International Journal of Computing and Digital Systems
ISSN ((2210-142X)
Int. J. Com. Dig. Sys. 15, No. 1 (Mar-2024)

http://dx.doi.org/10.12785/ijcds/150190

RDMAA: Robust Defense Model against Adversarial Attacks in Deep Learning for Cancer Diagnosis

Atrab A. Abd El-Aziz[1], Reda A. El-Khoribi[2] and Nour Eldeen Khalifa[2]

[1]*Department of Information Technology, Faculty of Computers and Information, KafrelSheikh University, Egypt*
[2]*Department of Information Technology, Faculty of Computers and Artificial Intelligence, Cairo University, Egypt*

Received 14 Nov. 2023, Revised 28 Feb. 2024, Accepted 1 Mar. 2024, Published 10 Mar. 2024

Abstract: Attacks against deep learning (DL) models are considered a significant security threat. However, DL especially deep convolutional neural networks (CNN) has shown extraordinary success in a wide range of medical applications, recent studies have recently proved that they are vulnerable to adversarial attacks. Adversarial attacks are techniques that add small, crafted perturbations to the input images that are practically imperceptible from the original but misclassified by the network. To address these threats, in this paper, a novel defense technique against white-box adversarial attacks based on CNN fine-tuning using the weights of the pre-trained deep convolutional autoencoder (DCAE) called Robust Defense Model against Adversarial Attacks (RDMAA), for DL-based cancer diagnosis is introduced. Before feeding the classifier with adversarial examples, the RDMAA model is trained where the perpetuated input samples are reconstructed. Then, the weights of the previously trained RDMAA are used to fine-tune the CNN-based cancer diagnosis models. The fast gradient method (FGSM) and the project gradient descent (PGD) attacks are applied against three DL-cancer modalities (lung nodule X-ray, leukemia microscopic, and brain tumor magnetic resonance imaging (MRI)) for binary and multiclass labels. The experiment's results proved that under attacks, the accuracy decreased to 35% and 40% for X-rays, 36% and 66% for microscopic, and 70% and 77% for MRI. In contrast, RDMAA exhibited substantial improvement, achieving a maximum absolute increase of 88% and 83% for X-rays, 89% and 87% for microscopic cases, and 93% for brain MRI. The RDMAA model is compared with another common technique (adversarial training) and outperforms it. Results show that DL-based cancer diagnoses are extremely vulnerable to adversarial attacks, even imperceptible perturbations are enough to fool the model. The proposed model RDMAA provides a solid foundation for developing more robust and accurate medical DL models.

Keywords: Adversarial Attacks,Cancer Diagnosis, Deep Learning,Deep Convolutional Autoencoder

1. INTRODUCTION

Deep learning (DL) has gained popularity as a result of a trillion-fold increase in processing power. DL models are sophisticated that can outperform a variety of natural image analysis tasks, including image retrieval, object recognition, and image classification. Additionally, DL has been shown to work best in a variety of medical image applications, including cancer diagnostics, oncology, and radiology. This allowed the automation of some medical processes and the integration of DL systems in clinical settings. Therefore, the robustness and reliability of DL models are critical issues that must be addressed. For example, deep models frequently make incomprehensible mistakes in noisy conditions which result in serious unexpected effects [1].

Despite the higher performance of DL techniques, the research community has discovered a serious security problem in existing DL algorithms. According to recent research, DL systems are subject to so-called adversary attacks. DNNs can be fooled into producing inaccurate predictions with high confidence by slightly altered input instances. This has prompted safety concerns concerning the use of deep learning models in healthcare systems. Adversarial examples are slightly perturbated photos that look like the originals but were purposely created to fool previously trained models. Such attacks use a small well-crafted perturbation into the model inputs to produce a misclassification. The perturbations that are imperceptible to human vision are enough to cause the model to generate a high-confidence misclassification prediction [2],[3]. Medical safety is so important in clinical practice. So, the vulnerabilities and the security risks that come with implementing DL algorithms have received a lot of attention. It should be considered how deep diagnostic models are vulnerable to adversary attacks if the clinician is not involved in the diagnostic procedure at all. As a result of this vulnerability, additional chances for fraud may arise. Figure 1 provides an illustration scenario of an adversarial

E-mail address: atrab_ahmed@fci.kfs.edu.eg, r.abdelwahab@fci-cu.edu.eg, nourmahmoud@cu.edu.eg

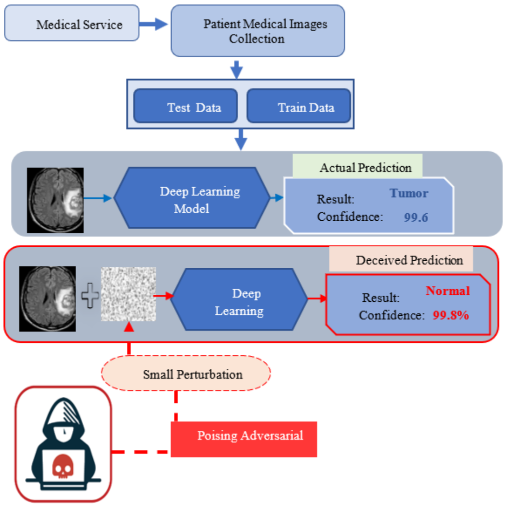


Figure 1. Representation of An Adversarial Attack Such as Malware on An Online DL-Based Medical Application

attack that can deceive a DL model into failing to identify the tumor type by adding a small perturbation to medical images. For instance, Diagnostic errors will aggravate the condition of patients and, at the same time, tarnish the credibility of healthcare organizations. Any online clinical system that uses a machine-learning algorithm for diagnosis could be fooled by these attacks [4]. While much of the existing research on adversary machine learning has focused on natural images, a complete understanding of adversary attacks in medical images is still a work in progress. Medical images can include domain-specific properties (different biological textures) that are different from natural photos [5].

According to a recent industry report: 30% of all cyberattacks on artificial intelligence (AI) systems would use AI model theft, training data poisoning, or adversary attacks through 2022. This is of particular concern for the healthcare business, which is expected to experience two to three times the average number of cyberattacks as other industries [6]. As a result, adversarial attacks could be a major concern in the medical field. This is due to two major factors: financial interests and technical weaknesses. The most common attacks are FGSM [7], PGD [8], Basic Iterative Method (BIM) [9], L-BFGS Attack [10], Carlini and Wagner (C&W) [11], DeepFool [12], and One Pixel Attack [13].

