5-31-2022

Adversarially robust and accurate machine learning for image classification

Yanan Yang
New Jersey Institute of Technology

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

ADVERSARIALLY ROBUST AND ACCURATE MACHINE LEARNING FOR IMAGE CLASSIFICATION

by
Yanan Yang

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. This dissertation presents the gradient-free trained sign activation networks to detect and deter adversarial attacks on medical imaging AI (Artificial Intelligence) systems. Experimental results show a higher distortion value is required to attack the proposed model than other state-of-the-art models on brain MRI (Magnetic resonance imaging), Chest X-ray, and histopathology image datasets. Moreover, the proposed models outperform the best existing models and are 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 33.89%. It is concluded that the sign network is a solution to defend against adversarial attacks due to high distortion and high accuracy on transferability. In addition, different models have different tolerance abilities on adversarial attacks.

This dissertation develops a novel detecting module to defend against adversarial attacks proactively. The proposed module uses the adaptive noise removal process to reconstruct the input and detect adversarial attacks without modifying the models. Experimental results show that the proposed models can successfully remove most noises and obtain detection accuracies of 97.71% and 92.96%, respectively, by comparing the

classification results on adversarial samples of MNIST and two subclasses of ImageNet datasets. Furthermore, the proposed adaptive module can be used as part of an ensemble with different networks to achieve detection accuracies of 70.83% and 71.96%, respectively, on the white-box adversarial attacks of ResNet18 and SCD01MLP. The best accuracy of 62.5% is obtained for both networks when dealing with black-box attacks.

ADVERSARIALLY ROBUST AND ACCURATE MACHINE LEARNING FOR IMAGE CLASSIFICATION

by
Yanan Yang

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

Department of Computer Science

May 2022

.

APPROVAL PAGE

ADVERSARIALLY ROBUST AND ACCURATE MACHINE LEARNING FOR IMAGE CLASSIFICATION
Yanan Yang

Dr. Frank Y. Shih, Dissertation Advisor — Date
Professor of Computer Science, NJIT

Dr. Zhi Wei, Committee Member — Date
Professor of Computer Science, NJIT

Dr. Usman W. Roshan, Committee Member — Date
Associate Professor of Computer Science, NJIT

Dr. Dimitri Theodoratos, Committee Member — Date
Associate Professor of Computer Science, NJIT

Dr. William Graves, Committee Member — Date
Associate Professor of Psychology, Rutgers University

BIOGRAPHICAL SKETCH

Author: Yanan Yang

Degree: Doctor of Philosophy

Date: May 2022

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2022

- Master of Science in Computer Science,
  New Jersey Institute of Technology, Newark, NJ, 2015

- Bachelor of Science,
  Hubei University, Wuhan, Hubei, P. R. China, 2013

Major: Computer Science

Presentations and Publications:

Y.Yang, F. Y. Shih, U. Roshan, "Defense against adversarial attacks based on stochastic descent sign activation networks on medical images," *International Journal of Pattern Recognition and Artificial Intelligence*, 2022.

Z. Yang, Y. Yang, Y. Xue, F. Y Shih, J. Ady, U. Roshan, "Accurate and adversarially robust classification of medical images and ECG time-series with gradient-free trained sign activation neural networks," *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2020.

Y. Yang, F. G. Farhat, Y. Xue, F. Y. Shih, U. Roshan, "Classification of histopathology images with random depthwise convolutional neural networks," *7th International Conference on Bioinformatics Research and Applications*, 2020.

Y. Xue, Y. Yang, F. Farhat, F. Y. Shih, O. Boukrina, A. M. Barrett, J. R. Binder, U. W. Roshan, W. W. Graves, "Brain tumor classification with tumor segmentations and a dual path residual convolutional neural network from MRI and pathology images," *5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019 (The Medical Image Computing and Computer Assisted Intervention Society)*, Shenzhen, China, pp. 360-367, Oct, 2019.

To my family.

Mr. Ruiling Yang, Mrs. Hulian Chen, Ms. Yaqin Yang,

who have given me invaluable educational opportunities, always loved me unconditionally, been my emotional ancho, and whose good examples have taught me to work hard for the things that I aspire to achieve. I am truly thankful for having you in my life.

ACKNOWLEDGMENT

It is a genuine pleasure to express my deep sense of thanks and gratitude to my advisor, colleague, and my friend forever, Dr. Frank Y. Shih. Your rigorousness, passion, and perseverance in the research process have shown me a scientific researcher's outstanding character. Behind is the story of us working together towards the same goal, bravely challenging problems, and committing impactful work.

I would like to offer my sincerest thanks and praise to the dissertation committee members: Dr. Zhi Wei, Dr. Usman W. Roshan, Dr. Dimitri Theodoratos, and Dr. William Graves. Thank you for your guidance on my research. I truly enjoy working with you.

Special thanks to Dr. Roshan and Dr. Graves, for providing guidance and feedback throughout these projects, without whom I would never understand medical images and have no content for my work.

A big thanks also go out to the Computer Science Department for providing me with financial support. Their support was essential to undertaking my studies. Furthermore, I would like to thank NJIT IST Academic and Research Computing Systems (ARCS) administrators for providing reliable service and high-performance computing resources.

I would like to express my sincere gratitude to my labmates: Shaobo Liu, Yunzhe Xue, Meiyan Xie, and Fadi Farhat, for your technical support and encouragement during these four years.

Finally, I heartily appreciate my family: Mr. Ruiling Yang, Mrs. Hulian Chen, Mrs. Yaqin Yang, and Mrs. Shiran Xiao; my best friends: Mrs. Qianwen Ye, Ms. Cuicui Zheng, Mr. Xuan Li, Ms. XinYin, and Mr. YaoXing Li. You are my strong backing, the source of my strength, and my harbor of tenderness.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF TABLES
(Continued)

Table **Page**

LIST OF FIGURES

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Objective

This dissertation aims to present a set of accurate and adversarially robust deep learning models for medical image datasets and an adaptive detection scheme to defend against adversarial attacks.

The following accurate deep learning models for medical image datasets are presented, the dual path residual convolutional neural network and the random depthwise convolutional neural network. The classification performance analysis shows that both networks have higher accuracy when trained on medical image datasets.

Furthermore, the stochastic descent sign activation networks are implemented to defend against adversarial attacks, which are SCD01, SCDCE, and SCDCEBNN. The different evaluation experiments suggest that the SCD models have the more powerful defense ability compared with the state-of-the-art.

Last but not least, an adaptive detection scheme with an adaptive image reconstruction algorithm is deployed to defend against adversarial attacks more actively. This proposed module has a competitive detection rate and can ensemble into any model.

1.2 Background

Machine learning is a branch of artificial intelligence, which can learn the structures from data, identify patterns and make decisions with minimal human intervention. Many machine learning applications have been developed for a long time. They have achieved

success in many fields such as image recognition, video detection, autonomous driving, voice recognition, and fraud detection. Some widely used machine learning algorithms include linear regression, logistic regression, support vector machine(SVM), multiple layer perceptron (MLP), decision tree, random forest, and neural networks.

Deep learning has emerged in the past few decades as a new approach is shown in machine learning, such as Google's latest automatic translator, which achieved impressive results in handling large amounts of data. Deep learning models use the convolutional neural network (CNN). The role of "convolution" can reduce the data dimensions into a more accessible processing form while keeping the features used for a better prediction.

LeNet [1] was first introduced by LeCun and Bengio in 1995. This network is a straightforward convolutional neural network and is used to learn the complex, high-dimensional, and non-linear mappings from the sizeable data collections. In 2012, the CNN became known as AlexNet [2], which was developed by Krizhevsky *et al*, and won the 2012 ImageNet vision contest with an impressive 85% accuracy. From AlexNet, the state-of-the-art CNN architectures are going deeper and deeper. In 2014, two deep learning networks were proposed, one is GoogLeNet [3] with 22 layers, and another is VGG [4] with 19 layers, while the AlexNet [2] has only five layers. GoogLeNet and VGG both have more outstanding performances on classification than previous models.

However, increasing the depth of the network brings the vanishing gradient problem, which causes deep networks to be hard to train. Microsoft Research Asian developed ResNet [5] that achieved 96.4% accuracy on ImageNet competition in 2015. The classification rate is higher than GoogLeNet [3] and AlexNet [2]. Moreover, the most exciting thing is that the residual block avoids the vanishing gradient problem.

The neural networks also show more impressive performances than clinicians in many healthcare tasks. For example, medical images, such as magnetic resonance imaging (MRI), computational tomography (CT), and histopathology provide detailed information for diagnosing various diseases. Medical experts typically have to browse numerous images to diagnose diseases, which requires considerable training. Furthermore, the diagnosis process is time-intensive and is prone to manual errors. The deep learning methods can help the experts decide and accelerate treatment processes.

Recent research has shown that deep learning has a high accuracy, reasonable prediction, and high sensitivity to automatical medical tasks. Since 2016, the International Skin Imaging Collaboration (ISIC) has begun to aggregate a large-scale publicly accessible dataset of dermoscopy images and hosted the challenges on disease classification and segmentations [5]. The top-ranked participant in 2017, Yan *et al* implemented a fully convolutional network ensemble approach to achieve an average accuracy of 93.4% and a Dice coefficient of 0.849 on Lesion segmentation [6]. On average classification performance characteristics, the top three winners are 91.1%, 91%, and 90.8%, respectively [7-9]. Oktay *et al* [10] proposed an anatomically constrained convolutional neural network (ACNN) model according cardiac anatomy, a generic training strategy for super-resolution medical images. As a result, the classification accuracy on the cardiac MR dataset was up to 91.6%.

Deep learning models train algorithms efficiently to outperform other approaches on medical tasks. However, it is worth noting that deep neural networks are vulnerable to adversarial examples, which are crafted by adding imperceptible perturbations on the original images. Adversarial attacks bring some unforeseen losses in the real world. For

instance, an attacker might manipulate their examination report on a computer to cause a misdiagnosis of the disease, a false medical reimbursement claim, or commit insurance fraud. Much research has shown that adversarial attacks can force deep learning neural networks to make a wrong decision [11-14]. Therefore, the robustness and security of machine learning are still open problems.

Researchers investigated adversarial attacks on medical images and mainly focused on testing the robustness of deep learning models for medical image analysis [15, 16]. Paschali *et al* [17] showed that classification accuracy drops from 87% on the regular medical images to almost 0% on the adversarial examples. Hokuto *et al* [18] demonstrated that their attack method achieved over 80% success rates on the deep learning models. Goodfellow *et al* [19] presented that an image of adding imperceptible perturbations can be misclassified with very high confidence by GoogLeNet. Papernot *et al* [20] showed that a crafted stop sign is incorrectly classified as a yield sign.

All of the above studies indicate that the adversaries can potentially use the crafted images to inflict severe damage. The success of adversarial attacks leads to security threats. It is crucial to ensure that neural networks detect abnormal inputs more safely and securely. In other words, the robustness of deep learning algorithms needs to be reevaluated before deploying them in the real world.

To defend against adversarial attacks, some researchers proposed the techniques of improving model robustness and detecting malicious behaviors. However, most of them require modifying the target model. One of the most straightforward methods is called *adversarial training*, which uses as many adversarial examples as possible to retrain the network and improve classification accuracy [21-23]. Papernot *et al* [24] introduced

another defense technique, named *defensive distillation*, to train two networks as a distillation. The first network produces probability vectors to label the original dataset, while the newly labeled dataset is used to train the other network.

Recent studies have been focused on detecting adversarial examples directly. Xu *et al* [25] proposed training the detection modules using a set of adversarial examples and the corresponding benign ones. However, this technique requires modifying the model and a sufficient number of adversarial examples.

Although the above researchers have shown definite achievements in defending against adversarial attacks, the cost of training time and the computational ability to retrain the models and modify the network architectures are impractical. Generating appropriate adversarial examples for training and statistical testing is quite expensive, and the successful results depend on a comprehensive prior knowledge of various potential adversarial techniques. Even worse, attackers can compose adversarial examples by the revamped attack algorithms, usually unknown to the defender in advance. In this way, there is a good chance for the adversarial examples to evade the classifier. Moreover, most of the existing defense techniques are model-specific. Rebuilding or retraining a classifier would consume significant extra time. Therefore, attackers can easily craft efficacious fake examples.

Other methods intend to reconstruct the input images to defend against adversarial attacks. Liao *et al* [26] proposed a pixel denoiser method to remove the noise on high-level representation. As adversarial noises also interfere with the features constructed on the networks, Xie *et al* [27] developed a feature denoising architecture to smooth input images by applying non-local means or other filters directly on the feature level. Zhang *et al* [28]

proposed an image reconstruction network based on residual block structures to suppress the noises. Zhao *et a*. [29] found that pixels can be divided into the low-sensitivity and the high-sensitive groups based on the contributions to image classification. Inspired by that observation, they developed a structure-preserving low-rank image completion method on high sensitivity points to remove noises.

However, there exists a problem with these techniques. Reconstruction and denoise methods apply a smoothing filter on the input images to reduce noises but inevitably remove some details on the object's interest. Different adversary attackers magnify different scales of perturbations even on the same images. The perturbations on a compelling adversarial image from the semantic and grayscale datasets are very different for attackers. In other words, using the same parameters on an adversarial attack to generate an adversarial sample on ImageNet images cannot generate a successful one from MNIST. As a result, using the same reconstructed strategy may overly reduce the noise and lead to a new misclassification.

This dissertation first evaluates the classifying performances of machine learning on medical images. Two networks are presented. A dual path residual convolutional neural network is proposed for classifying brain tumor types. The model is trained simultaneously from both MRI and pathology images and achieves a validation accuracy of 84.9%. In addition, a depthwise convolutional neural network with random weights (RDCNN) is investigated on four popular open-source medical datasets. The experimental results show that the proposed model has a 95% average accuracy across all datasets, higher than state-of-the-art models.

Furthermore, this work presents the sign activation networks with a gradient-free stochastic coordinate descent algorithm (SCD) to address the earlier secure machine learning challenges. Experimental results show a higher distortion value is required to attack the proposed model than other state-of-the-art models on MRI, Chest X-ray, ECG, and histopathology image datasets. Moreover, the average accuracy of SCD models in classifying the adversarial examples is 88.89%, which outperforms the best and even twice superior.

Based on the preliminary experiments, a novel defend scheme is proposed to proactively defend against the adversarial attack, which reconstructs the input image with an adaptive method to avoid excessive noise reduction and effectively detects the adversarial samples in advance. Futhermore, the framework is lightweight to make an ensemble into any network, and only a few arguments need to be set up. Experimental results show that the proposed model can successfully remove most noises and obtain higher detection accuracy by comparing the classification results on different adversarial attackers' samples.

CHAPTER 2

A DUAL PATH RESIDUAL CONVOLUTIONAL NEURAL NETWORK FROM MRI AND PATHOLOGY IMAGES

2.1 Related Work on Brain Tumor Segmentation and Classification

Brain cancer tumors fall into different categories given by the World Health Organization [30–32]. Predicting tumor types correctly plays a crucial role in diagnosis and treatment. The automated classification of tumor types can significantly speed up physician diagnosis and give patients better care and treatment.

The CPM-RadPath 2019 MICCAI challenge is to predict three tumor types automatically, and the contest provides MRI and pathology images simultaneously. Figure 2.1 shows a cropped pathology image with a Grade IV tumor (class G) and a radiology image (MRI) from this challenge.

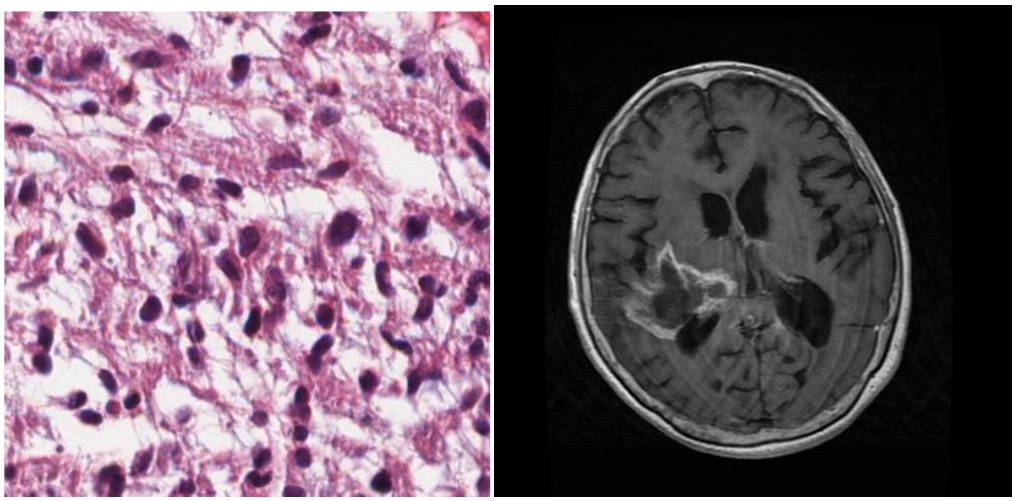


Figure 2.1 A typical cropped pathology image with a Grade IV tumor (class G) (left) and a radiology image (right) taken from the CPM-RadPath dataset.

Brain tumors' classification and segmentation are challenging tasks because MRI scans share a highly heterogeneous appearance and shape. Although increasing scientific

literature on machine learning algorithms address this critical task, comparing the different segmentation strategies that have been reported so far is problematic. There are many reasons. For example, the private datasets differ widely, and there are rarely open manually-annotated datasets for designing and testing on machine learning algorithms.

The top-ranked schemes on brain tumor segmentation in the BraTS 2017 and BraTS 2018 challenges are based on machine learning algorithms, particularly the convolutional neural networks [33]. Kamnitsas *et al* ensemble multiple convolutional neural networks and achieved the highest accuracies in 2017 challenges. The winner in the 2018 contest designed a convolution network with 32 filters.

Inspired by the success of convolutional neural networks in image recognition tasks, a dual path residual convolutional neural network solution is proposed to solve the prediction of tumor type problems. The preliminary experiments using predicted tumor segmentation of each MRI image show higher overall validation accuracy than the MRI images without masks.

2.2 Proposed Networks

2.2.1 Custom Designed U-Network for Predicting Tumor Segmentations

Figure 2.2 shows custom designed U-Network to predict tumor segmentation from MRI images [34]. The proposed network takes images in four modalities and is trained on the Brain Tumor Segmentation (BraTS) 2019 MICCAI challenge [33, 35].

