best after SCDCE.

Table 3: Average Minimum Estimated L2 Adversarial Distortion of on Chest X-ray Datasets as Given by HopSkipJump When Attacking Different Models

	SCD01	SCDCE	SCDCEBNN	MLP	LeNet	ResNet18	Random Forest
Image 1	10.59	18.40	17.50	14.78	4.28	1.03	18.53
Image 2	10.48	16.39	12.64	15.08	2.86	0.45	11.28
Image 3	9.00	17.55	10.50	14.49	4.19	0.64	9.18
Image 4	9.26	7.68	11.68	10.71	0.34	0.07	12.01
Image 5	7.24	14.02	10.16	12.91	4.10	0.43	2.39
Average	9.31	14.81	12.49	13.60	3.15	0.52	10.68

To get a visual feel for the distortions, Fig. 6 shows the original and adversarial images of “Image 1” from Chest X-ray dataset. They all have higher distortions, among which SCDCE has the highest.



(a) (b) (c)

(d) (e) (f)

(g) (h)

Fig. 6. Visualizations of original images and adversarial images among different networks from Chest X-ray dataset. (a) The original image, (b) the adversarial example which will fool SCD01, (c) the adversarial example which will fool SCDCE, (d) the adversarial example which will fool SCDCEBNN (e) the adversarial example which will fool MLP, (f) the adversarial example which will fool LeNet, (g) the adversarial example which will fool Resnet18, and (h) the adversarial example which will fool Random Forest.

Table 4 lists the average adversarial distortions of random test images from the colorectal dataset. The average distortion of SCDCEBNN is highest.

Table 4: Average Minimum Estimated L2 Adversarial Distortion of on Colorectal Cancer Histopathology Datasets as Given by HopSkipJump When Attacking Different Models

	SCD01	SCDCE	SCDCEBNN	MLP	LeNet	ResNet18	Random Forest
Image 1	28.3	41	41.32	9.9	29	31.6	19.9
Image 2	4.4	6.3	9.2	2.8	7	6.2	3.9
Image 3	35.8	36.1	44.71	9.9	36.8	39.8	30.4
Image 4	30	38.6	43.02	12	24.1	19.1	28.7
Image 5	17.2	26.5	28.97	7.7	17.1	19	13.4
Average	24.1	29.7	33.44	8.5	22.8	23.1	19.2

Fig. 7 shows the original image and adversarial images of human colorectal histopathology dataset which shows a visual feel for the distortions. All three SCD models have higher distortions than other models. Compared with other adversarial images, SCDCE adversary is full of more colorful ditties. The morphology is hard to identify such that it would be potentially abnormal.