The main contributions of this paper are as follows:

- Introducing a novel defense model, RDMAA, specifically designed for securing DL-based cancer diagnosis systems against white-box attacks. This defense model employs fine-tuning of CNN with pre-trained parameters from a Deep Convolutional Autoencoder DCAE.
- Implementing parameter sharing between the convolution and max-pooling layers in both the DCAE encoder and CNN models, leading to a substantial reduction in training computations.
- Demonstrating that the proposed model effectively enhances prediction accuracy and model robustness by mitigating noise introduced by perturbed samples from prominent white-box attacks. This improvement is particularly notable when faced with adversarial samples.
- Evaluating the robustness of the defense technique across three cancer modalities MRI, lung nodule (X-ray), and Acute leukemia (microscopic) employing both binary and multi-class classifications.
- Conducting a comparative analysis between the proposed RDMAA model and the adversarial training technique. The results indicate that RDMAA outperforms adversarial training, underscoring its superiority in defending DL-based cancer diagnosis models against adversarial threats.

The rest of this paper is organized as follows: A brief literature review of related work will be discussed in Section 2. Next, the proposed model RDMAA will be discussed in detail in section 3. Then, the experimental results will be dis-cussed in section 4. Finally, a brief conclusion of the proposed work and future directions will be discussed in Section 5.

2. LITERATURE REVIEWS

While there have been many studies on adverse examples in natural images, there are far fewer in medical images. A typical scenario is a clinic that could alter medical images to force all patients to undergo surgery. In situations like this, algorithms must be confirmed to be accurate and adversarial examples that may have unfavorable results must be resolved. This section begins with a quick review of model threats, attack strategies, and many adversarial attacks and detections, followed by descriptions of defenses for medical images with different modalities.

A. Model Threat

The protection of any system is evaluated in association with the objectives and capabilities of its possible attacks. The threat model concept captures the attacker's capabilities, including his or her knowledge and goals. When assessing the adversary resilience of machine learning systems, specifically establishing the threat model under consideration helps to clearly define the attack area against which robustness is assessed, providing for claims that can be refuted. The following is the threat model that was used in this study:

Goal: It's supposed that the attacker's purpose is to create a widespread misclassification, which is known as an untargeted adverse attack. The purpose of an untargeted adversary is to change the entry image so that it can be predicted as any class other than the ground truth class. However, the objective of the directed adversary is to change the entry so that it can be classified as a special class.

Capabilities: It's supposed that the capabilities of the attacker are as follows: the attacker has only the ability to change the input images of the victim system which is fed directly to DL networks. The attacker can alter the input images in a very crafty way, and imperceptible to the human eye.

Knowledge: situations are created where the attacker is aware of the medical data information and the victim model network architecture. In all attack scenarios we investigated, the attacker cannot modify the structure of the target model.

B. *Attacks Strategies*

Maliciously constructed inputs are used to attack a target model in the adversarial scenario. They cause small changes to the original inputs which can deceive the target model. One attack strategy is to maximize the classification error of the DNN model given a pre-trained DNN model.

1) *White-box Attack*

In a white-box situation, the attacker has full access to the architecture, parameters, gradients, and other data of the target neural network. The adversary can deliberately build adversarial examples using the knowledge of the network. Because exposing the architecture and parameters of the model helps people understand the weaknesses of DNN models explicitly and can be mathematically evaluated, white box attacks have been intensively investigated [14].

2) *Black-box Attack*

The internal configuration of DNN models isn't available to adversaries in a black box attack scenario. The adversaries can only supply data to the models and view their results. They often attack models by continuing to feed samples into the box and analyze the output to exploit the input-output relationship of the model and identify its flaws. Black box attacks are more viable in applications than white-box attacks, as model designers rarely open their model parameters for proprietary reasons [15].

3) *Gray Attack*

In a semi-white box or gray box attack, the attacker develops a generative model to generate examples of adversaries in a white box environment. After training the generative model, the attacker no longer needs a victim model and can create adverse samples in a black box situation [14],[15]

C. *Existing Adversarial Attacks on Medical Images*

The authors in [16] have investigated the effects of FGSM and Jacobian-based Saliency (JSMA) attacks against brain segmentation and classification of skin lesions. Three pre-trained models were used (MobileNet, InceptionV3, and InceptionV4) for classification. Additionally, three DL segmentation models (Dense-Net, SegNet, and U-Net) were used for segmentation. The experiments revealed that the best robust models for classification and segmentation tasks were InceptionV3 and DenseNet, respectively. For classification, the authors showed that the strength of a model is proportional to its depth, but for segmentation, jump connections and dense blocks improve the efficiency of the model. The Structural Similarity (SSIM) ranged from 0.97 to 0.99, making the adverse samples imperceptible. Finlayson et al. [17] applied an attack against the ResNet50 model using PGD black and white box attacks on dermoscopy, chest radiography, and funduscopy images. In both attacks, the accuracy of the model was drastically reduced. Huq et al. [18] applied two white-box attacks (PGD, and FGSM) against two pre-trained (VGG16 and MobileNet) models for the classification of skin cancer. They experimented to classify an image into seven categories. After attacking, the accuracy decreased significantly. In mammographic images, the authors in [19] applied the FGSM attack. They used the (Digital Database for Screening Mammography) DDSM, which has two classifications: normal and malignant. While the SSIM fell below 0.2, the accuracy dropped by as much as 30Pal et al. [20] investigated the accuracy of COVID-19 classification based on computerized tomography (CT) and X-rays scans. The adversarial examples were generated using the FGSM attack and evaluated their impact on VGG-16 and InceptionV3 models. The results indicate the vulnerability of these models, with a reduction in accuracy of up to 63% and 90% for InceptionV3 and VGG-16, respectively. Transfer learning (TL) and self-supervised learning (SSL) introduced by authors in [21] were used to examine the robustness of the biological-image analysis. MRI was used for the cardiac segmentation dataset and Chest radiography was used for the pneumonia detection dataset. For transfer learning, a pre-trained ImageNet model was used. They tested PGD and FGSM attacks, as well as VGG11 and U-Net models. SSL outperforms TL because it learns stronger features, based on the findings. To improve performance on small, tagged data sets and adversary training, the authors advocate SSL in combination with adversary training as the default technique. Rahul et al. [22] applied the FGSM white-box attack and the one-pixel black box against the lung nodule classification model. In addition, three different architectures were used to train a custom model [23]. They reported a 28 to 36% decrease in accuracy after the FGSM attack. However, the model was significantly more robust in the black-box attack, with a reduction of only 2-3%. Kotia et al. [24] investigated the classification vulnerability of brain tumors to adversarial attacks. They applied noise-based attacks and FGSM white-box attacks. The most successful attack was FGSM, which reduced accuracy by 69%, while noise-based attacks reduced accuracy by 34 and 24%, respectively. The influence of adversarial attacks on retinal images was investigated by Shah et al. [25] to identify diabetic retinopathy, they analyzed image-based CNN and hybrid lesion-based algorithms for medical image analysis. To create adverse images, I-FGSM was used. The results demonstrate that the CNN models are relatively sensitive, while the lesion-based hybrid models are more robust, with 45 and 0.6% accuracy reductions, respectively. The relationship between the size of images in datasets, control parameters, and the efficacy of adversarial attacks was investigated by Kovalev et al. [26] White-box attack PGD was used to create the adversarial samples. The Inception V3 was used and two modalities for their experiments: chest X-ray and histology for eight distinct classification tasks. Histology images were found to be less vulnerable than X-rays. Furthermore, the greater the magnitude of the perturbation, the greater the success of the attack. Lastly, they demonstrated that the success rate of the attacks isn't influenced by the size of the training set. Allyn et al. [27] performed adversarial attacks on dermoscopic imaging. They tested the test set of data set HAM10000 with DenseNet201 after perturbing it. Overall, there was a 17% drop in accuracy. Li et al. [28]suggested a thick-to-thin deep 3D frame to address the problem of the NIH and JHMI datasets. ResDSN F2C is the model's name and is based on V-Net, U-Net, and VoxResNet. However, because FGSM and I-FGSM induce a significant drop in accuracy, this frame is subject to adversary attack (85.83%). To solve this problem, the authors recommended contradictory training for this model, which only reduced accuracy by 13.11 %. Shukla, et al. [29] investigated an adversarial attack against medical image segmentation models. Where the Loss function backpropagated to minimizes the error metrics. based the test performed on several popular models using various surrogate loss functions. Through a higher attack success rate, this attack outperforms other attacks against medical image segmentation.