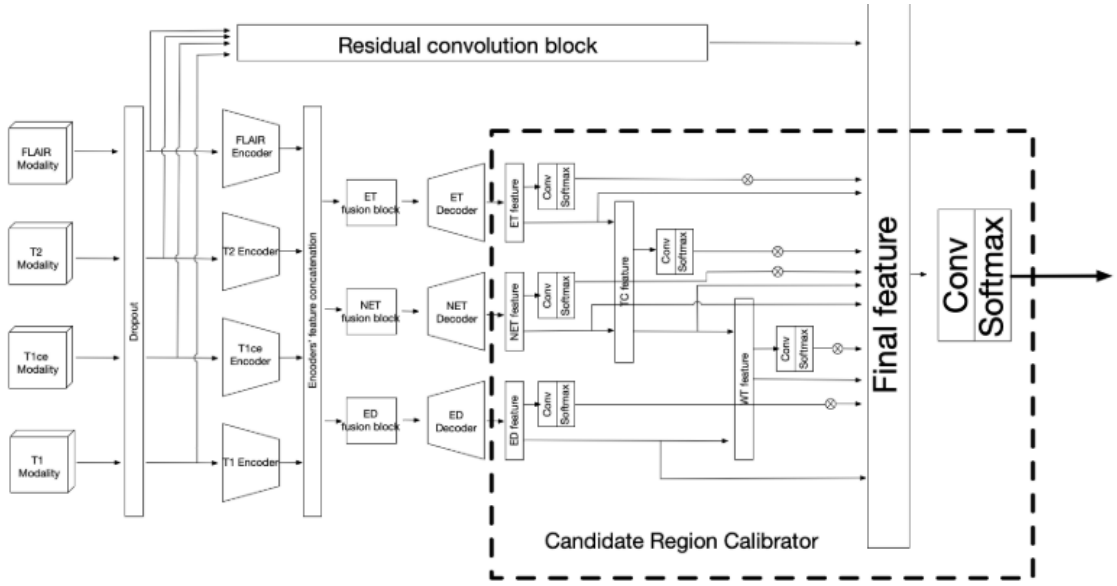


Figure 2.2 The architecture of custom-designed multi-modal tumor segmentation network.

2.2.2 Dual Path Residual Convolutional Neural Network

The ResNet18 architecture [5] uses residual connections between layers to prevent gradient vanishing problems and is a highly successful approach. Figures 2.3 (a) and (b) show the ResNet18 convolutional neural network architectures used separately on MRI and pathology images. The dual path model is in Figure 2.4.

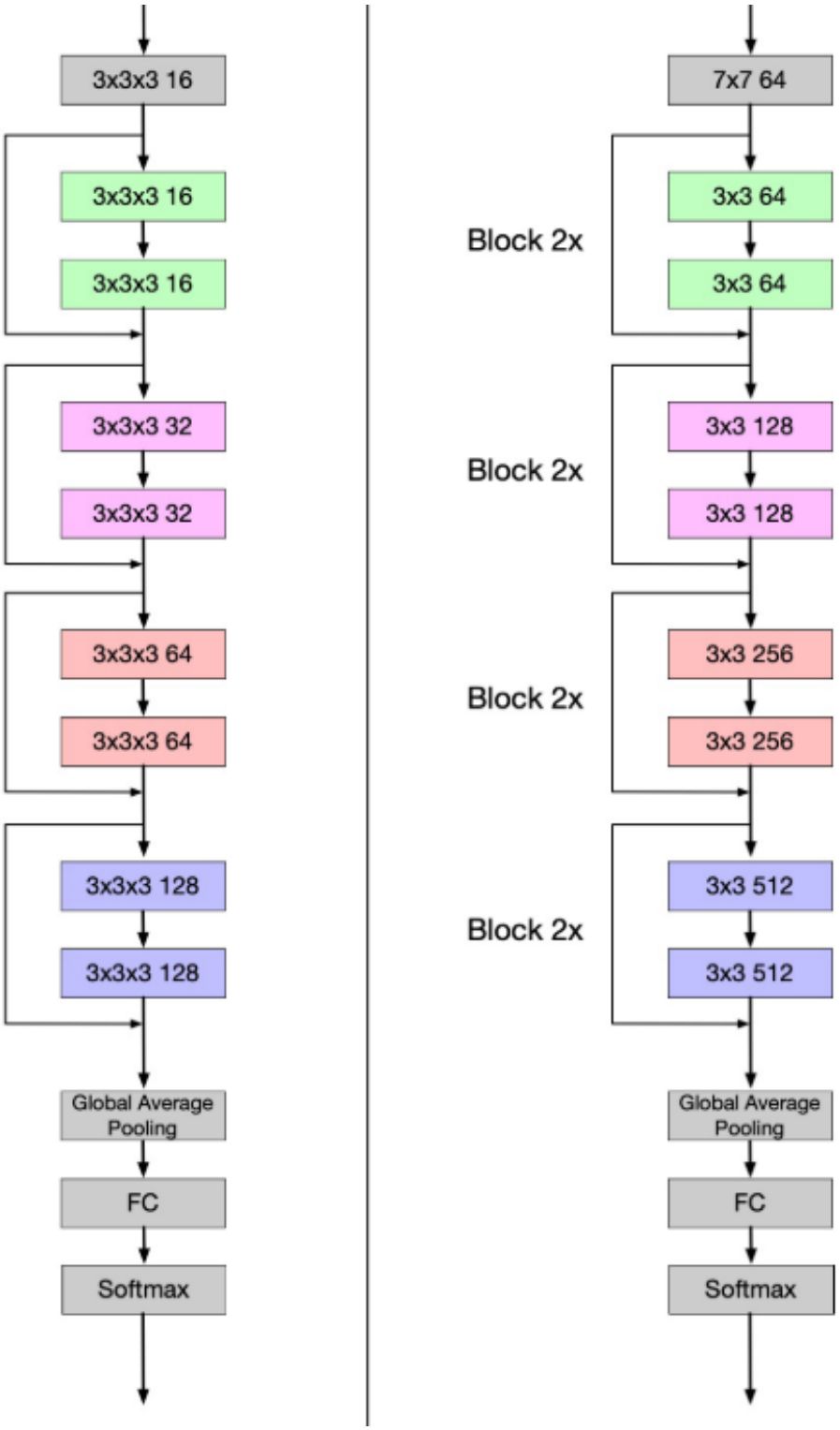


Figure 2.3 ResNet18 networks for 3D tumor and pathology images. FC is fully connected layer. (a) ResNet18 for 3D brain MRI (b) ResNet18 for pathology images and tumor segmentation. Each block shows the size and number of convolutional kernels, all with stride 1 except for the first convolutional block that has stride 2.

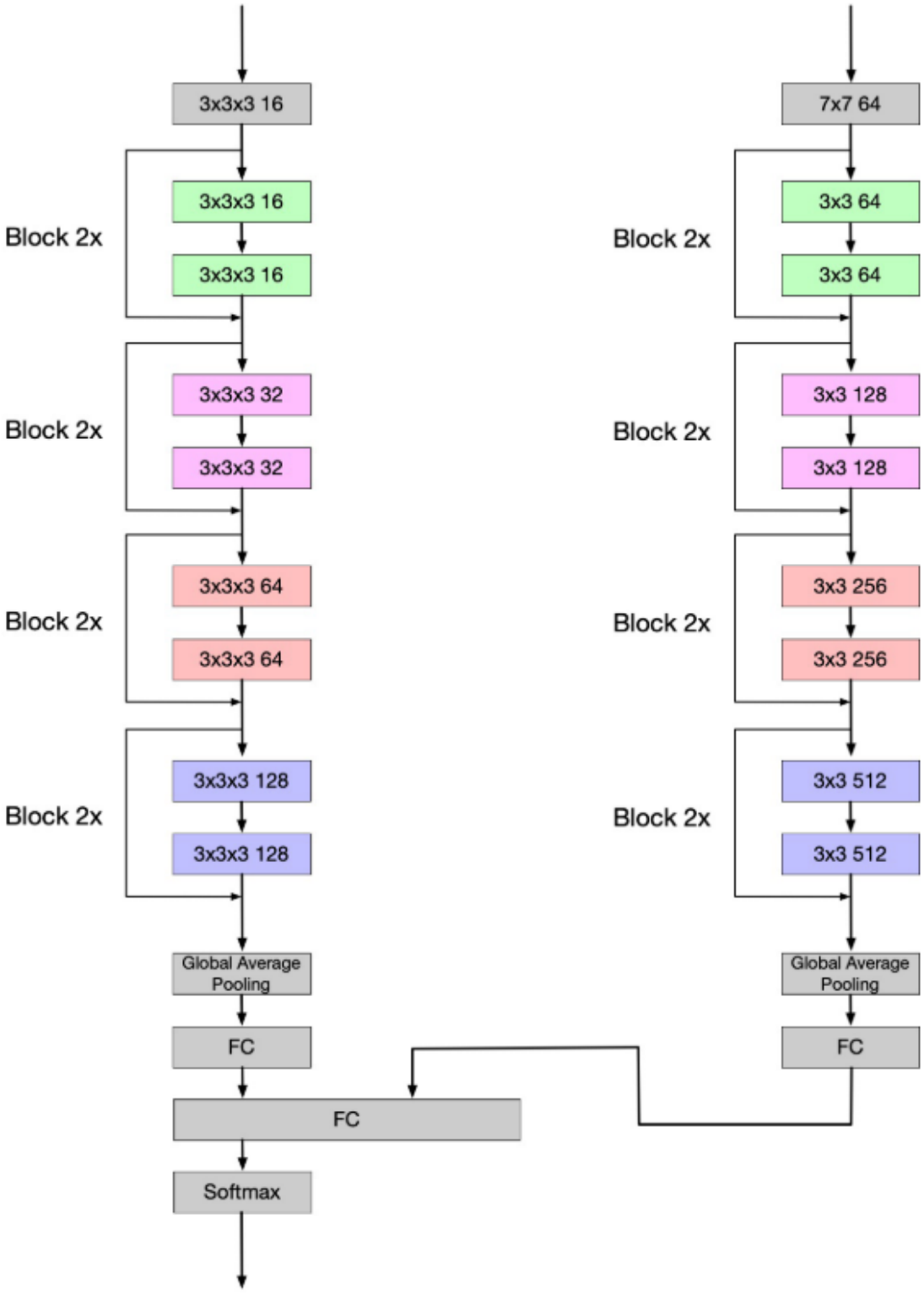


Figure 2.4 Dual path residual convolutional neural network for tumor segmentation and pathology images. Each block shows the size and number of convolutional kernels all with stride 1 except for the first convolutional block that has stride 2.

2.3 Setting Parameter

The networks use the standard cross-entropy loss function to predict the three tumor classes, they are Lower grade astrocytoma, IDH-mutant (Grade II or III); Oligodendroglioma, IDH-mutant, 1p/19q codeleted (Grade II or III); and Glioblastoma and Diffuse astrocytic glioma with molecular features of glioblastoma, IDH-wildtype (Grade IV).

2.3.1 Training Network

The combined model simultaneously takes in tumor segmentation and pathology images from each patient as input for dual path model training. Each tumor segmentation trainer randomly picks eight pathology images of the patient that go into the same batch during training. If a patient has less than eight pathology images (which occur in some cases), the algorithm selects random ones with replacement. At the end of the 2D network is the average operation that averages the features of the eight images into one layer and then concatenated into the 3D part, sees Figure 2.4.

The 3D ResNet18 network is trained with 60 epochs, a learning rate of 0.01, stochastic gradient descent with Nesterov, a batch size of 8, and no weight decay.

The 2D ResNet18 network is trained with 100 epochs, a learning rate of 0.01, stochastic gradient descent with Nesterov, a batch size of 128, and no weight decay.

The dual path network is trained with 50 epochs, a learning rate of 0.01, stochastic gradient descent with Nesterov, a batch size of 8, and no weight decay.

Early stopping is adopted to prevent overfitting. After the training accuracy reaches 90%, the training process will stop when the loss increases in the following one.

2.3.2 Dataset

The dataset is provided by CPM-RadPath 2019 MICCAI challenge, and which includes three classes, Lower-grade astrocytoma, IDH-mutant (Grade II or III); Oligodendroglioma, IDH-mutant, 1p/19q codeleted (Grade II or III); and Glioblastoma and Diffuse astrocytic glioma with molecular features of glioblastoma, IDH-wildtype (Grade IV).

The images are from 221 patients as training data and 35 as validation. Each patient has 3D MRI images in four modalities: native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR). All brain scans were obtained with different clinical protocols and various scanners from different institutions. However, all images were co-registered to the same anatomical template, interpolated to the same resolution (1 mm3), and skull-stripped.

Each patient also owns varying number of pathology images. These are digitized whole slide tissue images captured from Hematoxylin and Eosin (H&E) stained tissue specimens. The tissue specimens were scanned at 20x or 40x magnifications.

To train 3D images on ResNet18, the preprocessing will normalize the data by subtracting the mean and dividing by the standard deviation to give 0 mean and unit variance. The original images are cropped and padded from dimensions 240×240×155 to 160×192×160.

When training on 2D ResNet18, each image is randomly cropped from dimensions 512×512 to 224×224, and a center crop variant is recorded. Data augmentation performs as a random horizontal flip on images during training and inference processes.

When training the dual path model, the MRI images and pathology ones use the same methods described above in the individual networks.

2.4 Experimental Results

Figure 2.5 shows tumor segmentation of a given slice of an MRI image by the BraTS model for each of the three different axial planes. The predicted tumor is highly accurate compared to the true tumor segmentation across all four image modalities. It can be conjectured that the position and size of the tumor play a more significant role in determining the tumor type than the entire MRI image. Therefore, the proposed models will take these as inputs vs. the original MRI images.

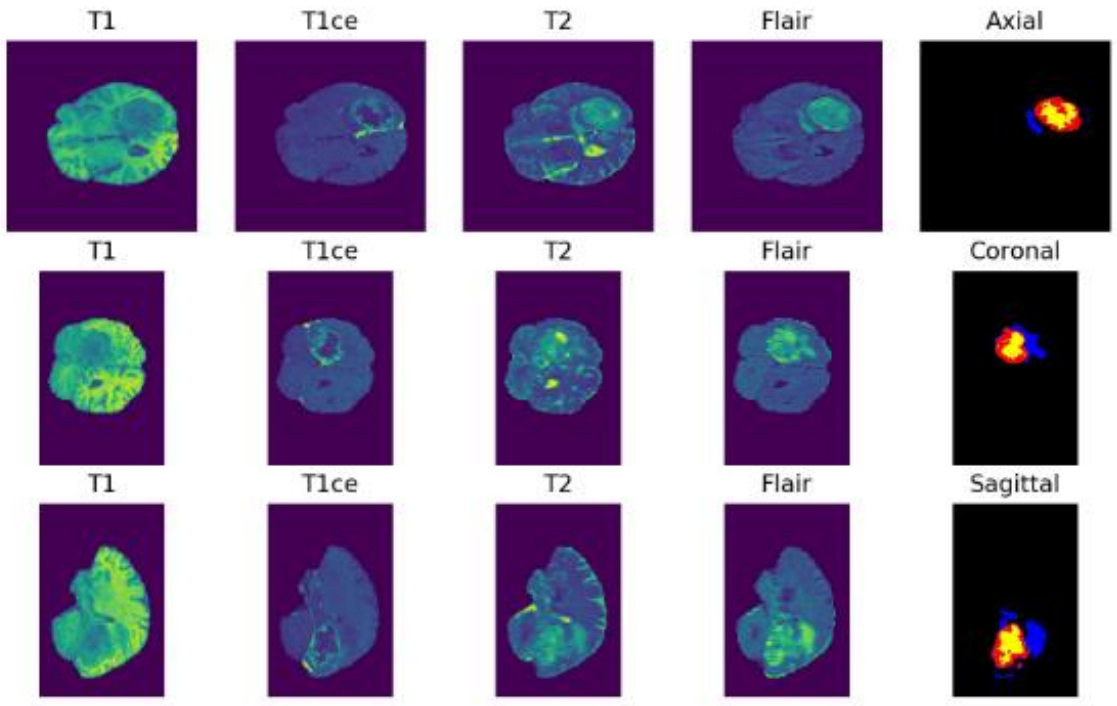


Figure 2.5 Tumor segmentations given by our BraTS model in all three axial planes for a given slice across four image modalities. The proposed models will take the predicted tumor segmentations that are highly accurate in these examples as input to classify the tumor type.

The first evaluation is about the training loss. Figure 2.6 shows the training loss and accuracy of the dual path model on the predicted tumor segmentation, pathology images,

and combined images model. There is a high training accuracy in all three cases that suggests the models may be overfitting. Therefore, early stopping was applied as described above to avoid this situation.

(a) Training loss (b) Training accuracy

Figure 2.6 The training losses and accuracies of the individual and dual path models.

The second evaluation is about the validation accuracies with different training datasets. Table 2.1 shows using the tumor segmentations gives a higher validation accuracy of 77.1% than using MRI images alone, which gives 69.8%. On the other hand, the validation accuracy on pathology images alone is lower than that of MRI and tumor images. In the case of random crops on pathology images, it varies between 66.2% and 69.2%. Combining the MRI images with pathology images under random crops gives us 78.7% validation accuracy, whereas combining with tumor segmentations gives us 81.6%. Finally, combing MRI images with pathology images under center crop also gives 78.7%, while combining tumor segmentation with pathology images under center crop gives the best validation accuracy of 84.9%.

Table 2.1 Validation Accuracy from Different Training Datasets

Custom Dataset	Accuracy
Brain MRI images	69.8%
Predicted tumor segmentation	77.1%
Pathology (center crop)	66.2%
Pathology (random crop)	66.2% - 69.2%
Combined MRI + pathology (random crop)	78.7%
Combined MRI + pathology (center crop)	78.7%
Combined tumor + pathology (random crop)	81.6%
Combined tumor + pathology (center crop)	84.9%

2.5 Conclusion

This chapter presents a dual path residual convolutional neural network that can be trained on both tumor segmentation and pathology images simultaneously. Experimental results show a higher accuracy for predicting tumor category than using the original MRI images alone. Furthermore, the validation accuracies are improved much more. The best result is 84.9%, an increase of 18.7% at most compared with the other seven different cases.

CHAPTER 3

CLASSIFICATION OF HISTOPATHOLOGY IMAGES WITH RANDOM DEPTHWISE CONVOLUTIONAL NEURAL NETWORKS

3.1 Background

The classification of histopathology images plays a crucial role in diagnosing and understanding cancer. Pathologists take a long time training before determining the disease types but cannot avoid the manual errors. Moreover, the diagnosis is time-sensitive. Convolutional neural networks that attain state-of-the-art image recognition have previously been proposed for this problem. The automated classification techniques can promise a more efficient and accurate diagnosis and better treatment.