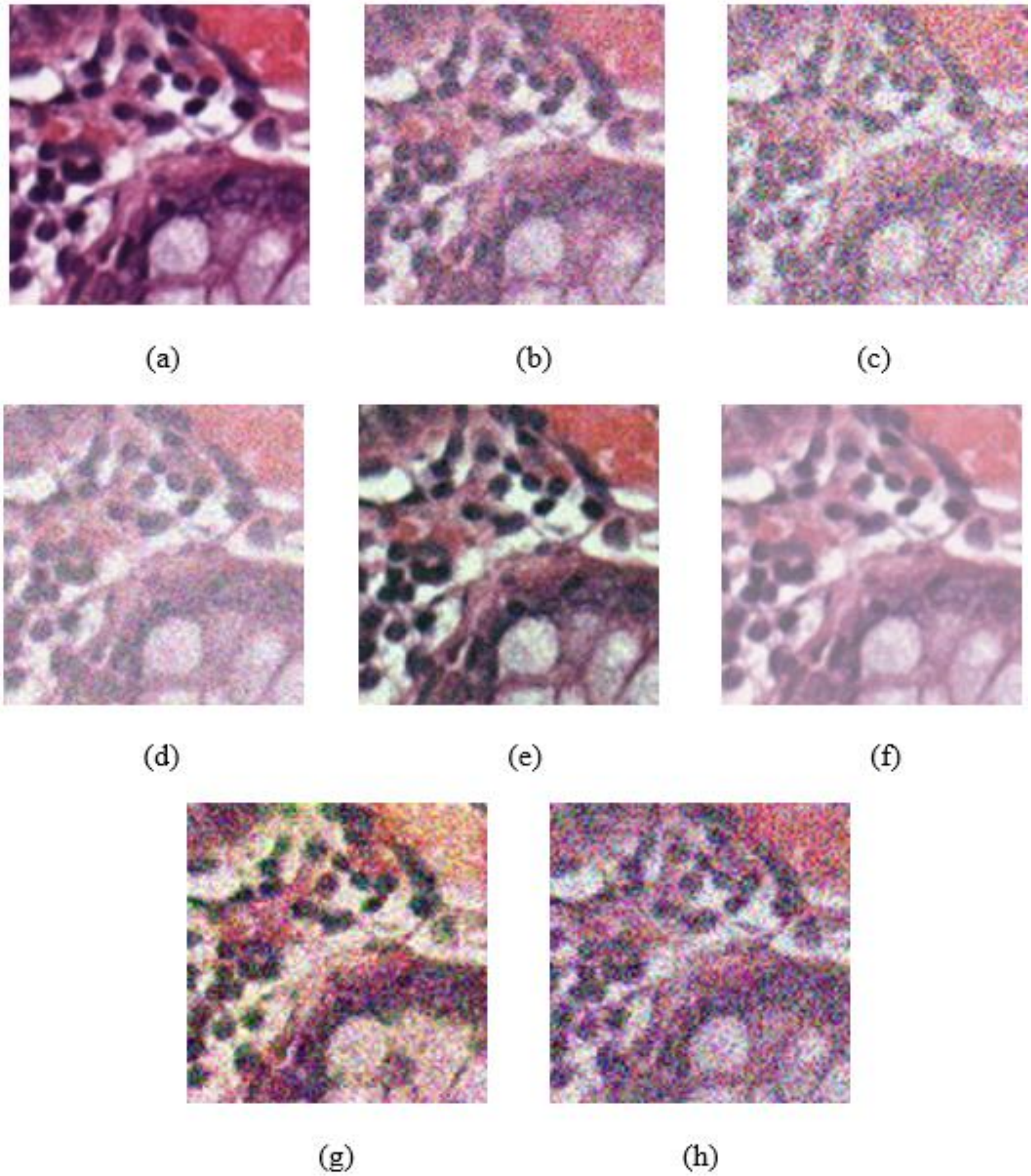


Fig. 7. Visualizations of original images and adversarial images among different networks from colorectal histopathology dataset. (a) The original image, (b) the adversarial example which will

fool SCD01, (c) the adversarial example which will fool SCDCE, (d) the adversarial example which will fool SCDCEBNN, (e) the adversarial example which will fool MLP, (c) the adversarial example which will fool LeNet, (e) the adversarial example which will fool ResNet18, (f) the adversarial example which will fool Random Forest.

Table 5 lists the average minimum estimated L2 adversarial distortion on all three datasets. The distortions of the SCD models are even higher with SCDCEBNN taking the lead and twice better than all other models.

Table 5 Average Minimum Estimated L2 Adversarial Distortion on All Three Datasets

	SCD01	SCDCE	SCDCEBNN	MLP	LeNet	ResNet18	Random Forest
Average	13.67	18.26	19.12	8.59	10.96	8.43	10.74

3.2.3 Evaluation of Defense Ability by Transferability

Another evaluation is to make use of the transferability property [10]. Given two models, F (·) and G(·), an adversarial example trained on F will be an adversarial example on G, even if they are trained in completely different manners or on different datasets. There has been a significant amount of available methods to construct adversarial examples [3, 10, 31-34] and to make networks robust against adversarial examples [35-38]. No defenses have been able to classify adversarial examples correctly. Thus, correctly classifying adversarial examples is difficult.

In the previous section, attackers generate adversarial samples on different models, and we test all these adversarial examples on all models. If a model G can detect the adversarial examples from another model F and classify them correctly, the model G is more robust against adversarial attack. Table 6 shows the results for classifying one random image and all adversarial examples. We can see that a random image can be classified by all models, which is marked as 'Y.' The targeting adversarial samples are misclassified by their targeting models, respectively. If the model can identify the adversarial example correctly, it is marked as 'Y,' otherwise; it is marked as 'N.' Our models 01MLP and SCDCE can detect all adversarial examples and classify them correctly.