D. Existing Attack Defense techniques

Ren et al. [30]used the adversarial defense in brain MRI segmentation to deal with small datasets. The subject of segmentation is extremely difficult due to the small datasets, especially in 3D MRI. However, the authors of this study showed that adding adversarial cases to the data can increase the robustness of the model. They used FGSM in an anisotropic CNN cascade to create adverse samples. Additionally, other studies used adversarial attacks during the training process to produce more robust models overall. Liu et al. [31] examined the effects of ad-versarial training on computed tomography nodules in the lungs. Three 3DResUNets were used in training, and data were obtained from the LUNA and NLST cohort. They employed the PGD attack to uncover the patterns leading to misclassification with great confidence and then used this data to train the network. The authors propose augmenting adversarial data to reduce the susceptibility of nodule detection inadequately represented nodule features and unexpected noise. Vatian et al. [32] conducted a notable study on adversarial examples, characterizing them as “natural” adversarial attacks. Their experimentation involved MRI and CT scan of the brain for lung cancer using a CNN model. The study revealed that in advanced medical imaging systems, noise acting as a ”natural” adversarial example could manifest itself. To counter these attacks, three defense methods were implemented. Adversarial training with JSMA and FGSM proved to be the most effective defense, outperforming the other two methods: layer activation function replacement with Bounded ReLU and data augmentation with Gaussian noise. Another study attempted to address the problem of limited-angle tomography, which can cause problems in CT reconstruction due to lack of data, resulting in image misinterpretation. To address this problem, Huang et al. [33] developed solid adversarial training. Because venom noise is prevalent in CT scans, they used it as a disturbance in the images for training. The trials were conducted using the U-Net model and low-dose CT grand challenge data. The results demonstrated that poisonous noise retraining is effective for limited-angle reconstruction, but insufficient for non-local adversaries. The resilience of three pre-trained deep diagnostic models was investigated by Xu et al. [34] Melanoma detection with IPMI2019-AttnMel, diabetic retinopathy detection with InceptionV3, and ChestX-ray classification with CheXNet. They used PGD and GAP (generative adversarial perturbations) assaults to test their theories. Both attacks significantly reduce model accuracy, with PGD attacking with 100% accuracy. To deal with attacks, the authors offered two defense techniques. The first is multi-perturbation adversarial training (MPAdvT), which involves training models with multiple perturbation intensities and iteration stages. In the adversarial training process, all samples are treated equally while according to Papernot et al. [35] the perturbation of misclassified examples is more important for model resilience, and in the realm of natural images, minimizing techniques are more crucial than maximization. The second defense method, misclassification-aware adversarial training (MAAdvT) is based on these observations. The authors added a misclassification aware regularization to adversarial loss. They use Kullback–Leibler (KL) divergence for the classifier to be stable against misclassified adversarial examples. The proposed defense methods present better results than standard adversarial training. All samples are treated equally in the contradictory training process, however, the authors state that disturbance in misclassified examples is more crucial for the robustness of the model and that minimization strategies are more relevant than maximizing in the field of the natural image. Many of the defensive approaches listed above involve adversary training as a defense strategy. Therefore, to address the aforementioned problems, this defense model is proposed, especially for cancer image reconstruction based on fine-tunning, in which the DCAE in the pre-training stage is integrated with CNN. There are two parts of our model: (1) Perturbated samples reconstruction with DCAE model. After that, the learning weights of the DCAE encoder are retrieved and used as initial weights for CNN initialization in the next stage. (2) Modeling with three different CNN architectures. Lastly, the proposed model is compared with the adversarial training technique.

3. THE PROPOSED MODEL RDMAA

In this section, the proposed technique for evaluating the robustness of CNN-based cancer diagnosing models will be introduced against two white-box adversarial attacks. It will include the process of generating the perpetuated adversarial image and constraint on the perpetuated image in section (3.1), the convolutional autoencoder (section 3.2), and also the CNN-based cancer classification models in (section 3.3). The proposed defense model against adversary attacks consists of two main parts. The first part of the RDMAA model is the DCAE image reconstruction pre-trained component and CNN is the second part as shown in Figure 2.

DCAE: This part is responsible for pre-training the convolutional layer using clean and noisy (adversarial FGSM and PGD) images to reconstruct the perpetuated images. It also transfers the learned weights to the CNN classification part. The DCAE architecture consists of three main computational layers which are the Convolutional Layer (2D Conv), the Max-Pooling layer, and the Upsampling Layer (UpSampling2D).

CNN: This part uses the weights-tunning technique based on DCAE pretrained weights of the encoding phase to be able to predict the adversarial images correctly with high classification accuracy. The architectures of CNN-based models are composed of five types of layers which are (convolutional 2D, max-pooling, Dense, Dropout, and Flatten layer).

A. *Adversarial Attacks Generation*

In this paper, two of the most frequent and effective adversarial attack techniques are used. The fast gradient sign method (FGSM), which was first introduced in 2014, and the projected gradient descent (PGD), which was first introduced in 2017, are tested on three different medical datasets. Three deep learning models for cancer image classification are attacked using these adversarial generation algorithms.