This chapter presents a depthwise convolutional neural network with random weights (RDCNN) [36]. Previously this has been shown to classify images with a similar background, color, and texture accurately as evaluated on existing benchmarks, Corel Princeton Image Similarity Benchmark [37]. Therefore, it is hypothesized that image similarity may play a role in classification, which may be helpful in the problem of histopathology images. The experimental study on the accuracy of trained convolutional networks compared with RDCNN show that RDCNN used similarity can improve the classification rates and the average accuracy are higher than previouse models.

3.2 Method

3.2.1 Convolutional Neural Networks

Briefly, a convolution layer performs as a moving non-linearized dot product against pixels given by a fixed kernel size $k \times k$ (usually 3×3 or 5×5). The dot product is usually non-

linearized with the sigmoid or hinge (ReLu) function since both are differentiable and fit into the gradient descent framework. The output of applying a $k \times k$ convolution against a $p \times p$ image is an image of size $(p - k + 1) \times (p - k + 1)$.

The convolutional neural network is inspired by multiple layer perceptron (MLP), and it is typically composed of alternating convolution and pooling layers followed by a final flattened layer. The computational engine of the MLP is an arbitrary number of hidden layers that are placed between the input and output layers. MLP is trained with the backpropagation learning algorithm and can solve non-linearly separable problems.

A traditional convolution network is formed by convolutional layers, pooling layers, and fully connected layers and specified by a kernel size and the number of kernels in each layer. Recently, famous modern networks may include residual layers, inception, and more complex structures. In addition, some optimization algorithms also accelerate convolutional neural networks to achieve higher accuracy. These methods all involve randomnesses, for instance, stochastic gradient descent, data augmentation, regularization, dropout, and cutout. Unsurprisingly, the random weights are essential to network architecture in achieving high accuracy.

3.2.2 Random Depth Wise Convolutional Neural Networks (RDCNN)

Consider applying random convolutional blocks repeatedly and then averaging all the values in the final representation of the image. After repeating this step k times, it will generate k new features in k dimensional space. These steps describe a random depthwise convolutional neural network (RDCNN) [36]. During generating the new feature space, no label information is used. Therefore, RDCNN can be considered as an unsupervised feature learning method. Figure 3.1 shows a simple toy example [36]. In (a), Four images, I0, I1,

I2, and I3, contain various objects but are very similar. In (b), four random hyperplanes divide each image into different groups. Each related hyperplane is used to calculate the sign of each patch, and the result is a 2 × 2 matrix for each image shown in (c). The matrix obtains a single feature value for the image given by the hyperplane through the average pooling in (d). Finally, a representation capture these similarities is found during these processes, I0, I1, and I2 are more similar than image I3.

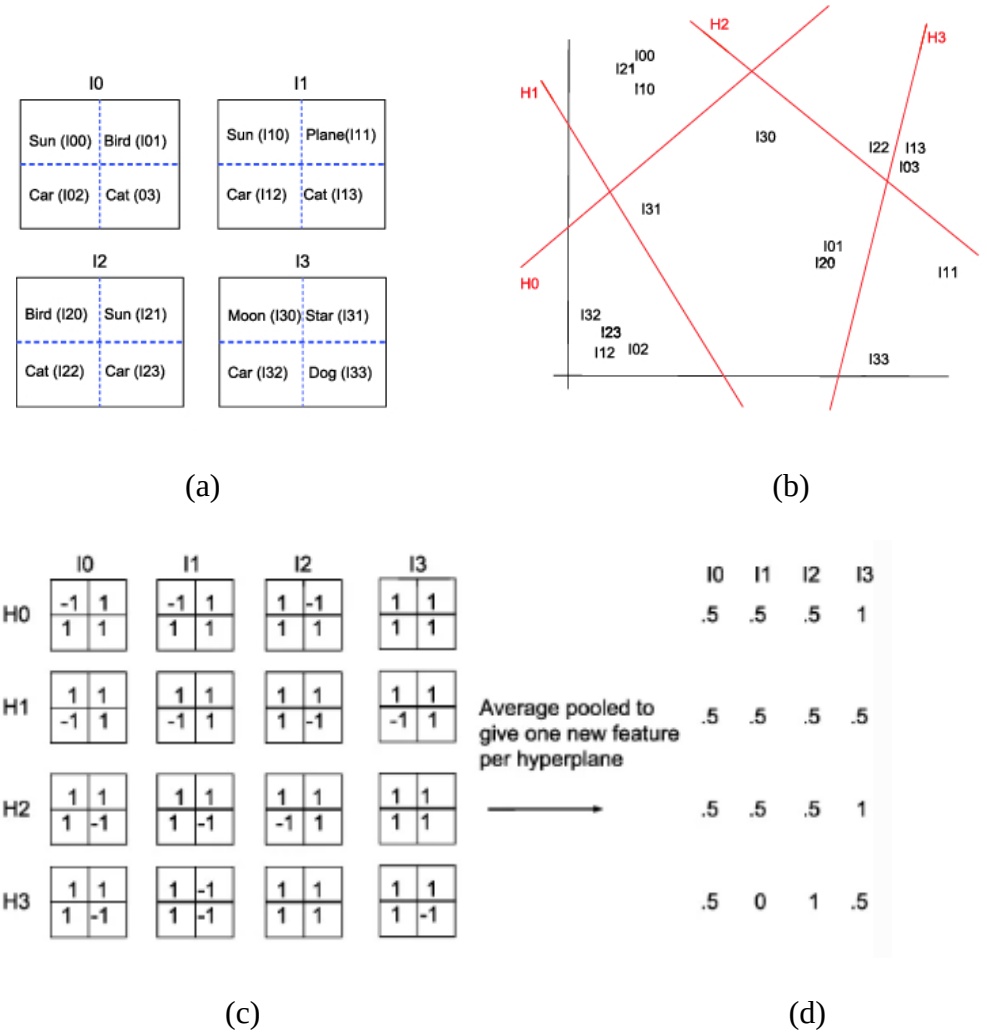


Figure 3.1 The process of randomness. (a) Four images contain objects in different parts of the image and are divided into four partitions. (b) For random hyperplanes (in red) on the input space of features from all patches of the images. (c) The outputs can be considered four partitions. (d) New feature values were obtained from average pooling.

Source: [36].

There is no theoretical guarantee that random hyperplanes would avoid a linearly separable space even repeatedly applying on image patches. However, Figure 3.1 shows that the four random hyperplanes would partition the space, and the patches will have the same outputs if they are in the same space. Other neighbor patches are likely to be less similar. As shown in (d), there is only one different output in these four hyperplanes.

The parameters in the proposed network are the number of convolutional blocks *b*, the size of each kernel $k \times k$, and the number of kernels *m* in each layer (this is the same in each layer). Figure 3.2 shows an example of the RDCNN network with two layers ($l = 2$) and five 3×3 convolution blocks in each layer ($m = 5$, $k = 3$). The values in each convolutional kernel are randomly from the Normal distribution with mean 0 and variance 1.

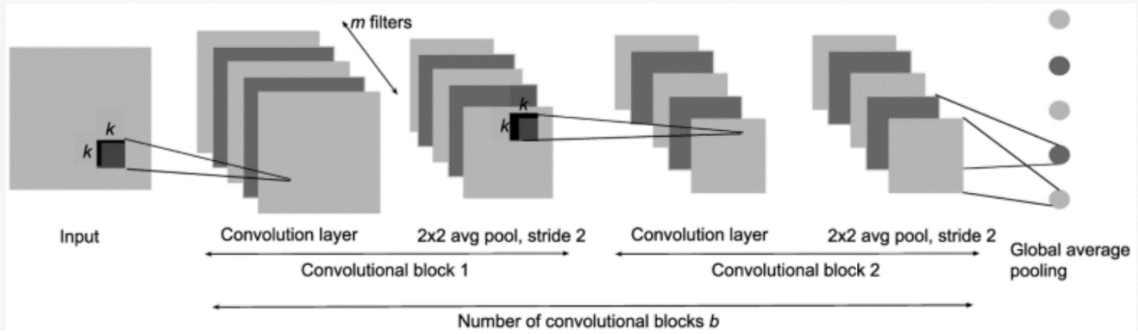


Figure 3.2 A random depthwise convolutional neural network with two convolutional blocks, five kernels with size $k = 3$ in each layer.

The output of each convolution with the sign function is non-linearized, and the convolution is depthwise. The *i-th* convolution is applied only on the *i-th* kernel of the previous layer. In the input layer, however, the convolution is applied in the conventional way to account for three layers of RGB images. After the convolutions, a global average

pooling layer as the final layer generates a flattened feature space. A linear support vector machine trains the final feature space.

3.3 Datasets and Compared Networks

3.3.1 Datasets

There are four publicly available datasets spanning the different cancers. These are available upon request to reproduce experimental results in this paper.

The Invasive Ductal Carcinoma (IDC) dataset is provided by ICPR 2012 contest [4]. The original dataset consisted of 162 whole mount slide images of Breast Cancer histology specimens scanned *40 x 40*. From that, patches of size 50x50 were extracted by the ROI method, of which 198,738 were IDC negative and 78,786 IDC positive. The train and test ratio is 80:20 [38].

This ISIC dataset is provided by the ISIC 2019 Challenge [6]. This is for classifying skin cancer images among nine different diagnostic categories: Actinic Keratosis, Squamous Cell Carcinoma, Basal Cell Carcinoma, Seborrheic Keratosis, Solar Lentigo, Dermatofibroma, Nevi, Melanoma, and Vascular Lesions. This dataset includes a total of 25,331 images, each of size 600 × 400. The train and test ratio is 80:20.

Gleason 2019 dataset contains prostate cancer from H&E-stained histopathology images provided by Gleason 2019 challenge (https://bmiai.ubc.ca/research/miccai-automatic-prostate-gleason-grading-challenge-2019). This challenge is part of the MICCAI 2019 Conference and will be one of the three challenges under the MICCAI 2019 Grand Challenge for Pathology. Data used in this challenge consists of 267 tissue micro-array (TMA) images. The size of each image is 5120 x 5120. Each TMA image is

annotated in detail by several expert pathologists. Map1 (the first expert pathologist labels) is selected as the proper labels and split these images into train and test with a ratio of 80:20.

The microscopic biopsy images in the BreakHis dataset were collected from 82 patients using different magnifying factors (40X, 100X, 200X, and 400X). The images are provided in their raw PNG (Portable Network Graphic) format, without normalization or color standardization, and are all the same size (700x460 pixels, 3-channel RGB, 8-bit depth per channel). The samples were collected using the Surgical Open Biopsy method, called partial mastectomy or excisional biopsy [39]. This procedure removes a large tissue sample and is done in a hospital with general anesthesia.

The benign and malignant image groups are further divided into sub-groups describing the specific kind of anomaly. For benign lesions, the anomalies present are fibroadenoma, Phyllodes tumor, and tubular adenoma. For the malignant lesions, the anomalies present are ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma.

The images at the 400X magnification level are used in the experiment. In total, there are 1,606 samples in dataset, the ratio of train and test is 80:20. Out of that total, 374 samples are benign, and 1,232 are malignant. Using augmentation (adding more samples by rotating and flipping the original images), there will be 8,120 samples (6,496 for training and 1,624 for testing).

3.3.2 Deep Networks Compared in Study

Two modern networks are used to evaluate the performance of the RDCNN network, VGG [4] and ResNet [5]. These two convolutional neural networks are designed to enable deeper

architectures and are trained with stochastic gradient descent. Both networks that we use have previously shown high accuracy on the ImageNet classification benchmark and thus serve as competitive baselines in the study. The models are implemented in Keras.

VGG16 is based on the deep convolutional neural network with several convolutions and pooling layers. VGG16 is one of the winners of the ImageNet contest in 2014 [4]. On the other hand, ResNet50 is the deep residual convolutional network [5] that contains connections from previous layers and not just the last one. ResNet won the ImageNet contest in 2015.

3.4 Experimental Results

As observed before [36, 40], Figure 3.3 shows that increasing features increases the test accuracy. Figure 3.1 shows that a new feature space will be generated in the final flattened layer when applied a kernel to each image. There are improvements in test accuracy as increasing the number of kernels on the STL10 and CIFAR10 benchmarks in Figure 3.3. It also shows that the training accuracy reaches 100% much faster. However, the test accuracy continues to improve.

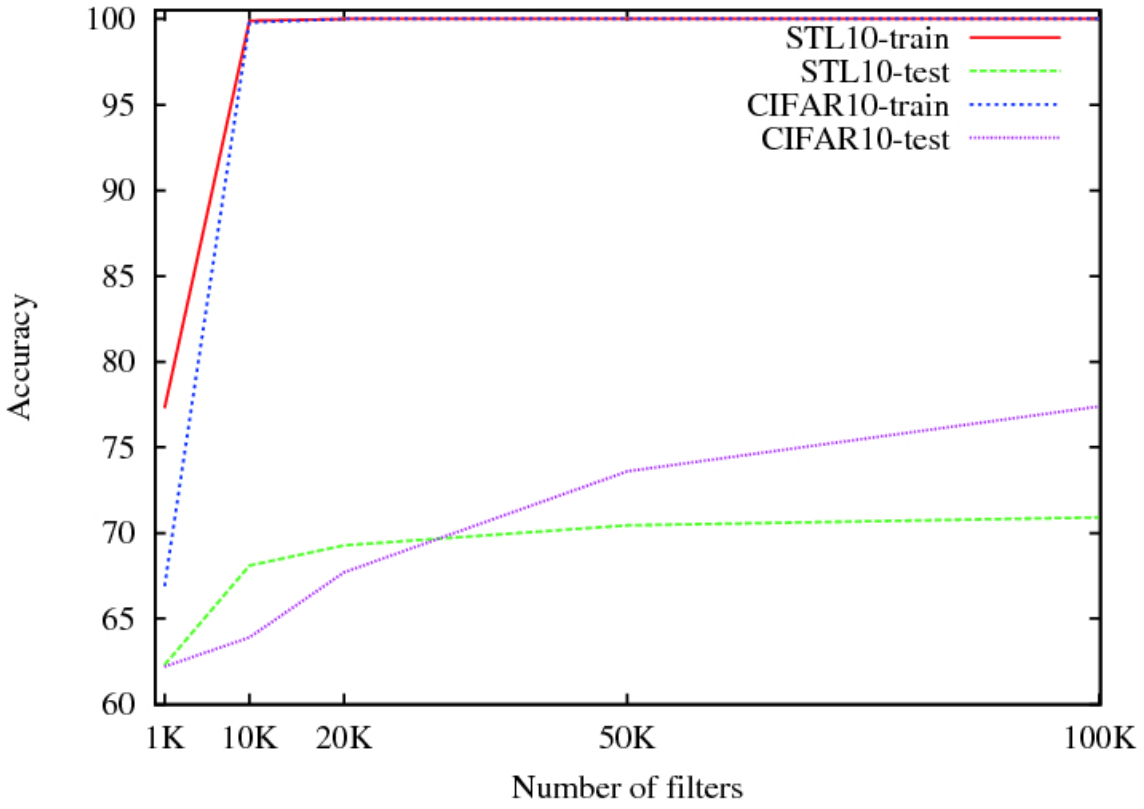


Figure 3.3 Effects of the increasing number of kernels (final features) on the test accuracy.

Source: [36]

Therefore, preliminary experiments are conducted on RDCNN to study the effect of increasing features. Both ResNet50 and VGG16 are trained with a batch size of 32 and center cropped images for the Gleason dataset (whose images are enormous in dimensions). ISIC images are also performed center cropping to improve its test accuracy. Images in IDC are cropped by the ROI method into small patches.

Table 3.1 shows the train and test accuracies of VGG16, ResNet50, and RDCNN on the four datasets as described before. On ISIC and Gleason, RDCNN achieves much higher accuracy than the remaining datasets. The kernel size of RDCNN has little effect on the datasets shown here. On the BreakHis (2-class and 7-class) and IDC datasets, ResNet50

has the highest accuracy, but RDCNN is only 1%, 3%, and 0.6% behind on test accuracy. After averaging the accuracy across all the datasets, RDCNN has the highest mean of 95%, whereas VGG16 and ResNet50 have 79% and 89.8%, respectively.

Table 3.1 Train and Test Accuracies (Shown as Percentages) of Fully Trained VGG16 and ResNet50 and Unsupervised RDCNN on Our Datasets

Dataset	Method	Train	Test
IDC	VGG16	92.2	83.3
	ResNet50	100	88.2
	RDCNN(30K features, k=3, 4 layers)	87.8	87.6
	RDCNN(50K features, k=5, 4 layers)	86.3	87.6
Dataset	Method	Train	Test
ISIC	VGG16	89.8	85.9
	ResNet50	90.3	87.5
	RDCNN (65K features, k=3, 4 layers)	100	100
	RDCNN (65K features, k=5, 4 layers)	100	100
Dataset	Method	Train	Test
Gleason	VGG16	83.3	73.4
	ResNet50	87.5	75
	RDCNN (68K features, k=3, 4 layers)	100	93.5
	RDCNN (70K features, k=5, 2 layers)	100	93.5
Dataset	Method	Train	Test
BreakHis (2 classes)	VGG16	82.8	81.8
	ResNet50	100	99.8
	RDCNN (10K features, k=3, 7 layers)	100	98.8
Dataset	Method	Train	Test
BreakHis (7 classes)	VGG16	99.14	70.61
	ResNet50	94.1	98.6
	RDCNN (10K features, k=3, 7 layers)	99.5	95.4

To determine why RDCNN performs better in this study, Figure 3.4 shows a random image from the ISIC dataset and its top ten similar images in each network's final

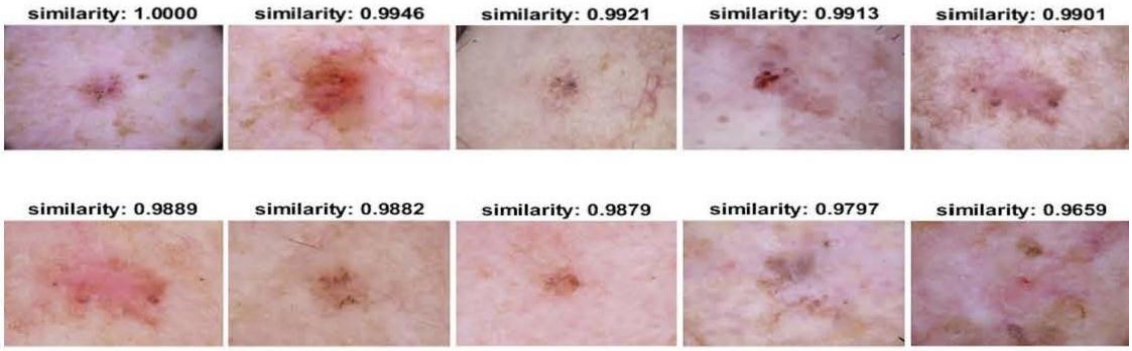
layer feature space. In the spaces of RDCNN, all ten similar images are in the same category as the query. However, in ResNet50 and VGG16, two and three are from different classes. In datasets such as ISIC, where similarity also implies the same category, the unsupervised RDCNN network outperforms trained models.