Table 6: Results for Classifying One Random Image and All Adversarial Examples

	SCD01	SCDCE	SCDCE BNN	MLP	LeNet	ResNet18	Random Forest
Original Test Image	Y	Y	Y	Y	Y	Y	Y
Adversarial Image from SCD01	-	Y	Y	Y	Y	N	Y
Adversarial Image from SCDCE	Y	-	Y	N	Y	N	Y
Adversarial Image from SCDCEBNN	Y	Y	-	Y	N	N	N
Adversarial Image from MLP	Y	Y	Y	-	Y	N	Y
Adversarial Image from LeNet	Y	Y	Y	Y	-	N	Y
Adversarial Image from ResNet18	Y	Y	Y	Y	Y	-	Y
Adversarial Image from Random Forest	Y	Y	Y	Y	Y	N	-

Table 7 shows the average accuracy of all models when classifying the adversarial examples. We can see that our proposed models have higher accuracies, which are 88.89% and 85.19%. They can identify fake examples and are hard to be fooled by adversarial attacks. In other models like MLP and LeNet are the best but are still lower than our proposed models.

Table 7. Average Accuracy of All Models When Classifying the Adversarial Examples

	SCD01	SCDCE	SCDCEBNN	MLP	LeNet	ResNet18	Random Forest
Average Accuracy	88.89%	88.89%	85.19%	81.48%	81.48%	38.89%	57.14%

4. Conclusions

In this paper, we present a model that is robust to adversarial attacks in MRI images, chest X-ray and histopathology images. We show that higher distortions are required when adversarial attacking is applied on the gradient-free trained sign networks with SCD compared with state-of-the-art models. Experimental results on classifying the adversarial samples show that our models' accuracy is more competitive, and thus, the adversarial attack can easily be detected on our models. To develop a robust medical machine learning models which can deter attack in advance, more research is required to verify the results on a larger cohort and show the results on the different adversarial attack, such as white-box. We plan to develop a medical AI imaging system which can detect and deter adversarial attack in advance in the future work.

REFERENCES

[1] C. Yeh, M. Lin, P. Chang and L. Kang, "Enhanced Visual Attention-Guided Deep Neural Networks for Image Classification," IEEE Access, vol. 8, pp. 163447-163457, 2020.

[2] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z.B. Celik, and A. Swami, “The limitations of deep learning in adversarial settings,” *Proc. IEEE European Symposium on Security and Privacy*, Saarbrücken, Germany, pp. 372–387, March 2016.

[3] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” *Proc. IEEE Symposium on Security and Privacy*, San Jose, CA pp. 39–57. March 2017.

[4] N. Papernot, P. McDaniel, and I. Goodfellow, “Transferability in machine learning: from phenomena to black-box attacks using adversarial samples,” arXiv preprint arXiv:1605.07277, 2016.

[5] P. Panda, I. Chakraborty and K. Roy, “Discretization based solutions for secure machine learning against adversarial attacks,” *IEEE Access*, vol. 7, pp. 70157–70168, 2019.

[6] W. Brendel, J. Rauber, and M. Bethge, “Decision-based adversarial attacks: reliable attacks against black-box machine learning models,” arXiv preprint arXiv:1712.04248, 2017.

[7] S.G. Finlayson, J.D. Bowers, J. Ito, J.L. Zittrain, A.L. Beam, and I.S. Kohane, “Adversarial attacks on medical machine learning,” *Science*, vol. 363, no. 6433, pp.1287–1289, March 2019.

[8] S. C. Wetstein, et al., “Adversarial attack vulnerability of medical image analysis systems: Unexplored factors,” arXiv preprint arXiv:2006.06356, 2020.

[9] M. Paschali, S. Conjeti, F. Navarro, and N. Navab, “Generalizability vs. robustness: investigating medical imaging networks using adversarial examples,” *Proc. Conference on Medical Image Computing and Computer Assisted Intervention*, Granada, Spain, pp. 493–501, September 2018.

[10] H. Hirano, A. Minagi, D. Soudry, and K. Takemoto, “Universal adversarial attacks on deep neural networks for medical image classification,” *BMC medical imaging*, pp.1-13, January 2021.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *Proc. 3th International Conference on Learning Representations*, ICLR, San Diego, CA, May 2015.

[12] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, “Ensemble adversarial training: attacks and defenses,” *Proc. 6th Intl. Conf. on Learning Representations*, Vancouver, BC, Canada, May 2018.

[13] Y. Dong, T. Pang, H. Su, and J. Zhu, “Evading defenses to transferable adversarial examples by translation-invariant attacks,” *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, California, pp. 4312–4321, June 2019.