1) *Fast gradient sign method (FGSM) Attack*

Goodfellow et al. et al. [17],[36] proposed the one-step strategy to quickly create adversarial examples. Attackers modify the data based on gradient malfunction to maximize loss. With just a minor perturbation, the loss can be increased. Figure 3 shows an example of applying the FGSM attack on the three medical modalities that mislead the DL model. The adversarial perturbations are computed as the sign gradient of the loss with respect to the input image. The equation is as follows:

$$X'\_Adv = X + \epsilon * sign(\nabla x L(\theta, X, Y)) \tag{1}$$

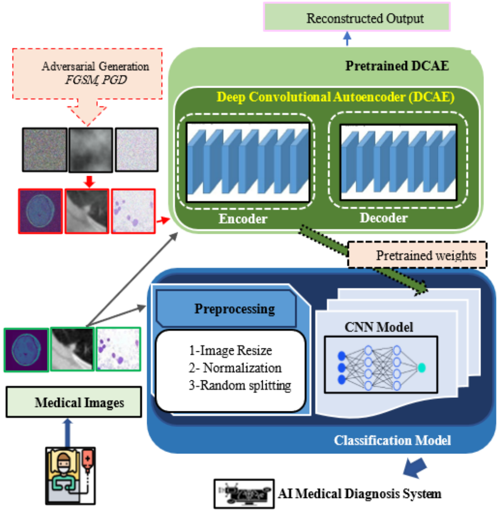


Figure 2. The Architecture of The Proposed Model RDMAA

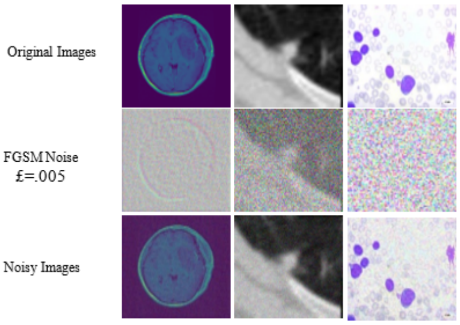


Figure 3. Example of FGSM Attack. 1st row displays the original modalities, 2nd row displays the FGSM patttern and the 3rd row displays the FGSM attack output

Where X'_Adv is the adversarial image, X is the original input image, Y is the label, L is the loss value, θ is the parameters, ∇x is the gradients and ϵ is the parameter controlling the maximum value of perturbation that is subtracted or added from every pixel in X. In this technique, the model is likely to be fooled by the sample produced in one step. FGSM generates adversarial samples faster than the other attacks as it only requires only one phase. As a result, FGSM meets the needs of experiments that require a large number of adversarial instances to be generated. Setting a high perturbation value will increase the probability that the adversary sample will be misclassified, but the resulting image will appear more distorted as a result. Therefore, the experiments were tested with two values of ϵ which are most common for medical images where ϵ =.002 and ϵ=.005. The parameter settings in Table 2 were used to create the adversarial images in Figures 3 and 4.

2) *Projected gradient descent (PGD) Attack*

Kurakin et al. [18] were the first to introduce PGD. It is a more iterative variant of the FGSM one-step approach.

$$X0 = X, X\_Adv^{t+1} = Clip_{x,\epsilon}(X^{t} + \alpha * sign(\nabla x * L(\theta, X^{t}, Y))) \tag{2}$$

Where t is the number of the iteration, ϵ is the perturbation degree, α is the step size, Clip is the function that clips its input so that it doesn't deviate from x by more than ϵ. This PGD attack searches the samples with the highest loss value. This type of adversarial is declared as "most adversarial". when the intensity of the disturbance (its norm) is limited, the sample points are more aggressive and more likely to mislead classifiers. Finding adverse instances is useful for identifying vulnerabilities in deep learning models. Three experimental parameters of the PGD attack were customized. The step size α= .002 and two perturbation values ϵ= .002 and .005 and t =3 iterations. To assess the robustness of AI-based cancer diagnostic models, a new dataset as adversarial input must be generated first. To do this, two adversarial attack techniques are applied, namely the fast gradient method (FGSM) and the projected gradient descent (PGD) method. For each original input image, the FGSM and PGD techniques generate perturbations by modifying the gradient of previously trained CNN diagnostic models. All details about those attacks are explained in the following sections. FGSM and PGD constitute solid foundations and are well known in the sector. FGSM is a rapid one-step method, while PGD is an iterative approach. By testing both, you can evaluate how well the defense model performs against a variety of attack strategies. FGSM represents a scenario where attackers may have limited time to generate adversarial examples, while PGD represents a scenario where attackers may have more computational resources to design attacks. Evaluating against a wide range of adversarial attacks can require a large amount of computation. Using a smaller set of representative attacks, such as FGSM and PGD, makes experimentation more manageable while providing valuable insights.

B. *Convolutional Autoencoders (CAE)*

1) *Autoencoder (AE)*

is a feed-forward neural network that seeks to recreate the input into the output under specific constraints. By first encoding and then decoding the inputs, the AEs execute unsupervised pre-training. The interconnections between layers are fully connected. The units of the previous layer are linked to the units of the next layer. The input and output layers are the same sizes as the image. The autoencoder is managed to learn compressed representation without loss of information by making the output target the same as the original image [37] .

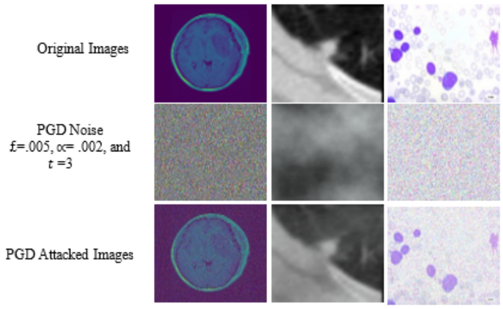


Figure 4. Example of FGSM Attack. 1st row displayes the original modalities, 2nd row displays the FGSM patttern and the 3rd row displays the FGSM attack output