Figure 3.5 shows the top ten similar images from a randomly selected Gleason dataset. In the RDCNN final layer space, seven of the ten images are in the same category as the query. In contrast, in the ResNet50 and VGG16 final layer space, only six and five are in the same category as the query.

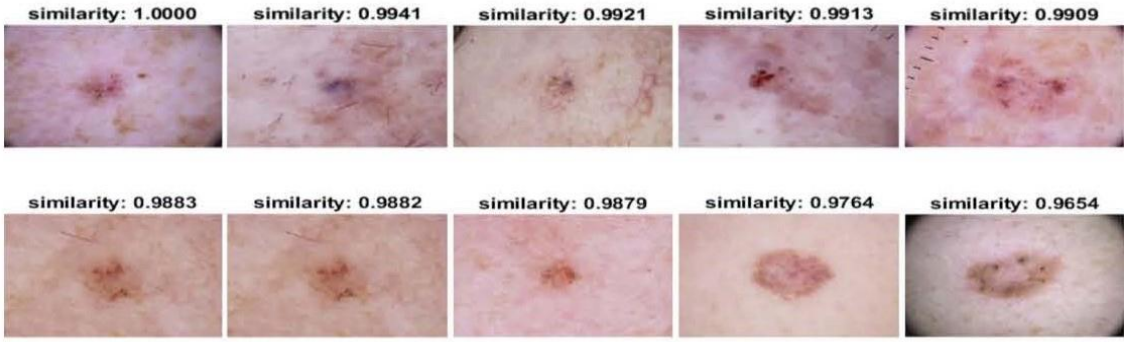
3.5 Conclusion

In this chapter, the unsupervised RDCNN is proposed. Compared with two state-of-the-art networks, the preliminary results suggest that RDCNN can be highly useful in classifying histopathology images where similarity also implies the same class membership. Furthermore, a kernel size of 3 and 4 layers works well in most cases on medical images.

(a) Top similar images to the query in RDCNN final layer feature space

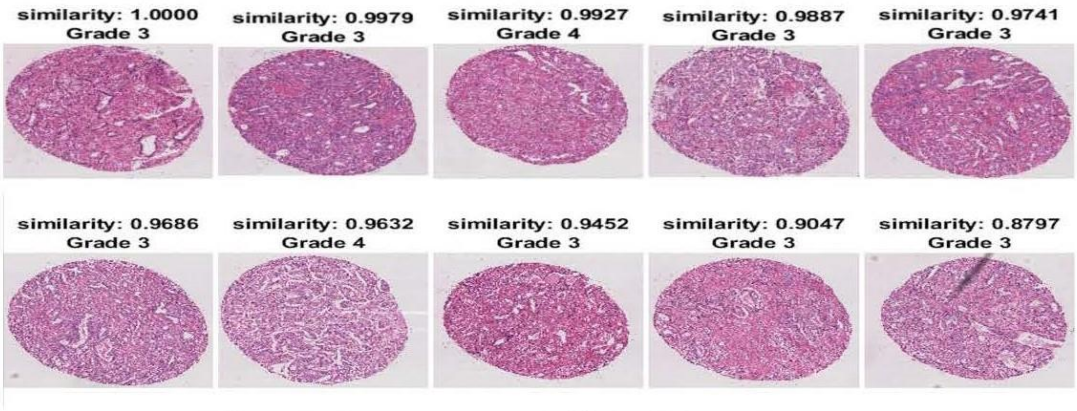


(b) Top similar images to the query in the ResNet50 final layer feature space

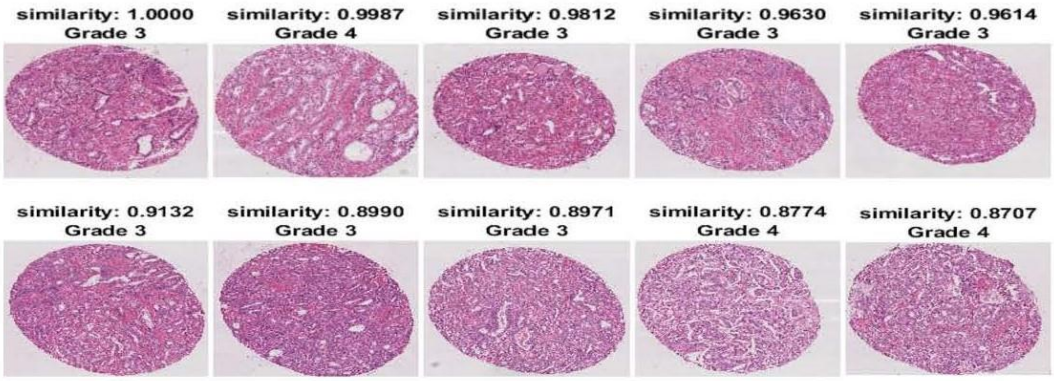


(c) Top similar images to the query in the VGG16 final layer feature space

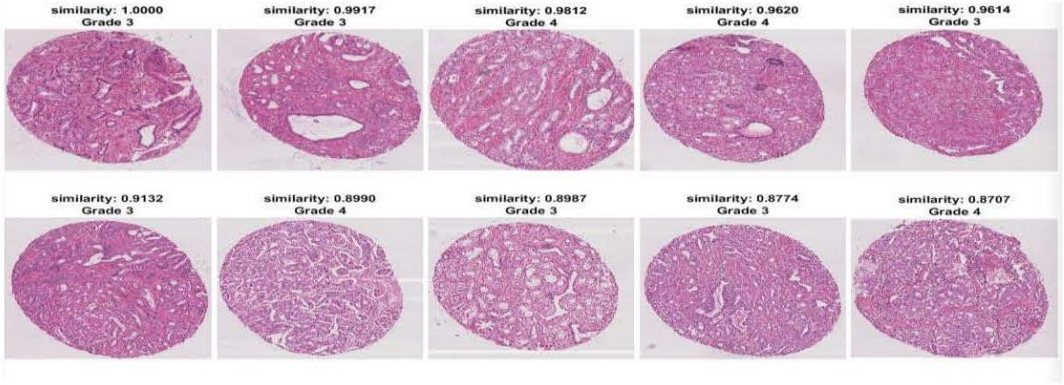
Figure 3.4 Top 10 similar images in the ISIC dataset. Randomly selected ones (also from ISIC) are in the same class as the query in the RDCNN feature space but not in ResNet50 and VGG16 space.



(a) Top similar images to the query in RDCNN final layer feature space



(b) Top similar images to the query in the ResNet50 final layer feature space



(c) Top similar images to the query in the VGG16 final layer feature space

Figure 3.5 Top 10 similar images in the Gleason dataset. Randomly selected ones (also from Gleason) are also in the same class as the query in the RDCNN feature space but less in the ResNet50 and VGG16 space.

Chapter 4

ACCURATE AND ADVERSARIALLY ROBUST CLASSIFICATION OF MEDICAL IMAGES WITH GRADIENT-FREE TRAINED SIGN ACTIVATION NEURAL NETWORKS

In the previous chapters, machine learning algorithms have been proven to achieve high accuracy in classification tasks, and more new modules have been proposed to enhance accuracy. However, they could misclassify some data added by minor perturbations known as adversarial attacks [11-23]. The attackers can fool machine learning systems with adversarial images, 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 cause insurance fraud.

4.1 Adversarial Attack

Adversarial examples have been shown to transfer across models, making it possible to perform transfer-based (substitute model) black-box attacks [13]. Transfer adversarial attacks and boundary attacks are the most lethal as they can be performed effectively without access to the model's parameters.

Researchers have investigated adversarial attacks on medical images and mainly focused on testing the robustness of deep learning models for medical image analysis [15 - 17]. Paschali *et al* [17] showed that classification accuracy drops from above 87% on the regular medical images to almost 0% on the adversarial examples. Hokuto *et al* [18] demonstrated that the attack method achieved over 80% success rates on the DNNs model. Many defense methods have been proposed to defend against adversarial attacks, in which

adversarial training is most prevalent. However, this tends to lower accuracy on clean test data. To overcome this problem, transfer-based methods were developed [13, 14], but they are still vulnerable. Therefore, adversarial robustness is still an open problem in machine learning.

Generally, the adversarial attack injects minor distortion into original data to fool a machine learning system. For example, assuming a network *F* can correctly classify the clean data *x,* Equation (4.1) represents the projection from data *x* to corresponding classification result *y*.

$$F(x) = y \tag{4.1}$$

Let the fake data be *x'*. The adversarial attacker *G* generates an adversarial example *x',* so the adversarial example is in Equations (4.2) and (4.3).

$$x' = G(x) \tag{4.2}$$

$$d(x, x') \leq \varepsilon(x) \tag{4.3}$$

$$F(x') \neq y \tag{4.4}$$

The G aims to generate a successful *x'* where the Euclidean distance *d* between *x* and *x'* should be smaller than a threshold value *ε(x),* so F cannot detect the *x.'* Finally, the classifier will misclassify the data ‘x as in Equation (4.4).

The images shown in Figure 4.1can help understand the harm of adversarial attack. By adding an imperceptibly small noise, the classification result of the image is changed. Here the value *.007* corresponds to the magnitude of the tiniest bit of an 8-bit image

encoding after GoogLeNet [3] conversion to real numbers. However, the adversarial example on the third column looks the same as the first clean image for humans.

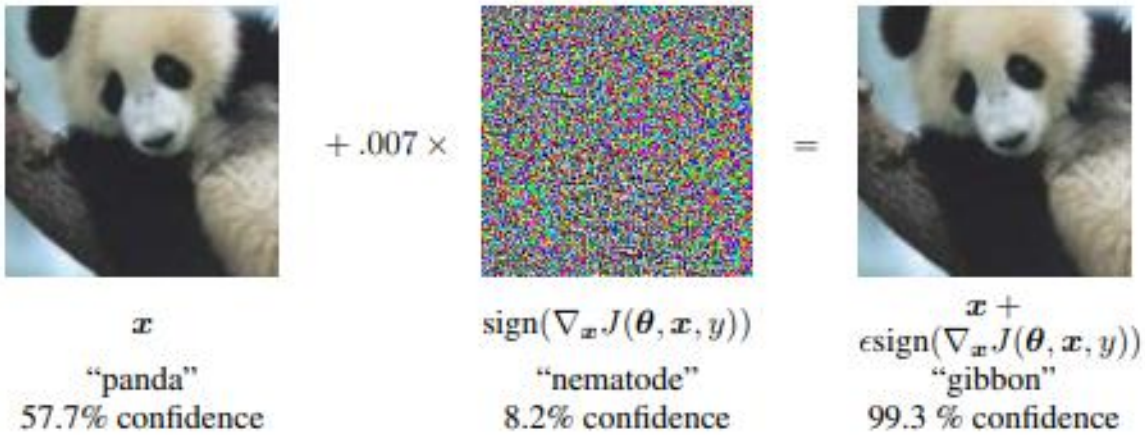


Figure 4.1 A example of adversarial example generation applied to GoogLeNet [3] on ImageNet.

Source: [19]

4.1.1 Black-box Attack

The black-box attack method is also called Decision-based attack. The attackers cannot access neural networks and the accurate probability in the real world but know the predicted category. Brendel *et al* [41] proposed that a decision-based adversarial attack achieves similar performance as a white-box attack. However, it is hard to practice on networks in the real world, since the massive queries to model are time consuming. Jianbo *et al* [42] proposed a query-efficient decision-based attack named HopSkipJump. It asked fewer queries to the target model and competitive performance in attack compared with Brendel.

In the following experiment, the HopSkipJump implementation in the IBM Adversarial Robustness Toolkit [42] is used to evaluate the robustness of the model. It is a family of algorithms and includes both untargeted and targeted attacks optimized for L2 and L∞ similarity metrics. 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 that 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.

4.1.2 White-box Attack

Unlike black-box attacks, white-box attackers can access all the information of the target models, such as architecture, parameters, and gradients. Therefore, the white-box attack can carefully craft adversarial examples by using this crucial information. Nonetheless, it is hard to use in practice since disclosing model architecture and parameters used in industries is rarely public; only some academic research will be open.

Recent researches on White-box attacks aim to help people understand the weakness of DNN models. Commonly used white-box attack algorithm includes Fast Gradient Sign Method [19], Deep Fool [18], and Projected Gradient Descent Attack [43]. Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attack will be used in this dissertation to study the defense ability of models.

4.2 Methods

With many kinds of research introduced above illustrating the incredible power of adversary examples, the adversarial robustness of machine learning models has achieved colossal progress [17, 20-29]. In the beginning, the gradient-based white-box attacks are used to improve the security of models. Then there are more strategies developed to defend against the adversarial attacks based on searching minimum distortion and black-box attacks.

Gradient-free trained sign activation networks have been proven a higher possibility to defend against adversarial attacks [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. This chapter adopts the gradient-free stochastic coordinate descent algorithm for training sign activation networks on medical image datasets to defend against black-box attacks.

4.2.1 The Stochastic Coordinate Descent (SCD)

Assume 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 Equation (4.5),

$$L_{scd} = \sum L\,(w, w_0, x_i, y_i) \tag{4.5}$$

where starting 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, a random set of features (coordinates) from w is selected called F. For each feature $w_i \in F$, we add or subtract a learning rate η and then determine w_0 that optimizes the risk. Finally, all possible values of w_0 are computed as in Equation (4.6),

$$w_0 = \frac{w_i^T x_i + w_{i+1}^T x_{i+1}}{2},\ (i = 0, 1, ..., n - 2) \tag{4.6}$$

where 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.

The above search algorithm is stochastic coordinate descent and is abbreviated by SCD. SCD will be applied to the final node and then a randomly selected hidden node in

each algorithm iteration to train a single hidden layer network. In practice, parallelism and several heuristics are used to speed up the run time.

4.2.2 Loss Function

The loss function is derived from the mathematical optimization theory, and it usually maps one or more values to the actual numbers. Intuitively it means the event's cost. The loss function related to machine learning is used to evaluate the difference between the predicted results of the model and the actual value. The lower value of loss function indicates the more accurate model performance obtained from training. Two general loss functions will be used in the proposed networks in classification tasks: Zero-one loss, and Cross entropy loss.

1. Zero-One Loss

Zero-one loss describe as in Equation (4.7), where *F* is a model which can correctly classify the input data *x*, and the predicted output is *F(x)*, *y* is the correct label related to *x*. When the predicted value is the same as the actual label, the loss value will be 1; otherwise, it is 0. Zero-one loss function is non-convex, and it is tough to solve. Therefore, it is more convincible when judging the number of errors in classification prediction.

$$L(y, F(x)) = \begin{cases} 1, if\ y = F(x) \\ 0, if\ y \neq F(x) \end{cases} \tag{4.7}$$

2. Cross-Entropy Loss

Cross-entropy loss is also named log loss; the output *L* is a probability value between 0 and 1. Cross-entropy loss values follow the same tendency as the predicted probability *a* of the

actual label *y*. For instance, when the actual observation label is 1, but the probability of predicting 1 is .012, that would be bad and result in a high loss value. Ideally, the loss of the perfect model would be 0.

$$\mathrm{L}\big(\mathrm{y}, \mathrm{f}(\mathrm{x})\big) = -\frac{1}{n}\sum_x [y * \ln a + (1-y) * \ln(1-a)] \tag{4.8}$$

The sigmoid function is usually used to obtain the probability of a binary class dataset.

4.2.3 Network Implementation

The following three types of sign activation networks using the proposed algorithm are trained in this dissertation:

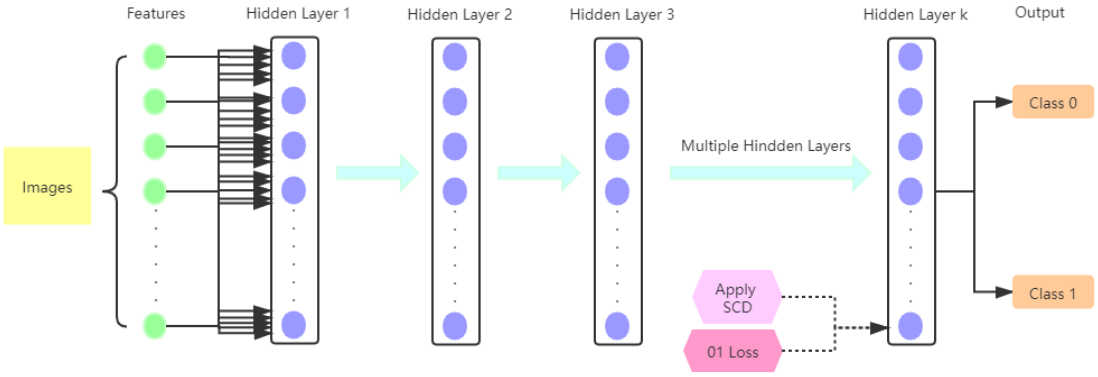
1. SCD01MLP: 01-loss in the final node with an MLP network.

2. SCDCEMLP: Cross-entropy loss in the final node with an MLP network.

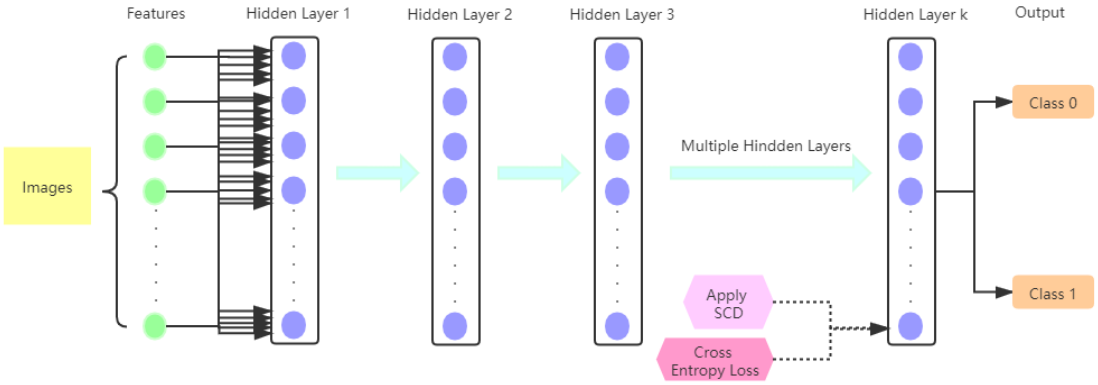
3. SCDCEBNN: Cross-entropy in the final node with binary weights throughout the model.

The basic architecture of SCD models is shown in Figure 4.2. The training procedure is implemented in Python, Numpy, and Pytorch. Since sign activation is non-convex, the training process starts from a different random initialization, runs 100 times, and outputs the majority vote.