[14] J. Chen, M. I. Jordan, M. J. Wainwright, “Hopskipjump attack: a query-efficient decision-based attack,” *Proc. IEEE Symposium on Security and Privacy*, San Francisco, CA, pp. 1277-1294, May 2020.

[15] M. Xie, Y. Xue, and U. Roshan, “Stochastic coordinate descent for 0/1 loss and its sensitivity to adversarial attacks,” *Proc. 18th IEEE Intl. Conf. on Machine Learning and Applications*, pp. 299-304, 2019.

[16] Y. Xue, M. Xie, and U. Roshan, “Towards adversarial robustness with 01 loss neural networks,” *Proc. 19th IEEE Intl. Conf. on Machine Learning and Applications*, Miami, FL, 2020.

[17] Y. Xue, M. Xie, and U. Roshan, “On the transferability of adversarial examples between convex and 01 loss models,” *Proc. IEEE Intl. Conf. on Machine Learning and Applications*, Miami, FL, 2020.

[18] Z. Yang, Y. Yang, Y. Xue, F.Y. Shih, J. Ady, and U. Roshan, “Accurate and adversarially robust classification of medical images and ECG time-series with gradient-free trained sign activation neural networks,” *Proc. IEEE Intl. Conf. on Bioinformatics and Biomedicine*, Seoul, South Korea, pp. 2456-2460, 2020.

[19] A. Paszke, et al, “PyTorch: an imperative style, high-performance deep learning library,” *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, pp: 8024-8035, December 2019

[20] S. Bakes, et al., “Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge,” arXiv preprint arXiv:1811.02629, 2018.

[21] J. N. Kather, et al., “Multi-class texture analysis in colorectal cancer histology,” Scientific Reports, Rep. 6, 27988, March 2016.

[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. of the IEEE*, vol. 86, no.11, pp. 2278–2324, 1998.

[23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, pp. 770–778, June 2016.

[24] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001.

[25] G. Katz, C. Barrett, D. L. Dill, K. Julian, M. and J. Kochenderfer, “Reluplex: an efficient smt solver for verifying deep neural networks,” *Proc. Intl. Conf. on Computer Aided Verification*, Heidelberg, Germany, pp. 97–117, July 2017.

[26] A. Sinha, H. Namkoong, and J. Duchi, “Certifiable distributional robustness with principled adversarial training,” arXiv preprint arXiv:1710.10571, 2017.

[27] A. Kantchelian, J. D. Tygar, and A. Joseph, “Evasion and hardening of tree ensemble classifiers,” *Proc. Intl. Conf. on Machine Learning*, New York City, New York, vol. 48, pp. 2387–2396, June 2016.

[28] T. Weng, H. Zhang, H. Chen, Z. Song, C. Hsieh, D. Boning, I. S. Dhillon, and L. Daniel. “Towards fast computation of certified robustness for relu networks,” arXiv preprint arXiv:1804.09699, 2018.

[29] M. Nicolae, et al., “Adversarial robustness toolbox” v1.0.0. arXiv preprint arXiv:1807.01069, 2018.

[30] F. Tramer and D. Boneh, “Adversarial training and robustness for multiple perturbations,” *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, pp. 5858-5868, December 2019.

[31] B. Biggio, et al., “Evasion attacks against machine learning at test time,” *Proc. Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Würzburg, Germany, pp. 387-402, September 2013.

[32] S.M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: a simple and accurate method to fool deep neural networks,” *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, pp. 2574-2582, June 2016.

[33] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, “The limitations of deep learning in adversarial settings,” *Proc. IEEE European Symposium on Security and Privacy*, Saarbrücken, Germany, pp. 372-387, March 2016.

[34] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” arXiv preprint arXiv:1312.6199, 2013.

[35] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, “Distillation as a defense to adversarial perturbations against deep neural networks,” *Proc. IEEE Symposium on Security and Privacy*, San Jose, CA, pp. 582-597, May 2016.

[36] A. Rozsa, E. M. Rudd, and T. E. Boult, “Adversarial diversity and hard positive generation,” *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, pp. 410-417, June 2016.

[37] U. Shaham, Y. Yamada, and S. Negahban, “Understanding adversarial training: increasing local stability of supervised models through robust optimization,” *Neurocomputing*, vol. 307, pp. 195-204, September 2018.

[38] S. Zheng, Y. Song, T. Leung, and I. Goodfellow, “Improving the robustness of deep neural networks via stability training,” *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, pp. 4480-4488, June 2016.