2) *Deep Convolutional autoencoder (DCAE)*

extends the basic structure of the AEs since it changes the fully connected layers to convolution (convolutional encoding and decoding) layers. DCAEs are better suited for image processing than classical autoencoders because they use the full power of convolutional neural networks to exploit the structure of the image. In DCAEs, the weights are shared between all input locations which helps to save the local spatiality. DCAEs merge the unsupervised pre-training of autoencoders with the advantages of convolutional filtering in CNN. Rather than the fully connected layers, the encoder incorporates convolutional layers, and the decoder has deconvolutional layers, in contrast to the structure of autoencoders. Deconvolutional filters can be transferred copies of convolutional filters, or they can be trained from scratch. Each deconvolutional layer must also be preceded by a layer that falls apart [38]. The design of a DCAE is divided into two parts: an encoding phase to represent features or to give a compressed version of the input, and a decoding section to reconstruct the input from the compressed form. The convolution and maximum pooling layers are used in the encoding phase, while the deconvolution and up-sampling layers are used in the decoding section as shown in Figure 5. The shared weights are a benefit of deep learning techniques. Autoencoders also can quickly reconstruct noisy images. As a result, the proposed defense model combines the benefits of DCAE and CNN for pre-training. Then the classification models are initialized based on fine-tuning of parameters. The encoding part of the DCAE consists of three layers of convolution 2D and three layers of Max_pooling as shown in Figure 6. On the other hand, the decoding part of DCAE has four convolution 2D layers and three up_sampling layers. Algorithm 1 describes the main steps to using the DCAE model to train the network on the medical images. Three different DCAEs are trained to separately reconstruct the three used medical modalities (Lung nodule, leukemia, brain tumor) simply with the same architecture. This means that the proposed

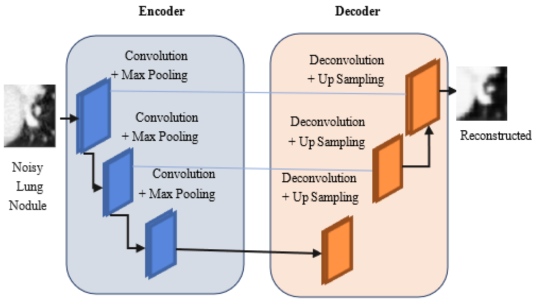

Encoder
Decoder
Convolution + Max Pooling
Convolution + Max Pooling
Convolution + Max Pooling
Deconvolution + Up Sampling
Deconvolution + Up Sampling
Deconvolution + Up Sampling
Noisy Lung Nodule
Reconstructed

Figure 5. General Structure of The Deep Convolutional Autoencoder

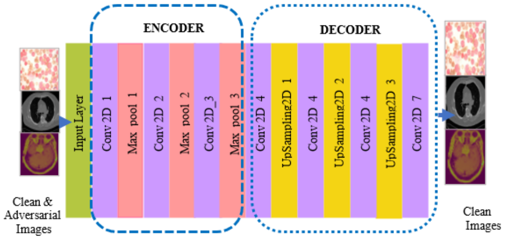


Figure 6. The Proposed DCAE-ADVs Framework

DCAE model can be used with different types of medical data and achieve a high reconstruction rate. Different filter sizes were tested on the convolutional layer and finally set to 64 with a kernel size of (3 * 3). This framework produced acceptable reconstruction accuracy and required less time for training. In the proposed DCAE model, three different size values are used for each type of medical data as the input image size. Generally, the input size is equal to (height * width * channels). For brain tumors and leukemia, the size of input images was (64*64*3) and for lung images it was (52*52*1). The input size was initially reduced three times at the encoder by a scale factor of two before being compressed. The decoder then up-scaled those potential feature maps three times to the original size but included 16 channels. The last 16-channels would be recreated to the same size as the three-channel input by adding three filters to a convolutional layer on top of the decoder.

Algorithm 1 The proposed DCAE-ADVs reconstruction model

```
Input: A clean and noisy medical image D.
Output: Encode (convolutional and max_pooling) weights $w_x$,$w_y$.
Initializing the $w_x$,$w_y$ randomly.
Encode the medical images.
for ¡ITIR= 1: N do
    Estimate the 2D convolution weights $C_i$.
    Estimate max_pooling weights $M_k$.
    Estimate up_sampling $U_k$.
    Estimate 2D deconvolution $Cd_i j$.
    Reconstruct the noisy image with the 2D deconvolution output.
    Update the $w_x, w_y$.
end for
return $w_x$,$w_y$.
```

C. Convolutional Neural Network

The proposed scheme first applies three steps of image preprocessing. The first is to resize the images so that they are the same size. The brain and leukemia images were resized to (64 * 64) while the lung images were resized to (52 * 52). Then, Medical images are normalized. To train and test the CNN model, each dataset is randomly divided into three parts: training, validation, and testing as shown in Table 1. The first six layers of the three CNN models (convolutional and max_pooling) layers were identical to the encoder part in the DCAE. Following these layers, the output features were flattened into a one-dimensional vector for the three classification models. This was then followed by two dense layers in the brain tumor Classification model. Whereas, the lung nodule model is followed by three dense layers. On the other hand, the flatten layer was followed by three dense, dropout, and dense layers in the leukemia model as shown in Figure 7. In the brain tumor classification model, the final dense layer had three neurons for multiclass classification (meningioma, glioma, and pituitary) tumors. whereas, it had four neurons in leukemia to classify (Benign, Early, Pre, and Pro) and only one neuron in the lung nodule model because it performs a binary classification (Benign and Nodule). In the proposed adversarial defense model, the weights estimated by the encoder from the pre-trained autoencoder (DCAE) had been reused as the initial weights for the identical part in the CNN model. The other layers had the weights randomly.

4. EXPERIMENTAL RESULTS

This section begins by discussing the experimental environment. Also, the evaluation metrics will be discussed to estimate the robustness of the proposed models. Then, all the details about the cancer datasets will be discussed. In addition, the performance of the three deep cancer diagnostic architectures will be estimated under two adversarial attacks. Finally, a comparative analysis between the proposed model with another famous defense technique will be shown.

A. Experimental Environment

All the experiments in this paper were executed on the COLAB Cloud Platform [39]. The processor runtime was CPU since the size of data wasn't very large. The software used in the experiments was the Jupyter notebook. All the experiments have been carried out using Python 3 with many data science libraries such as Matplotlib, Scikit-Learn, Pandas, NumPy, CV2, Keras, and TensorFlow.