To illustrate the run time and clean test accuracies, the experiments are designed to compare the proposed models with the convolutional networks LeNet [1], ResNet18 [5], 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 to approximate the sign activation.



(a)



(b)

Figure 4.2 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.

In addition, the HopSkipJump is used to evaluate the robustness of those target models. Theoretical analysis and experiments show that 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. Figure 4.3 shows the scheme of the evaluation process.

HopSkipJump attacks a predictive model to generate an adversarial image, which would fool the model. Each image will be attacked by HopSkipJump 10 times to increase the chance of obtaining an accurate estimation. Each time, the initial pool size is 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. Therefore, this dissertation can report the distortion of only five random points in this dissertation.

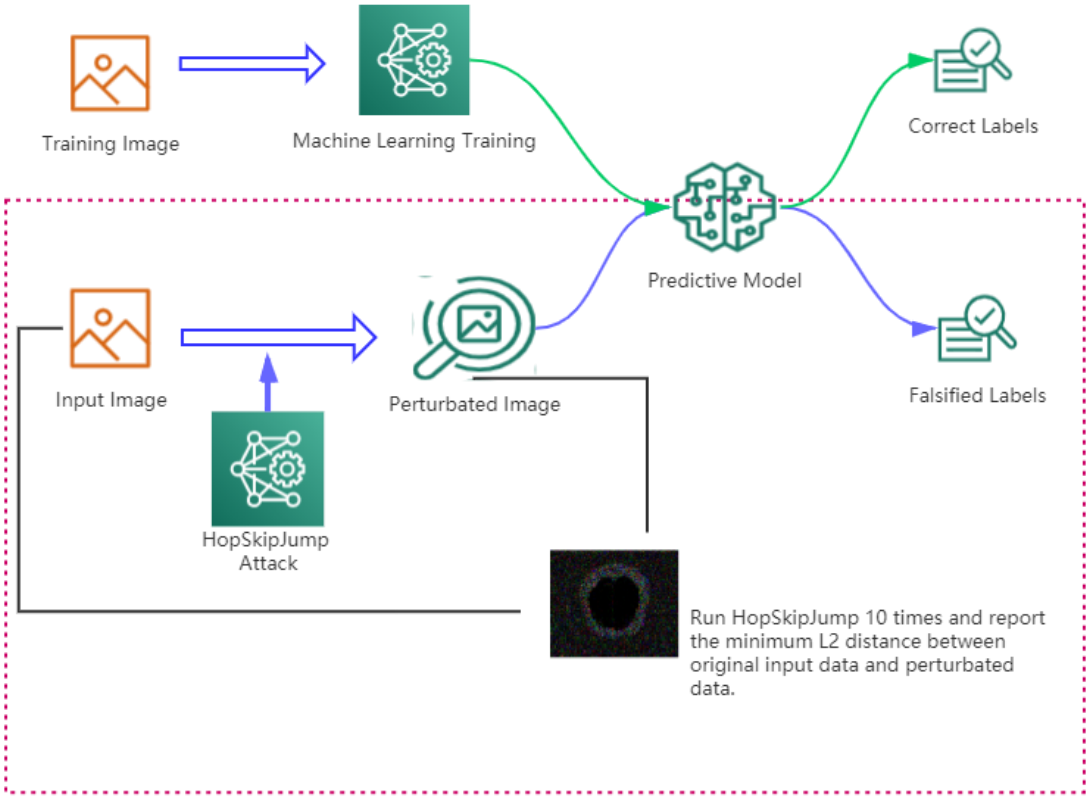


Figure 4.3 The procedure of attacking the models with HopSkipJump.

4.3 Dataset

Three popular medical imaging datasets are used to evaluate the classification accuracy, BraTs18, Chest X-rays, and Colorectal Histopathology. Aside from the difference in imaging tissue and modality of these three data sets, the images are shown in Figure 4.4.

4.3.1 BraTs18

The BratS18 dataset is 210 high-grade glioma (HGG) and 75 low-grade gliomas (LGG) MRI with binary masks for the tumor. Each 3D MRI contains 155 slices of size 240 × 240. The FLAIR modality images are used in all the experiments because the entire tumor is represented well by this modality. In total, there are 17,100 abnormal and 18,500 benign images for training. For testing, there are 1,800 abnormal and 1,900 benign images. Here also show more experimental results on other modalities, where ANT-GAN presents impressive synthesis quality. The two classes are down-sample to be a balanced dataset, and each class contains 1,462 images, which are resized to 96 × 96. The ratio of training and testing datasets is 80: 20.

4.3.2 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 two 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.

The evaluation set is checked by a third expert to account for any grading errors. The preprocessing resizes the images to 96 × 96 and down-sample to 1,584 for each category as the balanced dataset. There are 3,168 images in total, which are split into training and testing sets by a ratio of 80: 20.

4.3.3 Colorectal Histopathology

This dataset represents a collection of textures in histological images of human colorectal cancer. Ten anonymized H&E stained CRC tissue slides are obtained from the pathology archive at the University Medical Center Mannheim, Heidelberg University, 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 tissue types 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. The experiments randomly pick two classes, immune cells and normal mucosal glands, resize them to 96 × 96, the ratio of train and test sets is 80: 20.

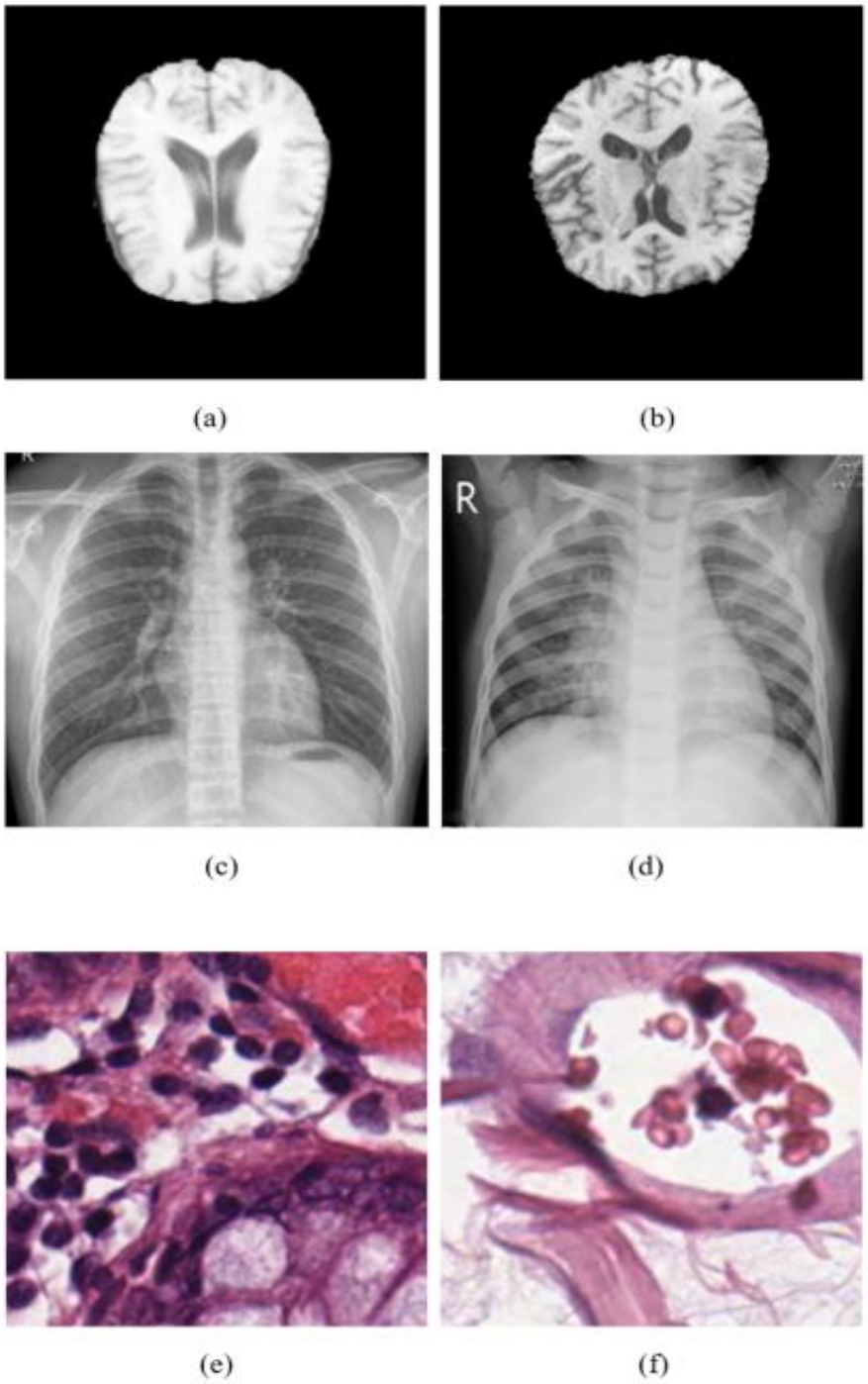


Figure 4.4 The sample images from three datasets. (a) Benign brain MRI, (b) abnormal brain MRI, (c) health chest X-ray, (d) pneumonia chest X-ray, (e) and (f) two different classes of human colorectal cancer, normal mucosal glands, and immune cells.

4.4 Qualitative Analysis

4.4.1 Evaluation of the Test Accuracy

Firstly, this work conducts experiments to compare all seven models' clean test accuracies on chest X-ray, histopathology, and BraTs18. The results are listed in Table 4.1. On the Chest X-ray dataset, the convolutional networks LeNet [1] and ResNet18 [5] have higher accuracies since they have the advantage of convolutions. On histopathology, the MLP and Random Forest have higher accuracies. Finally, on BraTs18, the ResNet18, LeNet, and Random Forest have higher accuracies, but other models are not too far behind.

Table 4.1 Average Accuracy of Validation Data on BraTs18, Chest X-ray, and Histopathology Image Datasets

	SCD01	SCDCE	SCDCEBNN	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%

4.4.2 Evaluation of the Defense Ability by L2 Distance

This work quantitatively measures the defense's robustness by measuring the distance between normal and abnormal samples under the *L2* metric, as most attacks did [44]. The *Lp* distance is the difference between original examples and adversarial examples, defined as

$$||d||_p=(\sum_{i=0}^{n}|vi|^p)^{1/p} \tag{4.9}$$

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 minor, the images will likely appear visually similar.

This section compares 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 and tree ensemble classifiers. Even the approximation of the minimum distortion in ReLu activated neural networks is NP-hard.

The distortions reported by HopSkipJump are lower (i.e., tighter and more accurate) than other boundary attack methods. Therefore, the HopSkipJump boundary-based black-box attack determined the adversarial distortion of randomly selected images from the BraTs18, chest X-ray, and colorectal cancer histopathology validation datasets.

The HopSkipJump is run ten times on each image to report the minimum value.

As shown in Figure 4.5, after 90 iterations, the distortions are minimum in all cases and become stable. Therefore, considering the best results and the computational ability, the experiments set 100 as the maximum iteration.

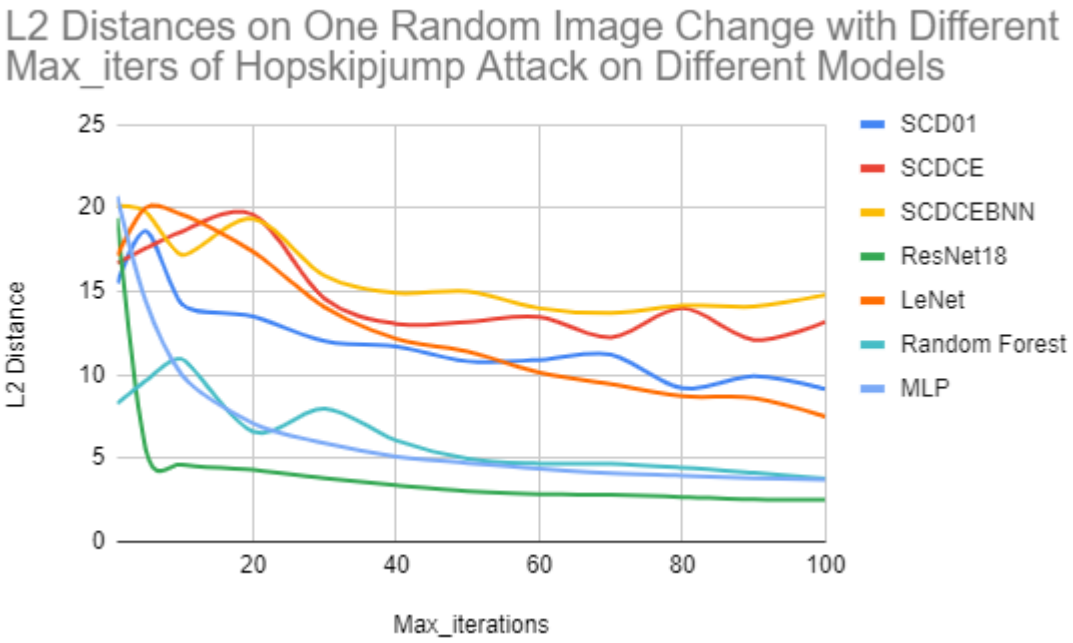


Figure 4.5 L2 distances on one image change with different max iterations when Hopskipjump attacks different models.

Table 4.2 and Figure 4.6 shows the average adversarial distortions of random test images from the BraTs18. Again, the gradient-free trained sign networks have higher distortions than other state-of-the-art models, and the SCDCEBNN has the highest distortion.

Table 4.2 Average Minimum Estimated L2 Adversarial Distortion 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	13.70	4.38	8.27	2.00	2.78

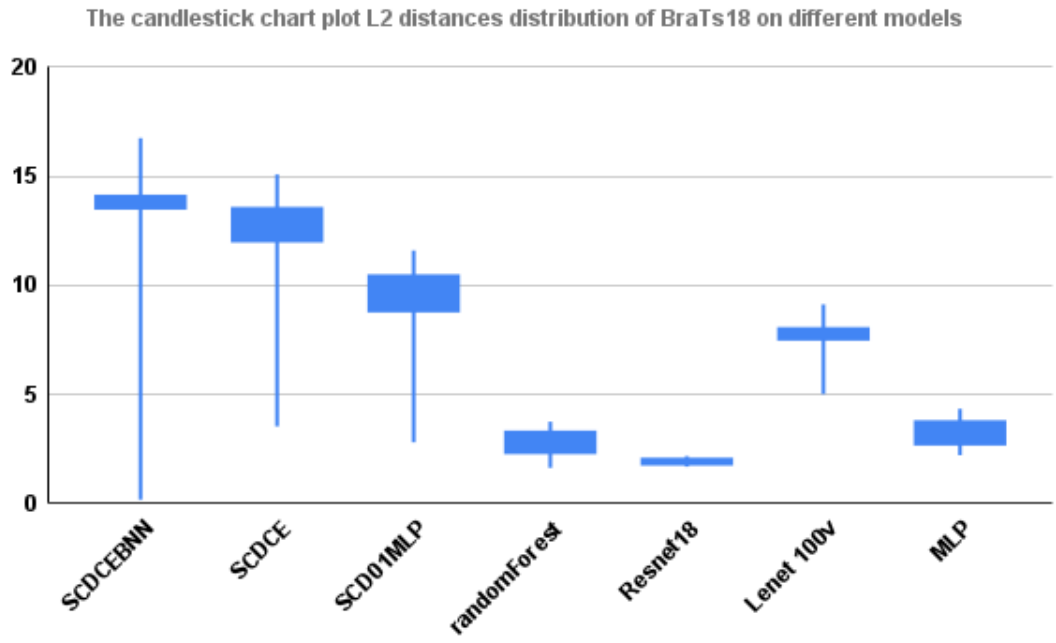


Figure 4.6 The candlestick chart plots the L2 distances on all adversarial images from BraTs18 on different models.

Figure 4.7 plot the original and adversarial images of “Image 1” from the BtaTs18 dataset 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.

Table 4.3 and Figure 4.8 lists the average adversarial distortions of random test images from the Chest X-ray dataset, where MLP is the second best after SCDCE.

Figure 4.9 shows the original and adversarial images of “Image 1” from the Chest X-ray dataset to get a visual feel for the distortions. They all have higher distortions, among which SCDCE has the highest.

Table 4.4 and Figure 4.10 lists the average adversarial distortions of random test images from the colorectal dataset. Again, the average distortion of SCDCEBNN is the highest.

Figure 4.11 shows the original and adversarial images of the 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 dots. The morphology is hard to identify, such that it would be potentially abnormal.