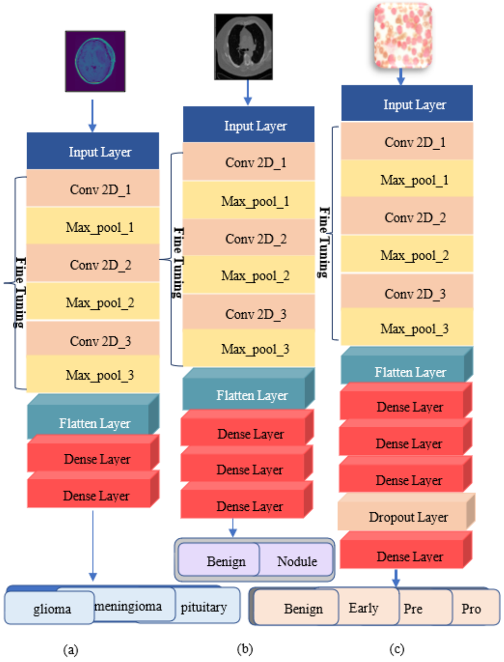


Figure 7. Proposed DL-based Cancer Diagnosing Models: a) shows the CNN of MRI classification, b) CNN-based X-ray classification and c) Leukemia classification framework

B. Evaluation Metrics

After the generation of the adversarial images, they are fed into the trained CNN classification models. To obtain a more complete assessment of the robustness of deep cancer diagnostic models, binary and multi-class models are used in this study. There are five evaluation metrics which are the accuracy, Precision *Pre*, Recall *Rec*, confusion-matrix, and F1_score ratio of the attack impact and the defense impact to assess the robustness of the models. They are described below [37]:

$$Precision = \frac{TruePositive(TP)}{TruePositive(TP) + FalsePositive(FP)} \quad (3)$$

$$Recall = \frac{TruePositive(TP)}{TruePositive(TP) + FalseNegative(FN)} \quad (4)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

$$F1\_Score = \frac{2 * (2 * TP)}{(2 * TP + FP + FN} \quad (6)$$

A true positive is when the model accurately predicts the positive class, while a true negative is when the model accurately predicts the negative class. Conversely, a false positive happens when the model incorrectly predicts the positive class, and a false negative occurs when the model inaccurately predicts the negative class.

C. Datasets Characteristics

Three publicly available cancer images datasets are used in this paper, including (1) LUNA16 [40] with X-ray images for lung nodule classification, (2) the Kaggle (CE-MRI) [41] dataset with MRI high contrast-colored images, and (3) the Acute Lymphoblastic Leukemia (ALL) [42] dataset with microscopic images. To verify that the proposed adversary defense model described in the previous section is generally appropriate for the AI-based cancer classification domain. The proposed defense model was tested using three DNN-based medical image classifications: (1) multi-classification of Acute leukemia from fundoscopy as (benign, pro, pre, early) (2) Binary classifying lung nodule from chest x-rays as (benign and nodule); and (3) multi-classification of the brain tumor as (glioma, meningioma, pituitary) from MRI as shown in Figure 6.

Acute Lymphoblastic Leukemia (ALL) [42]: This dataset is used to classify and identify ALL blasts in the most common cancer type of childhood and is publicly available at the Kaggle website. This dataset contains 3256 PBS images from 89 patients collected in the bone marrow laboratory of Taleqani Hospital (Tehran, Iran). The dataset consists of two main categories benign with three subtypes of malignant lymphoblasts leukemia (Benign, Early, Pre, Pro) lymphoblasts. DNNs should recognize the blasts of leukemia based on the presence of blood cell abnormalities.

LUNA16 Lung Nodule [40]: The LUNA16 is a subset of LIDC-IDRI dataset. For lung nodule classification, a subset of segmented lung nodules from LUNA16 X-ray dataset was used. This dataset contains X-ray images. The dataset contains 8106 images as nodules or non- nodules. Only the annotations categorized as nodules ≥ 3 mm as the other annotations (nodules ≤ 3 mm and non-nodules).

CE-MRI Brain Tumor [41]: The Kaggle CE-MRI brain tumor dataset for the tumor classification task is used. The data set contains MRI images and can be used as a training set for academic machine learning. CE-MRI was acquired between 2005 to 2010 from Nanfang Hospital, China. CE-MRI dataset includes 3,064 T1-weighted contrast-weighted images from 233 patients who had three different types of brain tumors: glioma, meningioma, and pituitary tumor as shown in Figure 8. A total of 64% of the X-ray image collection has been designated for training, 16% for validation, and 20% for testing. The collection of microscopic images has been divided into two parts: 71% for testing and 20% for training and validation. For training, validation, and subsequent testing purposes, the MRI data set was divided into three segments: 80%, 10%, and 10%. The number of samples used for training and testing are listed in Table 1.

D. Results and Discussion

In this part, the verification that the proposed RMDAA fine-tuning defense model is beneficial to improve the robustness of the AI-based cancer models. Compared to the non-medical DL models, the medical DL models appeared to be much more vulnerable to adversarial attacks. Three

TABLE I. Details Of The Three Cancer-Based Medical Datasets Splitting

Dataset	Classes	Train	Validate	Test	Test
				Original	AdversTest
X-ray	2	5187	1297	1622	1622
MRI	3	2481	276	307	307
Microscopic	4	2343	261	652	652

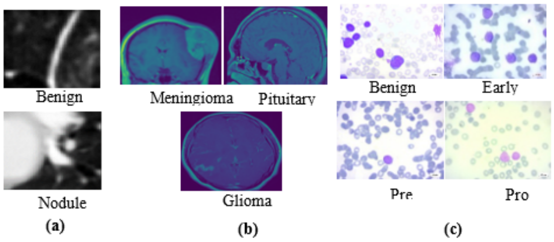


Figure 8. Samples of Each Class from The Three Cancer-Based Medical Datasets: a) X-ray lung images, (b) MRI Brain Tumor and (c) Microscopic Leukemia.

medical diagnostic models were trained for each medical modality. All three models have been fully trained to obtain an accurate classification rate for each modality. Then, two white-box attacks were performed separately against the well-trained CNNs. The generation of the adversarial sample requires certain constraints to the perturbation value to obtain adversarial samples that have an invisible difference compared to the original. When comparing the minimum perturbation size required for most attacks to be successful in medical imaging DL models with non-medical imaging DL models, a smaller ϵ size was given for medical attacks to be strong. The parameters of the two attacks are determined using the hyperparameter search method to find values that are appropriate for the experiments depending on the desired attack strength. We started by experimenting with different parameter settings. For example, we started with a range of perturbation values ϵ values since ϵ typically range from extremely small (e.g., 0.01) for minor perturbations to larger values for more powerful attacks. Then, different combinations of parameters for attacks are tested. Analyze how well our model performed against these attacks according to common evaluation measures such as Attack success rate, robustness, etc. Finally, the parameter values that result in the appropriate level of adversarial success rate are selected. The following are the specific set-tings for the two attack types mentioned above: two perturbation values ϵ of FGSM are tested .002 and 0.005. The PGD attack has a step size of .002 and two perturbation values .002 and .005 and three iterations. The generated adversarial images are equal to the number of samples in the test dataset. Table 2 shows the parameter settings of the FGSM and PGD algorithms.