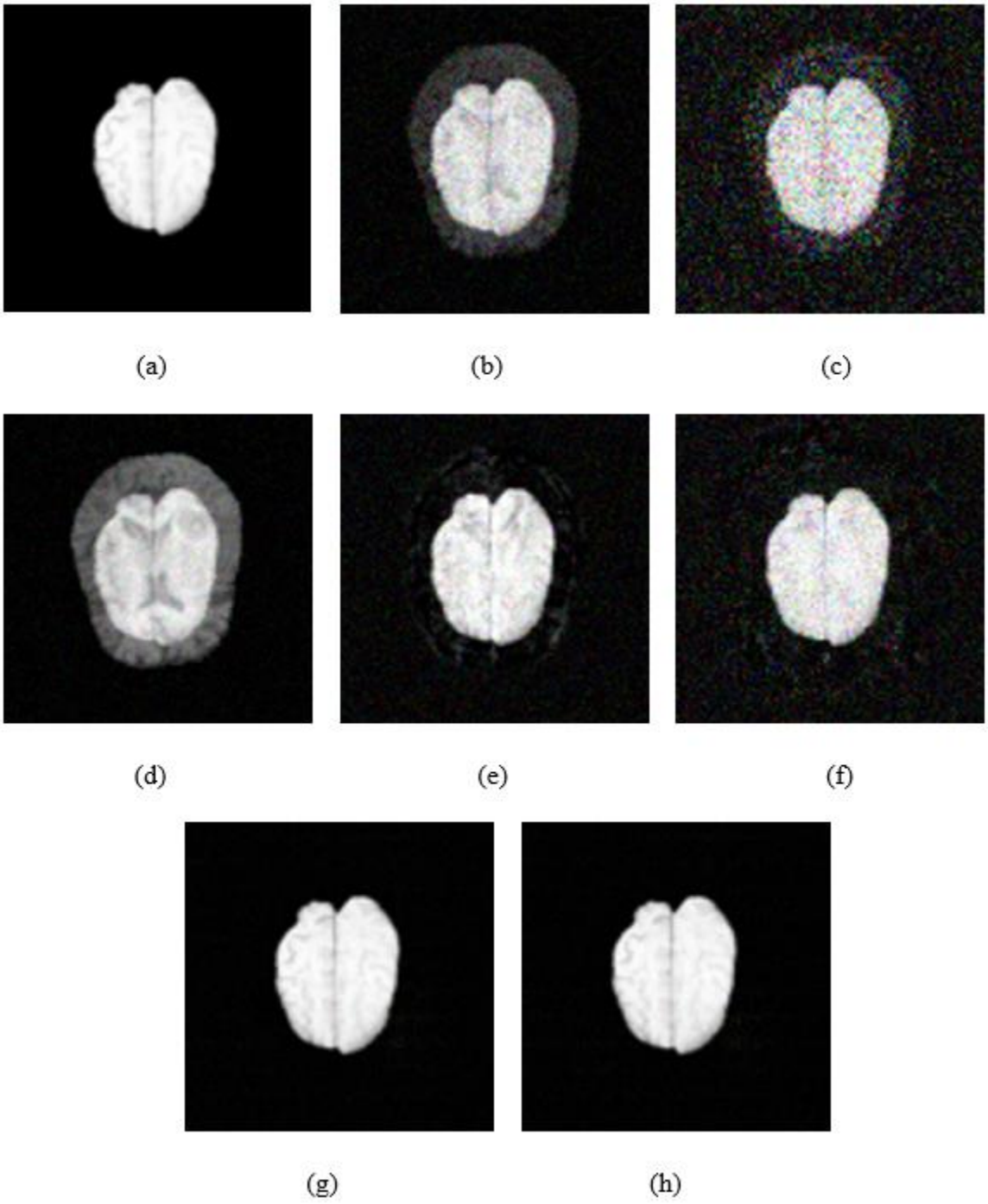


Figure 4.7 Visualization of original and adversarial images among different networks from BraTs18 dataset. (a) The original image, (b) the adversarial example which will fool SCD01, (c) the adversarial example which will fool SCDCE, (d) the adversarial example which will fool SCDCEBNN, (e) the adversarial 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 4.3 Average Minimum Estimated L2 Adversarial Distortion 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

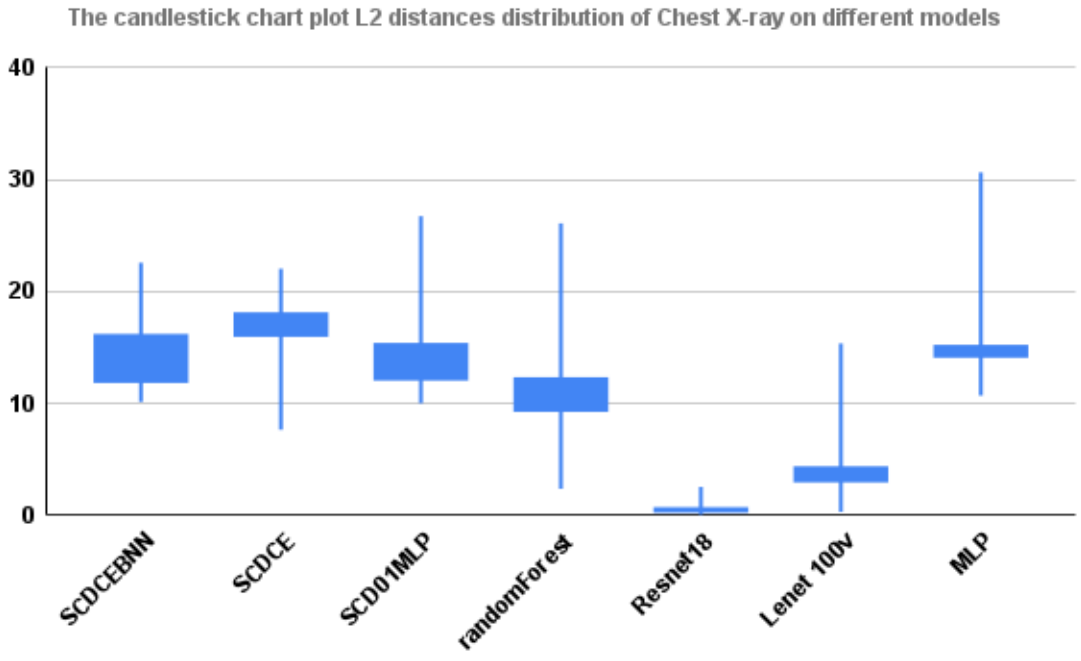


Figure 4.8 The candlestick chart plots the L2 distances on all adversarial images from BraTs18 on different models.

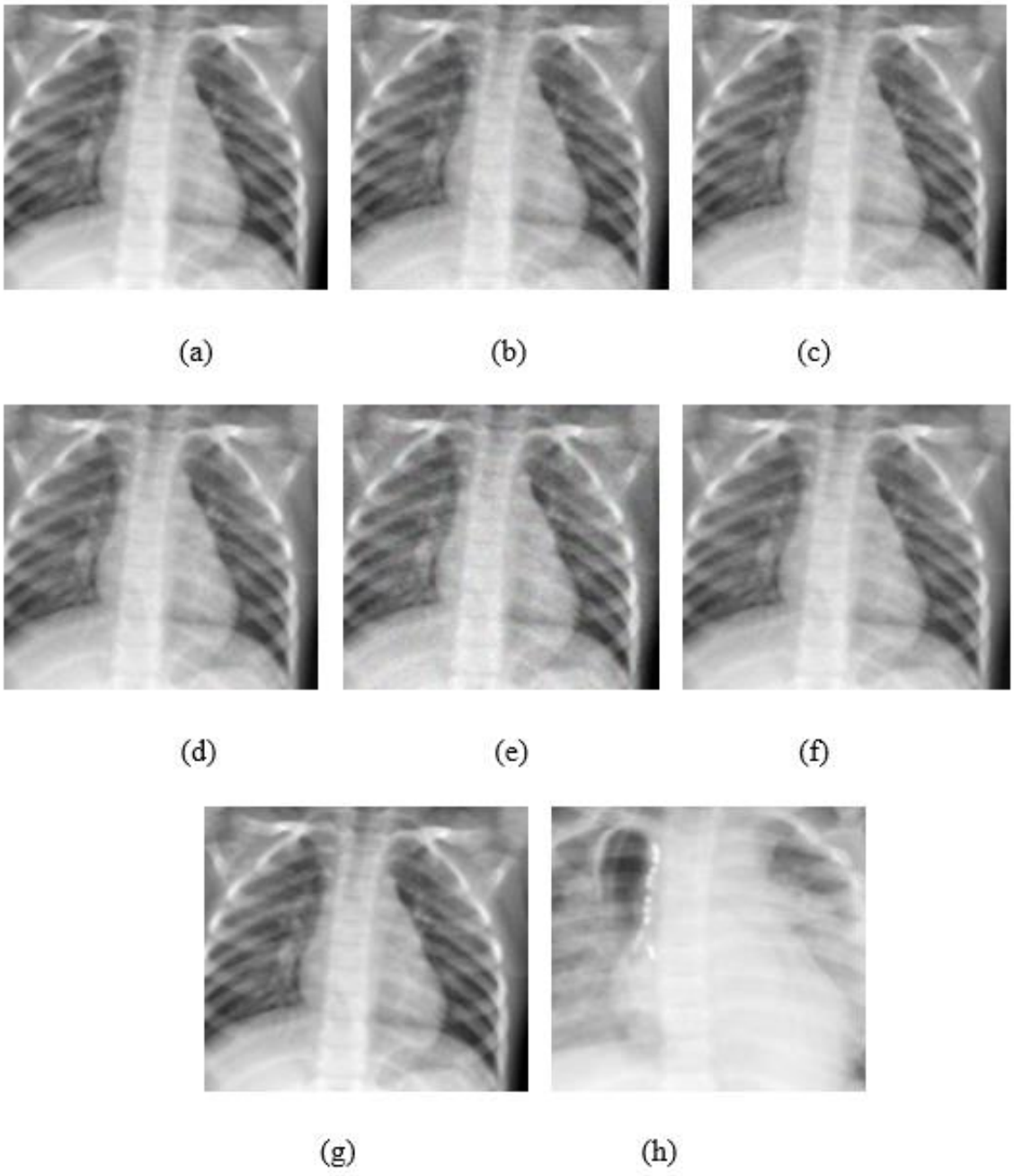


Figure 4.9 Visualizations of original and adversarial images among different networks from the Chest X-ray dataset. (a) The original image, (b) the adversarial example which will fool SCD01, (c) the adversarial example which will fool SCDCE, (d) the adversarial example which will fool SCDCEBNN (e) the 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 4.4 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

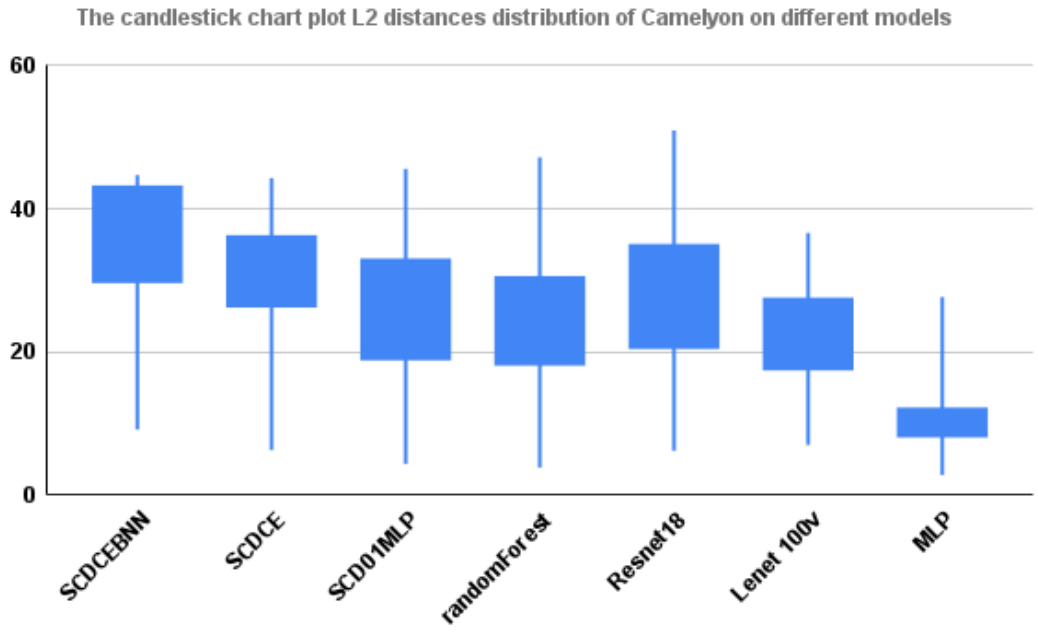


Figure 4.10 The candlestick chart plot the L2 distances on all adversarial images from BraTs18 on different models.

Table 4.5 lists the average minimum estimated L2 adversarial distortion on all three datasets. Again, the distortions of the SCD models are even higher, with SCDCEBNN taking the lead and twice better than all other models.

Table 4.5 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

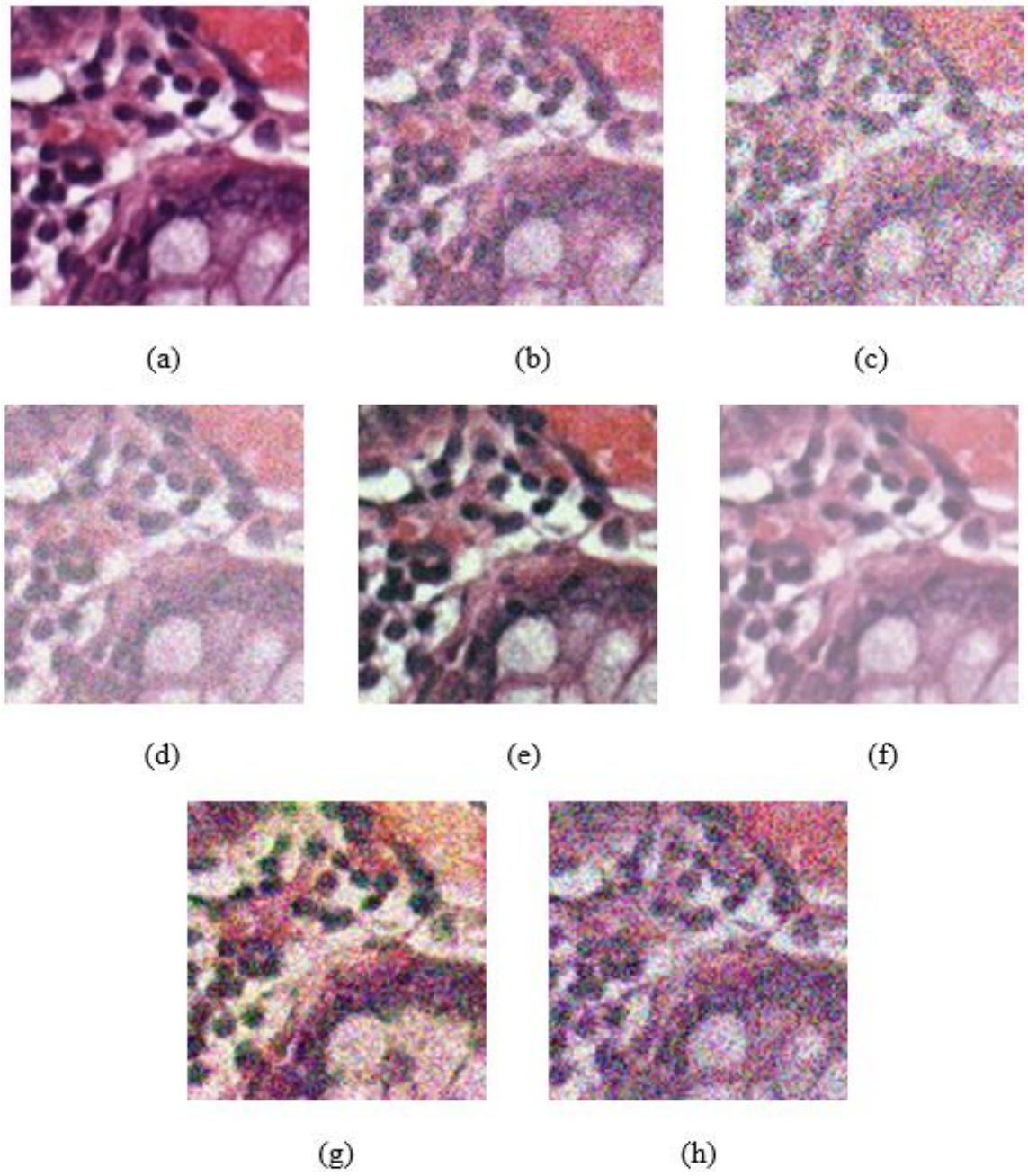


Figure 4.11 Visualizations of original images and adversarial images among different networks from colorectal histopathology dataset. (a) The original image, (b) the adversarial example which will fool SCD01, (c) the adversarial example which will fool SCDCE, (d) the adversarial example which will fool 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.

4.4.3 Evaluation of Defense Ability by Transferability

Another evaluation is to make use of the transferability property [13]. For example, 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 entirely different manners or on different datasets. There have been many available methods to construct adversarial examples and make networks robust against adversarial examples. However, 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 test these adversarial examples as misclassifications. If a model *G* can detect the adversarial examples from another model *F* and classify them correctly, it is more robust against adversarial attacks. Table 4.6 shows the results for classifying one random image and all adversarial examples. A random image can be classified by all models correctly, 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'. The proposed models SCD01MLP and SCDCE can detect all adversarial examples and classify them correctly.

Table 4.6 Results for Classifying One Random Image and All Adversarial Examples

	SCD01	SCDCE	SCDCEBNN	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 4.7 shows the average accuracy of all models when classifying the adversarial examples. The proposed models have higher accuracy, 88.89% and 85.19%, respectively. They can identify fake examples and are hard to be fooled by adversarial attacks. Other

models like MLP and LeNet are also not wrong but lower than our proposed models, Resnet18 is significantly lowest.

Table 4.7 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%

4.5 Conclusions

This chapter presents robust models to adversarial attacks in MRI images, chest X-rays, and histopathology images. The preliminary experiments 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 the proposed models' accuracies are more competitive than others, and the adversarial attack can easily be detected on our models.

CHAPTER 5

ADAPTIVE IMAGE RECONSTRUCTION FOR PROACTIVELY DEFENDING AGAINST ADVERSARIAL ATTACKS

5.1 Adversarial Attacks Defend Mechanism

The DNN-based image classifiers can misclassify the adversarial examples well-crafted by adversarial attacks, as discussed in previous chapters. To defend against adversarial attacks proactively, some researchers proposed the techniques of improving model robustness and detecting malicious behaviors.

5.1.1 Model-Specific Defense Mechanism

The model-specific defense mechanism is to normalize the parameters of a particular model through adversarial training. It makes the models more robust to the specific attacks corresponding to training samples. However, the disadvantage is obviously, for instance, the amount of training time consumed on individually training a model to the possible type of attack, and potential attackers can still attack the trained defense model by calculating its gradient.

The common defense methods include adversarial training (AT) [21] and equip CNN models to detect adversarial examples [21-25]. Nevertheless, *adversarial training* requires as many adversarial examples as possible to retrain the network and improve classification accuracy. Although Xu *et al* [25] proposed training the detection modules equipped into CNN models, this technique requires modifying the model and many adversarial examples. The retraining process requires a set of adversarial examples and the corresponding benign ones.

These methods achieve competitive results in defending against adversarial attacks. Nonetheless, the retraining time and the large requirements on computational ability are not realized. Besides that, the performances rely on the comprehensive prior knowledge of various potential adversarial techniques. The worst thing is revamped attack algorithms, which are usually unknown to the defender in advance. Once the brand new adversarial samples are generated, the model-specific defense techniques have to take extra time to retain or rebuild the models. This is a good chance for the adversarial examples to evade the classifier.

5.1.2 Model-Agnostic Defense Mechanism

Model-agnostic defense mechanism aims to eliminate or reduce adversarial perturbations by preprocessing the input, such as JPEG compression and high-level representation guided denoiser (HGD) [45]. It usually requires only a tiny calculation and retains the models for different attack types. Nevertheless, compression will reduce the clean classification accuracy, and one defend method cannot be effective on all attackers, for instance, the HGD [45] method only against BIM and FGSM attacks.

There are other recent methods to reconstruct the input images [26-29] to improve the robustness of the classifier. However, the noise added to the images varies widely. Some reasons lead to these problems. The first reason is that different adversary attackers generate different noises. Secondly, different target models will have different defense abilities. Moreover, the compelling adversarial images from the semantic and grayscale datasets are very different for attackers. Therefore, no one reconstructed strategy can remove all possible noise, which may sometimes lead to a new misclassification.

5.2 The Adversarial Perturbations and Removal

5.2.1 The Perturbations from Adversarial Attacks

The difference between the original and the crafted adversarial images looks like noises after visualizing. In Figure 5.1, the first column is the clean image. The middle column is the adversarial images generated by the Hopskipjump black-box attack on the ResNet18 model. The last column is the noise the attacker adds. It suggests that adversarial perturbations are difficult to detect by human eyes at the pixel level but lead to substantial noise at feature levels.

The perturbations from the adversarial attack are added layer by layer until the classifier makes a wrong decision. Different datasets require different scale perturbations even trained by the same model and attacker. For the same images trained on the same model, the effective fake samples are different with various attackers. We can observe that the noises added on the two different datasets in Figure 5.1 (c) and (f) are distinct, even though they are both crafted by the Hopskipjump attack on the ResNet18 model.

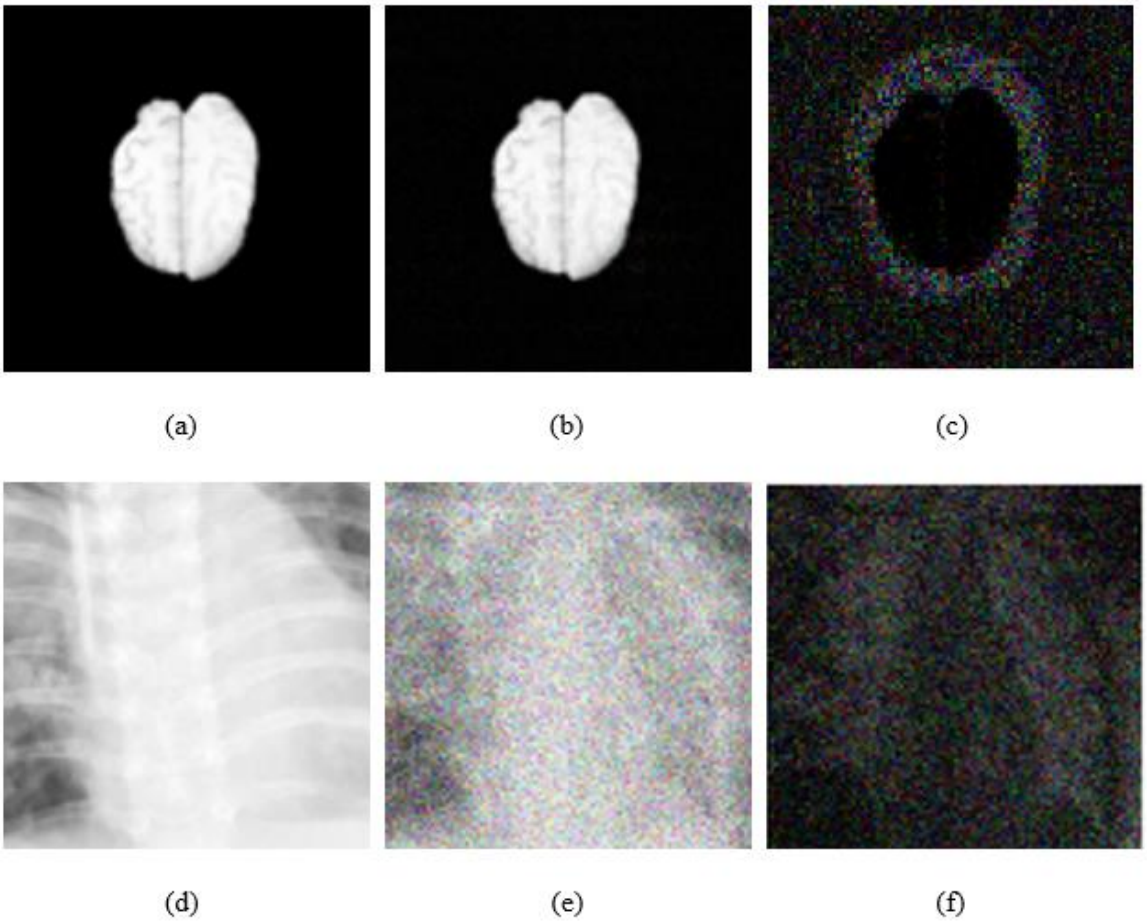


Figure 5.1 The perturbations on the image. (a) The clean image from BraTs18, (b) the adversarial image crafted by Hopskipjump black-box attack on the ResNet18 model, (c) the noise added by the attacker, (d) the clean image from ChestXray, (e) the adversarial image crafted by Hopskipjump black-box attack on the ResNet18 model, and (f) the noise added by the attacker.

5.2.2 Perturbation Removal

Let *G* denote an adversarial attacker, who generates a fake sample $I'(x, y)$ to fool the model *F* by adding perturbation $\varepsilon(x, y)$ on the original image $I(x, y)$ as

$$I'(x, y) = I(x, y) + \varepsilon(x, y) \tag{5.1}$$

where *x* and *y* denote the spatial coordinates. Note that the perturbation $\varepsilon(x, y)$ varies largely among attacker *G*, classifier *F*, and the clean input *I*. Assume that the model *F* can correctly

classify any clean instance. By applying the reconstruction techniques to remove $\varepsilon(x, y)$, we can generate a reconstructed image $I''(x, y)$, whose classified result is identical to the clean image I as

$$F(I) = F(I'') \tag{5.2}$$

Note that it is impossible to apply a single technique to remove all different perturbations. Image reconstruction techniques can reduce some noises, but over denoising happens since the noises are different among variant cases. Therefore, the reconstructed image $I''(x, y)$ is a higher quality image than the adversarial one $I'(x, y)$, but a lower quality than the original clean image $I(x, y)$. Our goal is to minimize the difference $d(x, y)$ as

$$d(x, y) = I(x, y) - I''(x, y) \tag{5.3}$$

In this way, the model F has a higher chance to make the correct decision. Therefore, the challenge is to remove as many perturbations $\varepsilon(x, y)$ as possible when an unknown input is given with added unknown noises.

5.2.3 Entropy Value

We use Shannon entropy [46, 47] to measure images' uncertain distributions and complexity features to analyze the internal information characteristics. If the entropy is high, it means the image includes more information. Let the gray level k have a probability P_k. The entropy H is calculated as in Equation (5.4),

$$H = -\sum_k P_k * log_2(P_k) \tag{5.4}$$

$$k = \sum_{(x,y)} k_{x,y} \tag{5.5}$$

The probability of the pixel $k(x, y)$ is $P_{xy} = k(x, y)/k$. Therefore, we have

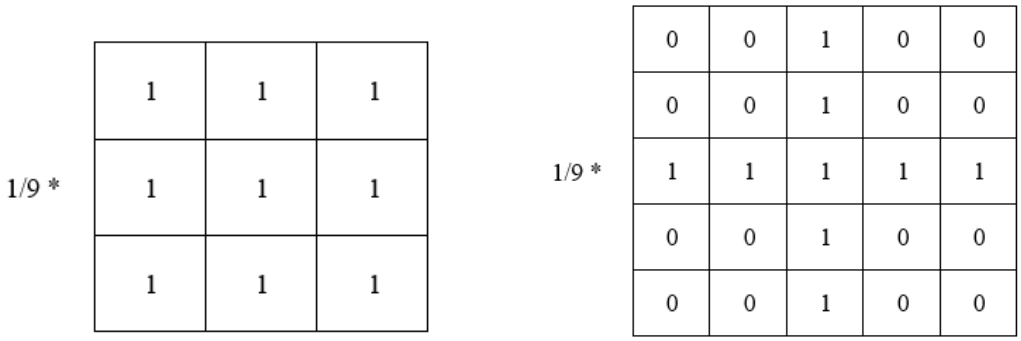
$$H(x, y) = log_2(k) - \frac{1}{k}\sum_{(x,y)} k_{x,y} * log_2(k_{x,y}) \tag{5.6}$$

The minimum entropy value of zero occurs when the image pixel value is constant in any location. The maximum entropy value is related to the total number of grayscales. For instance, in an image with 256 gray levels, the maximum entropy is $log_2(256) = 8$. Figure 5.2 shows the entropy values among different images, where (a) has a larger entropy than (c), indicating it contains less interest information. Note that the difference between the clean and adversarial images of (a) and (b) is greater than that of (c) and (d). We can conclude that the simpler images need larger perturbations to produce successfully adversarial samples. In contrast, color images are usually easy to craft with tiny noises.

Figure 5.2 The entropy values among different images. (a) The clean image with the entropy of 1.5222, (b) the adversarial image crafted by FGSM attacker with the entropy of 5.4241, (c) the clean image from ImageNet with the entropy of 4.9742, and (d) the adversarial image crafted by FGSM attacker with the entropy of 7.5302.

5.2.4 Adaptive Smoothing

Image smoothing intends to reduce and suppress image noises. The average smoothing filter is shown in Figure 5.3, where (a) is a square mask of 3 × 3 and (b) is a plus-shape mask of 5 × 5.



(a) **(b)**

Figure 5.3 Different average smoothing filter masks. (a) A square mask of 3 × 3 and (b) a plus-shape mask of size 5 × 5.

The Gaussian template is another method to reduce the blur in the smoothing process and obtain a more natural smoothing effect. The average smoothing treats the same weight to all the pixels in the neighborhood. Therefore, it is natural to think about increasing the weight of the neighbors close to the center and reducing the weight of distant neighbors. Figure 5.4 shows a 3 × 3 Gaussian template.

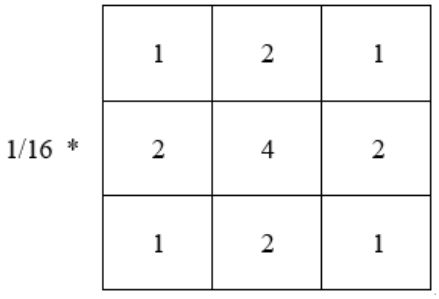


Figure 5.4 The 3 × 3 Gaussian mask template.

Increasing filter size to reduce the noises is an approach to remove as many noises as possible, but this method also causes the image over blurring. We use the adaptive Wiener filter to reconstruct data by removing noises without significantly blurring the structures in the image. The neighborhood $a(n1, n2)$ is used to estimate local mean μ and variance σ^2 as

$$\mu = \frac{1}{n*m}\sum_{n1,n2\in x} a(n1, n2) \tag{5.7}$$

$$\sigma^2 = \frac{1}{n*m}\sum_{n1,n2\in x}(a(n1, n2) - \mu)^2 \tag{5.8}$$

where x is the n-by-m local neighborhood of each pixel. The pixel-wise filter to suppress noise is represented as

$$b(n1, n2) = \mu + \frac{\sigma^2-\gamma^2}{\sigma^2}(a(n1, n2) - \mu) \tag{5.9}$$

where γ^2 is the noise variance, which is default as the average of all local variance in Equation (5.8).

5.3 The Proposed Method

5.3.1 The Architecture

Figure 5.5 shows the proposed framework. If the classified results of the original image *X(x, y)* and the related reconstructed image *X''(x, y)* are the same, we say *X(x, y)* is a clean image; otherwise, it is an adversarial sample.

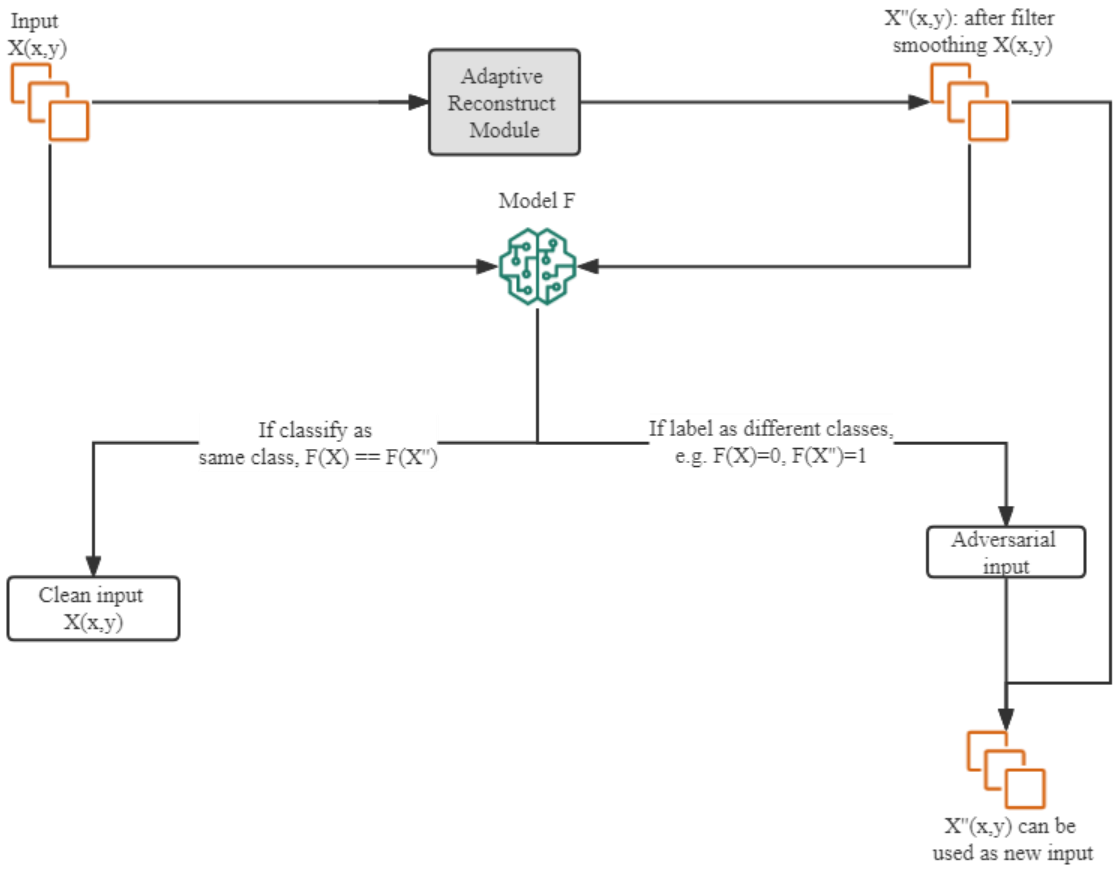


Figure 5.5 Detect adversarial samples by comparing the original classified result and the reconstructed one.

Two factors decide the reliability of this detection scheme. One is the robustness of model *F*. The previous chapters presented that the ability to defend against adversarial attacks differs vastly on different models. For example, LeNet has a higher tolerant ability than ResNet18, but is similar to MLP classifier. This means the reconstructed image *X'' (x, y)* cannot be classified by ResNet18, but may work on LeNet. The other is the reconstructed technique, such that the perturbations added on the clean images are random, and all the removed noises made by preprocessing methods are different in different scenarios.

This chapter proposes a novel reconstruction module with an adaptive process, as shown in Figure 5.6, which applies variant smoothing filters on different inputs. The

images of high entropy values require few distortions, and those of low entropy values need large perturbations. Based on different entropy values, we can decide if the smoothing filters should be applied or what kind of filters can be applied.

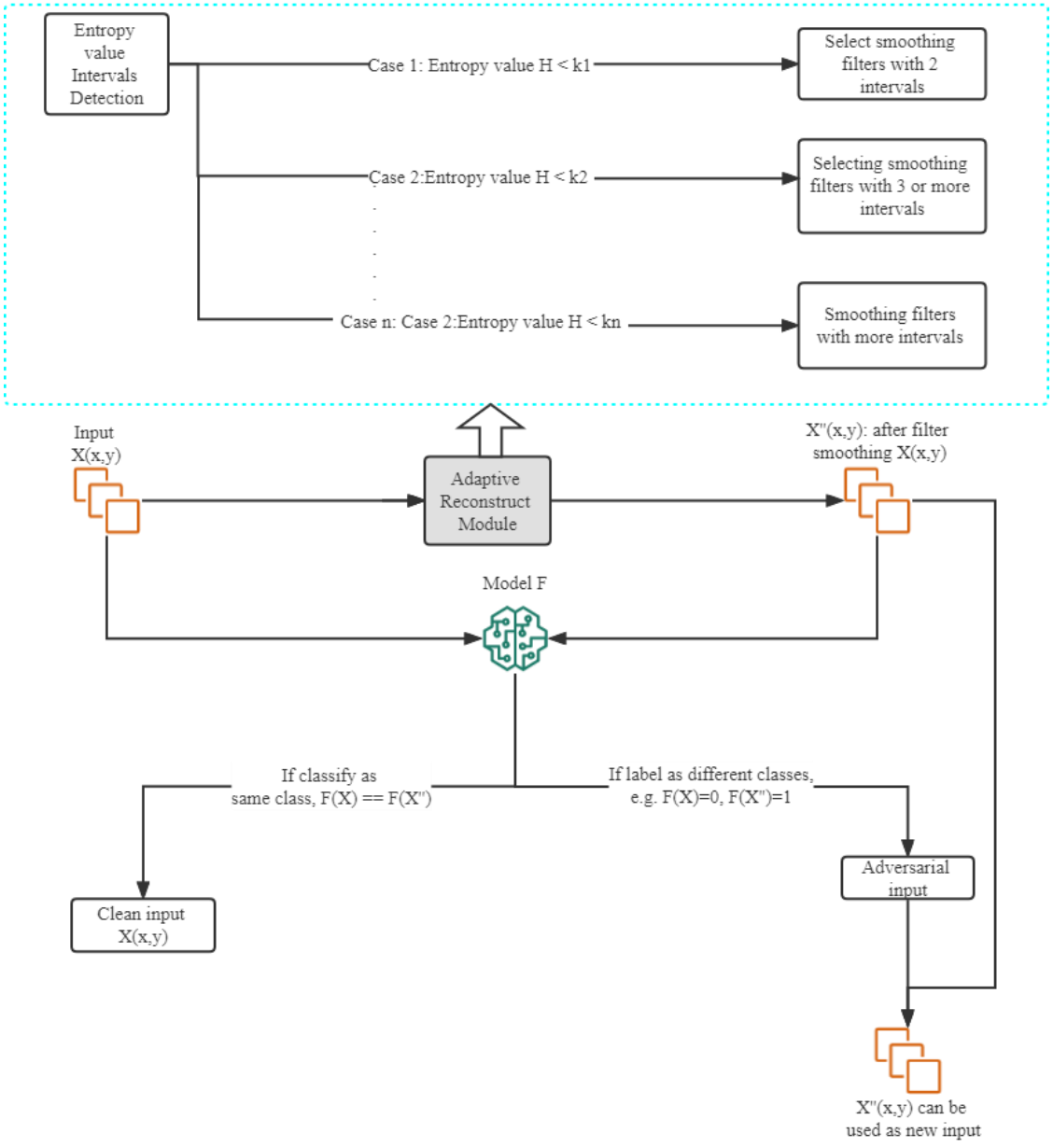


Figure 5.6 The architecture of adaptive reconstruction module. Computing and training the entropy values are divided into different intervals. In different cases, the module will apply different filters to reconstruct the images.