Before using adversary attacks, the classification accuracies reached 95% for the MRI dataset, 90% for X-ray, and 86% for the microscopic dataset. Then, apply attacks on the three models separately using the white-box adversarial examples that have effectively attacked the trained CNN. The impact of these attacks across the three datasets and the success rate of the proposed defense models are reported in Tables 3,4, and 5. As shown, the performance of the models under attacks had a greater absolute decrease for X-ray and leukemia and a moderate decrease for brain MRI under FGSM attacks. Lung X-ray: The success rate of the two attacks is higher in the X-ray datasets than in the brain MRI and leukemia datasets. However, the attack parameters used in the three datasets have the same design. This may be the result of the difference in biological texture between the cancer modalities. As shown in Figure 9, the success rates (the accuracy dropped up to) are more than 55% with FGSM and 50% with PGD attacks. On the other hand, this dataset achieves a high defense rate compared to the MRI dataset as shown in Table 4. The success rates are reduced to 3% and 8% under the FGSM and PGD respectively with other metrics as shown in Tables 3 and 6. Acute leukemia Microscopic: As seen in Tables 4 and 7, however, acute leukemia has a highly successful attack rate, it has the strongest defensive impact against FGM attacks. Table 4 shows that the attack success rates reached 50% for FGSM and 20% for PGD. While the defense model reduced the success rate to 89% under FGSM and 87% under PGD. This is because the proposed defensive model is based on the DCAE image reconstruction. When comparing the defense effects of the proposed defense models, it becomes clear that fine-tuning can significantly enhance the defense power of the model. Brain MRI: In contrast, the success rate of the two attacks in this dataset is lower than in the lung X-ray and leukemia datasets. Table 5 shows that the biological structure of the high contrast brain tumor images is more resistant to adversarial attacks. The success rate under FGSM is 25% and 18% under PGD. Tables 5 and 8 show that the proposed technique has a better defense impact against the two attacks. Of course, the success rate of the attack is related to the accuracy rate of the target model where the success rate decreased to 3% under both attacks. The accuracy of each target model will be displayed to confirm that it performs well in the classification of the medical diagnosis. As shown in tables 6,7 and 8, the proposed defense model achieves a very robust performance toward these attacks. Tables 6,7 and 8 show that the test accuracy of the PGD, and FGSM attacks dropped to a range

TABLE II. The parameters of the two attacks

Attack Method	Perturbation Values ϵ	Step Size α	Number of Iterations
FGSM	.002,.005	-	-
PGD	.002,.005	.002	3

between 35% and 77% accuracy, which is much lower than the clean set. This indicates that the attack is successful about 50% of the time. This result demonstrates that it is quite simple to undermine the performance of a neural network. On the other hand, the proposed model achieved a higher accuracy rate under the two attacks.

Adversarial training is a data augmentation defense technique that involves adding some adversarial samples to the original dataset. It's a powerful technique for developing a robust model resistant to per-instance attacks, in which adversarial instances are injected into the dataset during training. The original training dataset (I) and the adversarial datasets (S) are included in the new training dataset N, which serves as the training input of the DL models. Assume that the adversarial dataset (S) and the original dataset (I) have equal sizes. A simpler approach is to always train the DL models using the adversary dataset that ensures the highest accuracy of the model under attack for a specific range of adversary strengths used by an adversary attack. Multiple networks with multiple inputs for each type of DL model are trained. For each medical dataset modality, one network is trained using only the clean data, The other two networks are trained using adversarial samples augmented to the clean samples. For example, FGSM samples are added to the clean training set to adversarial train a network with FGSM. The training set now contains twice the number of samples as it did at the beginning. The same process is done with PGD adversarial training. Tables 9,10 and 11 show the accuracy of the test after the models have been trained using FGSM, PGD samples, and clean pairs. From Figure 10, It is observed that the network correctly categorizes most of the adversarial samples, reaching an accuracy similar to the clean samples. It shows a great enhancement from 40% to 50% for X-ray images, while, less enhancement was obtained for Microscopic and MRI modalities ranging from 10% to 20%. This indicates that the distribution of the original images and adversarial images are different from the MRI and Microscopic images. Tables 9,10 and 11 compare the effectiveness of our proposed defense model with the adversarial training standard defense techniques against the two white-box attacks. The robustness of the three cancer-based DNN classification models was tested against FGSM and PGD attacks using the two defense techniques. The results of classification for the three attacked architectures were greatly im-proved after employing the proposed fine-tuned model, as you can see in the tables. For example, the defense accuracy for a binary-class lung nodule classification task is 83%, and 88% for PGD and FGSM respectively, while the standard adversary training is 80%, and 86%. The same conclusion can be obtained with multi-class MRI classifiers for 93%, and 93%, while adversarial training achieved 78% and 86% respectively. one can witness that the proposed model is significantly better than the adversarial training technique.

Also, defense is judged to be efficient if it can resist a broad range of attacks and different parameters of attacks. Therefore, in Figure 8, the accuracy of the classifiers in the three medical datasets is demonstrated with and without defenses against FGSM and PGD white-box attacks, with various perturbations ϵ sizes, to analyze the effect of the parameters of the two attacks in the proposed fine-tuned CNN models. Figure 9 shows how the classifier's accuracy gradually drops as the value of perturbation ϵ increases with the three medical datasets. As a result, attacks with higher ϵ are more successful. Also, the accuracy of the proposed defense model under the same white-box attacks with the same perturbations values is shown in the same graph. The classifiers are shown to be very secure and robust against attacks with the proposed defense model, even at high ϵ values. According to the university of our model, the main goal of our defense strategy is to make deep neural medical models more resilient to adversarial attacks. Several factors determine whether a defense technique created for a DL model can be applied to other image classification models. These factors include the similarity of the models, the possibility of transferability, the vulnerabilities inherent to each model, the ability to respond to generic attack scenarios, and more. It is essential to understand that the proposed technique may not work in all situations. However, it is important to note that this technique had positive results when used with three different medicinal modalities, each of which is distinguished by certain characteristics. These results differ from those of other comparable models, which were only evaluated using one type of medical evaluation. Given the significant need for medical applications, our study mainly focuses on advocating medical DNN models. However, we realize that future research may be necessary to determine whether our defense is universally applicable to other domains and model types. We will consider upcoming research that examines its versatility for many uses. The results demonstrate that additional parameters affect the effectiveness of adversarial training techniques such as the algorithm used to create the adversarial samples. This means that if the network model is to be resistant to all potential attacks, the number of adversary instances required for adversary training can be increased significantly to account for all possible configurations. In general, it is quite challenging to train a resilient network that guarantees a certain level of robustness against all kinds of adversary cases. The RDMAA defense model is extremely robust

TABLE III. The Impact of Attacks On DL-Based X-Ray Datasets Classification and The Resistance of The Proposed Defense Model

	LUNG Nodule Classifier Performance			
	No Defense	RDMAA	Enhancement	Difference
No Attack	90%	91%		
PGD Attack ϵ=.005,α=.02	40%	83%	43%	8%
FGSM Attack ϵ=.005,	35%	88%	53%	3%