5.3.2 Parameters Setting

As discussed in previous sections, the filters may excessively smooth while reconstructing the images, and the models may return a new misclassification. To solve this problem, the uniform method is adopted to divide the entropy value into intervals evenly, and each interval shares the same size. Entropy values decide the quantities of intervals. The following adaptive filter selection algorithm is developed to select a filter for the image with different entropy values.

Algorithm 5.1

Adaptive Filter Selection Algorithm

Input: H: Entropy value; k: Intervals; F: Initiate Spatial filter; I(x,y): Original image input; F_i: another filter; i: the quantities of filters predefined; n: Entropy value; m1: Interval number; m2: Interval number

Output: I'(x, y): the smoothed image

```
if H < n and k ==m1,
    return I(x, y)
if H<n+1 and k ==m2,
    return I(x, y)
for all i filters:
    if abs(I(x, y) – F1(I (x, y)) ) <= abs( I(x, y) – Fi(I (x, y))),
        return I'(x, y) = F1(I(x,y)), break
    otherwise, F1(I (x, y)) = Fi(I (x, y))
end
```

The proposed algorithm tests MINIST and ImageNet-subset (i.e., two classes) images to set the interval size threshold values. FGSM and PGD attack ResNet18 to craft adversarial images and apply the standard convolution spatial filer of size 5 × 5. The entropy values of the MNIST dataset are smaller than 4, and those of the ImageNet is larger than 6.

The experiments test different interval numbers and use MNIST and ImageNet-subset as the experimental baseline. Since the entropy values of MNIST are all smaller than 4, and those of ImageNet are typical more extensive entropy value set; i.e., all are greater than 6. As shown in Table 5.1, the accuracy of MNIST is continuously high. Although it is improved with increasing intervals but not over 1%, it can conclude that two intervals are good enough. On the other hand, the accuracy of the ImageNet-subset is enhanced when the number of intervals starts from six.

Table 5.1 The accuracy of MNIST and ImageNet-subset under different numbers of intervals

	The number of intervals						
Dataset	2	3	4	5	6	7	8
MNIST	97.14%	97.28%	97.42%	97.57%	97.64%	97.65%	97.71%
ImageNet-Subset (2 classes)	54.64%	78.05%	81.96%	85.86%	92.10%	92.88%	92.96%

5.4 Experimental Results

This section presents experiments on the three popular medical imaging datasets: BraTs18, Chest X-rays, and Colorectal Histopathology, as described in previous chapters, to evaluate the proposed algorithm. Two networks, ResNet18 and SCD01MLP, are used to evaluate

the defense performance. It has been shown that ResNet18 has the worst defense ability, while SCD01MLP is the best.

5.4.1 Evaluation Against White-box Attacks

Both networks are trained on the clean datasets, and the classification accuracies are all higher than 90%. Because generating the PGD adversarial examples does not work on non-convex SCD models, the adversarial examples will not consider generated from SCD models. Table 5.2 evaluates the accuracy of ResNet18 on detecting white-box PGD attack examples. The adaptive smoothing filter works better by applying different filters and obtaining that. Table 5.3 shows the accuracy of SCD01MLP on detecting white-box PGD attack examples. The adaptive filters work better on BraTS18 datasets, but the Gaussian smoothing filters perform better on human colorectal histopathology images.

5.4.2 Evaluation Against Black-box Attacks

Tables 5.4 and 5.5 list the accuracies of detecting black-box BIM attack examples. The experiments set up four different filters and trained both networks on a clean dataset to achieve comparative classification accuracies. The adversarial examples are classified on the pertained models without defense methods. The detection accuracy shows that the adaptive and Gaussian filters work better on three datasets, but the adaptive filters perform better in removing black-box perturbations

Table 5.2 Performance of Detecting PGD Adversarial Examples with Different Smoothing Filters on ResNet18

Accuracy	Clean Accuracy on ResNet18	3*3 Spatial Average Smoothing	5*5 Spatial Average Smoothing	3*3 Gaussian Smoothing	Adaptive Smoothing with Mean
Colorectal Histopathology	99.60%	25%	25%	54.16%	54.16%
BraTs18	99.64%	29.16%	29.16%	66.67%	70.83%
ChestXray	94.32%	16%	25%	41.66%	41.66%

Table 5.3 Performance of Detecting PGD Adversarial Examples with Different Smoothing Filters on SCD01MLP

Accuracy	Clean Accuracy on SCD01MLP	3*3 Spatial Average Smoothing	5*5 Spatial Average Smoothing	3*3 Gaussian Smoothing	Adaptive Smoothing with Mean
Colorectal Histopathology	99.20%	29.16%	54.16%	66.67%	54.16%
BraTs18	98.38%	29.16%	66.67%	79.16%	79.16%
ChestXray	90.69%	25%	41.66%	41.66%	41.66%

Table 5.4 Performance of Detecting BIM Adversarial Examples with Different Smoothing Filters on ResNet18

Accuracy	Clean Accuracy on ResNet18	3*3 Spatial Average Smoothing	5*5 Spatial Average Smoothing	3*3 Gaussian Smoothing	Adaptive Smoothing with Mean
Colorectal Histopathology	99.60%	41.66%	45.83%	58.33%	58.33%
BraTs18	99.64%	41.66%	50%	62.50%	62.50%
ChestXray	94.32%	37.50%	37.50%	37.50%	45.83%

Table 5.5 Performance of Detecting BIM Adversarial Examples with Different Smoothing Filters on SCD01MLP

Accuracy	Clean Accuracy on SCD01MLP	3*3 Spatial Average Smoothing	5*5 Spatial Average Smoothing	3*3 Gaussian Smoothing	Adaptive Smoothing with Mean
Colorectal Histopathology	99.20%	45.83%	58.33%	58.33%	62.50%
BraTs18	98.38%	50%	62.50%	62.50%	62.50%
ChestXray	90.69%	45.83%	37.50%	37.50%	58.33%

5.4.3 Evaluation Against Non-Adversarial Samples

To evaluate whether the filters can excessively remove the information from images, the experiments evaluate the accuracies on the clean images. Table 5.6 shows that the filters decrease the accuracy, but not over 1%. The filter improves the classification results or promises the baseline accuracy.

5.4.4 Evaluation with Transfer Adversarial Samples

Transferability is another vital property of a machine learning system. A model has a robust defense ability if it can correctly classify the adversarial sample from other unknown models. We show the classification rates on transfer examples in Tables 5.7 and 5.8. The SCD01MLP can correctly classify most adversarial examples from ResNet18 attacked by the black-box attack. The best accuracy is 91.19%, and the worst is 83%. The classification rates of ResNet18 are improved up to 16.1%.

Table 5.6 Performance of Detecting Clean Examples with Different Smoothing Filters

Accuracy	Clean Accuracy on SCD01MLP	3*3 Spatial Average Smoothing	5*5 Spatial Average Smoothing	3*3 Gaussian Smoothing	Adaptive Smoothing with Mean
Colorectal Histopathology	99.20%	99.80%	98.69%	99.80%	99.20%
BraTs18	98.38%	99.41%	98.38%	99.41%	99.67%
Chest Xray	90.69%	92.23%	90.69%	93.78%	93.78%

Table 5.7 Performance of SCD01MLP on Detecting Transferred Adversarial Examples

Accuracy	Clean Accuracy on SCD01MLP	No Defense	3*3 Spatial Average Smoothing	5*5 Spatial Average Smoothing	3*3 Gaussian Smoothing	Adaptive Smoothing with Mean	Best case Improvement
Colorectal Histopathology	99.20%	85.19%	85.83%	83.33%	86.89%	86.89%	1.70%
BraTs18	98.38%	88.89%	89%	88.41%	89.00%	91.17%	2.28%
ChestXray	90.69%	82.13%	83.00%	83.00%	85.83%	83.00%	3.70%

Table 5.8 Performance of ResNet18 on Detecting Transferred Adversarial Examples

Accuracy	Clean Accuracy on ResNet18	No Defense	3*3 Spatial Average Smoothing	5*5 Spatial Average Smoothing	3*3 Gaussian Smoothing	Adaptive Smoothing with Mean	Best case Improvement
Colorectal Histopathology	99.60%	26.60%	31.66%	25.83%	38.33%	38.33%	11.73%
BraTs18	99.64%	38.89%	45.83%	43.92%	52.50%	52.50%	13.61%
ChestXray	94.32%	21.40%	31.66%	31.66%	37.50%	37.50%	16.10%

This section detects the PGD adversarial examples generated from ResNet18 models in the MNIST dataset to keep the fair scenario. Table 5.9 shows our approach is slightly lower than the adversarial training but higher than others. The classification performance on adversarial training is difficult to defeat. However, *adversarial training* takes longer time and requires a sufficient number of adversarial examples to retrain the networks. Furthermore, when a new attack technique is developed, *adversarial training* has to retrain the models again to improve the performance. Compared with that, the proposed method has a distinct advantage in consideration of training time and computation consumption.

Table 5.9 Performance Comparisons with Other Methods

Method	Performance
Feature Denoising for Improving Adversarial Robustness	55.70%
Defense against Adversarial Attacks by Reconstructing Images	48%~98%
Low-rank Completion of High-Sensitivity Points	62.20%
Adversarial Training	98.00%
Ours	**97.14%**

5.5 Conclusion

This chapter proposes a novelty adaptive framework that can reconstruct the images and proactively detect the adversarial attacks in advance. By comparing the classifying results of the original images and the reconstructed ones, the proposed method can detect the

adversarial samples successfully. Furthermore, an adaptive reconstruction process is developed based on the entropy values to avoid excessive noise reduction during reconstructing images. The significant contributions of the proposed method are that it does not need to retrain the model and can be assembled into any network.

The performances are evaluated in both white-box and black-box attack scenarios. In addition, different target models are conducted in the proposed mechanism to show the universal feature. Experimental results show significant improvement on correctly classifying the adversarial samples and providing a higher detecting accuracy than the existing techniques.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

This dissertation presents three neural network frameworks with higher classification accuracy than state-of-the-art models. The dual path residual network, random depth wise convolutional neural network, and Stochastic Coordinate Descent (SCD) networks perform better than state-of-the-art.

Meanwhile, the SCD networks are more robust than the traditional neural networks when conducting the black-box attack experiments. This evaluation demonstrates that the SCD can defend against adversarial attacks more efficiently.

A novel adaptive image reconstruction defense mechanism is proposed to address the misclassifications caused by defense schemes. This scheme does not require prior knowledge or retrain the models and can ensemble in any network. It is more proactive and efficient than the traditional defense against methods.

6.2 Future Work

Even though the adaptive image reconstruction defense mechanism achieves more outstanding performance in some datasets, it is necessary to extend future works on more adversarial attacks and target models, including larger cohort datasets. Moreover, the adaptive reconstruction algorithm needs to develop a more efficient optimization method for training entropy values and filter selection. Currently, only binary classification results

are shown in this dissertation. Because of that, the zero-one loss being directly applied in multi-class problems will result in massive local minima during training. Future work aims to find a better multi-classifier to solve secure multi-class classification problems.

REFERENCES

[1] L. Yann, Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, MIT Press, Cambridge, MA, USA, 3361(10):193202, 1995.

[2] A. Krizhevsky, I. Sutskever, G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in neural information processing systems*, 2012.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going deeper with convolutions," *Proc. the IEEE conference on computer vision and pattern recognition(CVPR)*, Boston, MA, USA, June, 2015.

[4] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[5] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition." *Proc. IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, June, 2016.

[6] N. CF. Codella, D. Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo et al, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," *IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pp. 168-172, Washington, DC, USA, May, 2018.

[7] K. Matsunaga, A. Hamada, A. Minagawa and H. Koga, "Image classification of melanoma nevus and seborrheic keratosis by deep neural network ensemble," *International skin imaging collaboration (isic) 2017 challenge at the international symposium on biomedical imaging (ISBI)*, Washington, DC, USA, May, 2018.

[8] I.G. Diaz, "Incorporating the knowledge of dermatologists to convolutional neural networks for the diagnosis of skin lesions," *International skin imaging collaboration (isic) 2017 challenge at the international symposium on biomedical imaging (ISBI)*, Washington, DC, USA, May, 2018.

[9] A. Menegola, J. Tavares, M. Fornaciali, LT. Li, S Avila, E. Valle, "RECOD Titans at ISIC Challenge 2017," *International skin imaging collaboration (isic) 2017 challenge at the international symposium on biomedical imaging (ISBI)*, Washington, DC, USA, May, 2018.

[10] O. Oktay, E. Ferrante, K. Kamnitsas, M. Heinrich, W. Bai, J. Caballero, Cook SA, A. De Marvao, T. Dawes, DP O'Regan, B. Kainz, "Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation," *IEEE transactions on medical imaging*, vol. 37, no. 2, pp. 384-395, Feb. 2018.

[11] 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.

[12] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," *Proc. IEEE symposium on security and privacy*, San Jose, CA, USA, pp. 39–57. March 2017.

[13] 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.

[14] 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.

[15] 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.

[16] G. Bortsova, C. González-Gonzalo, S.C. Wetstein, F. Dubost, I. Katramados, L. Hogeweg, B. Liefers, B. van Ginneken, J.P. Pluim, M. Veta, C.I. Sánchez, "Adversarial attack vulnerability of medical image analysis systems: Unexplored factors," arXiv preprint arXiv:2006.06356, 2020.

[17] 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, Sep, 2018.

[18] 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.

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

[20] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, “Ensemble adversarial training: attacks and defenses,” *Proc. 6th Internatianl. conference on learning representations (ICLR),* Vancouver, BC, Canada, May, 2018.

[21] J. Wang, T. Zhang, S. Liu, P.Y. Chen, J. Xu, M. Fardad, and B. Li, “Beyond adversarial training: Min-max optimization in adversarial attack and defense,” *Nuclear Physics*, Section A, 2019.

[22] Y. Shi and Y. Han, “Schmidt: image augmentation for black-box adversarial attack,” *IEEE International conference on multimedia and expo*, San Diego, USA, pp. 1–6, 2018.

[23] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, “The limitations of deep learning in adversarial settings,” P*roc. 2016 IEEE European symposium on security and privacy,* pp. 372–387, 2016.

[24] 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,* pp. 582–597, 2016.

[25] W. Xu, D. Evans, and Y. Qi, “Feature squeezing: Detecting adversarial examples in deep neural networks,” *Proc. Network and distributed system security symposium*, 2018.

[26] F. Liao, M. Liang, Y. Dong, and T. Pang, “Defense against adversarial attacks using high-level representation guided denoiser”. *2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR),* Salt Lake City, UT, USA, 2018.

[27] C. Xie, Y. Wu, L. V. D. Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," *Proc. IEEE/CVF conference on computer vision and pattern recognition,* Long Beach, CA, USA, 2019.

[28] S. Zhang, H. Gao, and Q. Rao, "Defense against adversarial attacks by reconstructing images," *IEEE transactions on image processing,* pp.6117-6129, 2021.

[29] Z. Zhao, H. Wang, H. Sun, J. Yuan, Z. Huang, and Z. He, "Removing adversarial noise via low-rank completion of high-sensitivity points," *IEEE Transactions on image processing,* pp.6485-6497, 2021.

[30] D. N Louis, A. Perry, G. Reifenberger, A. V. Deimling, D. Figarella-Branger, W. K Cavenee, H. Ohgaki, O. D Wiestler, P. Kleihues, and D. W Ellison, "The 2016 world health organization classification of tumors of the central nervous system: a summary," *Acta neuropathologica*, 131(6):803–820, 2016.

[31] P. Kleihues, D. N Louis, B. W Scheithauer, L. B Rorke, G. Reifenberger, P. C Burger, and W. K Cavenee, "The WHO classification of tumors of the nervous system," *Journal of neuropathology & experimental neurology*, 61(3):215–225, 2002.

[32] D. J Brat, K. Aldape, H. Colman, E. C Holland, D. N Louis, R. B Jenkins, BK Kleinschmidt-DeMasters, A. Perry, G. Reifenberger, R. Stupp, et al. "cIMPACT-NOW update 3: recommended diagnostic criteria for diffuse astrocytic glioma, idh-wildtype, with molecular features of glioblastoma," *WHO grade, Acta neuropathologica*, 136(5):805–810, 2018.

[33] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. Min Ha, M. Rozycki, 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.

[34] Y. Xue, M. Xie, F. Farhat, O. Boukrina, A. M. Barrett, J. R. Binder, U. W. Roshan, and W. Graves, "A multi-path decoder network for brain tumor segmentation," *MICCAI BraTS 2019 challenge*, 2019.

[35] B. H Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al, "The multi-modal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[36] Y. Xue, U. Roshan, "Image Classification and Retrieval with Random Depthwise Signed Convolutional Neural Networks," *International Work Conference on Artificial Neural Networks*, Springer, 492–506, 2019.

[37] Corel-Princeton Image Similarity Benchmark, Princeton University Computer Science Department http://www.cs.princeton.edu/cass/benchmark/, Retrieved on June, 2019.

[38] A. Janowczyk, A. Basavanhally, and A. Madabhushi, "Stain normalization using sparse autoencoders (StaNoSA): Application to digital pathology," *Computerized Medical Imaging and Graphics,* Vol 57,50–61, 2017.

[39] F. A Spanhol, L. S Oliveira, C. Petitjean, and L. Heutte, "A dataset for breast cancer histopathological image classification*," IEEE Transactions on Biomedical Engineering* 63, 7 ,1455–1462, 2015.

[40] M. Xie and U. Roshan, "Exploring classification, clustering, and its limits in a compressed hidden space of a single layer neural network with random weights," *International work-conference on artificial neural networks*, Gran Canaria, Spain, 2019.

[41] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," *International conference on learning representations (ICLR),* 2018.

[42] 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. 12771294, May 2020.

[43] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," arXiv preprint arXiv:1611.01236, 2016.

[44] F. Tramer and D. Boneh, "Adversarial training and robustness for multiple perturbations," Proc. Advances in neural information processing systems, Vancouver, Canada, pp. 58585868, December 2019.

[45] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," *Proc. IEEE conference on computer vision and pattern recognition*, pp. 1778-1787. 2018.

[46] D. Y. Tsai, Y. Lee, and E. Matsuyama, “Information entropy measure for evaluation of image quality,” *Journal of digital imaging*, vol. 21, no. 3, pp. 338–347, 2008.

[47] H. L. Cooper and M. I. Miller, “Information measures for object recognition accommodating signature variability,” *IEEE Transactions on Information theory*, vol. 46, no. 5, pp. 1896–1907, 2000.