TABLE IV. The Impact of Attacks On DL-Based Microscopic Datasets Classification And The Re-sistance Of The Proposed Defense Model

	Acute Leukemia Classifier Performance			
	No Defense	RDMAA	Enhancement	Difference
No Attack	86%	90%		
PGD Attack ϵ=.005,α=.02	36%	87%	40%	3%
FGSM Attack ϵ=.005,	66%	89%	23%	1%

TABLE V. The impact of attacks on DL-based MRI datasets classification and the resistance of the proposed defense model

	Brain Tumor Classifier Performance			
	No Defense	RDMAA	Enhancement	Difference
No Attack	95%	96%		
PGD Attack ϵ=.005,α=.02	77%	93%	16%	3%
FGSM Attack ϵ=.005,	70%	93%	23%	3%

TABLE VI. The Classification Performance Evaluation of the 2-Classes Lung Nodule DNN Model with No-Defense and Defense Technique

	No Defense Model				RDMAA			
	ACC	P	R	F1	ACC	P	R	F1
No Attack	90%	83%	86%	84%	91%	85%	82%	84%
PGD Attack ϵ=.005,α=.02	40%	50%	49%	38%	83%	72%	79%	74%
FGSM Attack ϵ=.005	35%	38%	29%	29%	88%	80%	77%	78%

TABLE VII. The Classification Performance Evaluation of The 4-Classes Acute Leukemia DNN Model With No-Defense And Defense Technique.

	No Defense Model				RDMAA			
	ACC	P	R	F1	ACC	P	R	F1
No Attack	86%	85%	84%	84%	90%	90%	88%	89%
PGD Attack ϵ=.005,α=.02	47%	50%	47%	47%	87%	86%	85%	85%
FGSM Attack ϵ=.005,	66%	64%	65%	66%	89%	88%	86%	87%

TABLE VIII. The Classification Performance Evaluation of The 3-Classes Brain Tumor DNN Model With No-Defense And Defense Technique.

	No Defense Model				RDMAA			
	ACC	P	R	F1	ACC	P	R	F1
No Attack	95%	95%	93%	93%	96%	95%	95%	95%
PGD Attack ϵ=.005,α=.02	77%	77%	79%	77%	93%	93%	93%	93%
FGSM Attack ϵ=.005,	70%	68%	65%	66%	93%	92%	93%	92%

TABLE IX. The Results of Two Defense Techniques over X-ray Dataset Under Two White-box Attacks

Lung Nodule Accuracy Comparison		
	Advrs_Training	RDMAA
PGD Attack ϵ=.005,α=.02	80%	83%
FGSM Attack ϵ=.005,	86%	88%

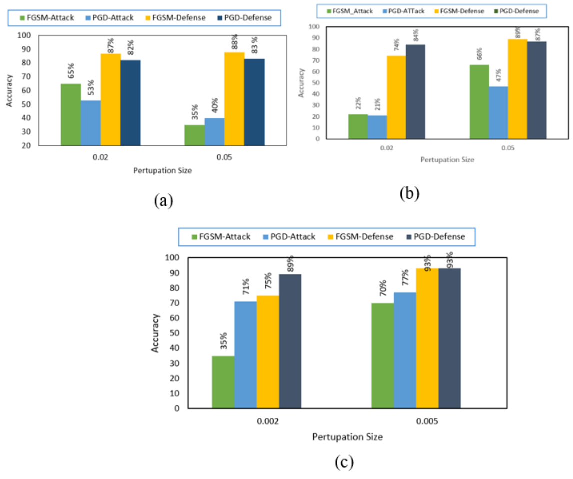
TABLE X. The Results of Two Defense Techniques over Microscopic Dataset Under Two White-box Attacks

Acute Leukemia Accuracy Comparison		
	Advrs_Training	RDMAA
PGD Attack ϵ=.005,α=.02	68%	87%
FGSM Attack ϵ=.005,	78%	89%

TABLE XI. The Results of Two Defense Techniques over MRI Dataset Under Two White-box Attacks

Brain Tumor Accuracy Comparison		
	Advrs_Training	RDMAA
PGD Attack ϵ=.005,α=.02	78%	93%
FGSM Attack ϵ=.005,	86%	93%

against FGSM and PGD attacks compared to the adversary training technique. Since the two white box at-tacks use geometric perturbations to target the most crucial areas in the benign samples, the proposed RDMAA model reconstructs the antagonistic samples to their original clean shape. Given the relatively limited research contributions in the field of defending medical adversarial attacks compared to natural images, and the absence of publicly available models tailored for our three datasets, we opted for two approaches. First, we applied a well-known public defense technique, called adversarial training (ADT), as a foundational step. This allowed us to make an initial comparison between our proposed model and this widely used technique. Our work not only addresses the current research gap but also establishes a baseline for evaluating the performance and potential advancements in medical adversarial defense. We plan to explore this avenue in future work to provide a more comprehensive assessment of the proposed defense method. As indicated in Tables 3-8, the results of the baseline model, without any defense mechanism, achieved accuracy rates of 91%, 86%, and 95% for the three medical modalities: X-Ray, Microscopic, and MRI datasets, respectively. However, when subjected to the PGD attack, the accuracy dropped significantly to 40%, 36%, and 77% sequentially. Similarly, when the FGSM attack was applied, it resulted in a substantial decrease in accuracy to 35%, 66%, and 70%, respectively. These results clearly demonstrate the vulnerability of the baseline model to adversarial attacks, which substantially degrade its performance on these medical datasets. The proposed defense mechanism aims to mitigate these effects and enhance the model's robustness. The tables also highlight the consistency of the defense model across the three medical modalities, even when subjected to different attack parameters. Notably, the model's accuracy without any defense increased by 1%, 4%, and 1% sequentially for the three modalities: X-Ray, Microscopic, and MRI datasets, respectively. The defense mechanism demonstrated enhanced performance against PGD attacks, resulting in accuracy improvements of 43%, 40%, and 16%. Similarly, the defense model effectively countered FGSM attacks, resulting in a 53%, 23%, and 23% increase in accuracy. These results showcase the effectiveness of the RDMAA defense model compared to the adversarial training (ADT) technique as introduced in tables 9-11. The proposed model not only enhances accuracy but also demonstrates its consistency across different medical modalities. This consistency is a valuable contribution for researchers, indicating that the defense model is a robust solution applicable to various medical domains. Also, the confusion matrices are plotted to evaluate the change in prediction samples as a result of comparing the effect of the adversarial training technique with the proposed defense model. The true and predicted classes are represented by the rows and columns in the arrays, respectively as shown in Figure 10. After attacking the systems, in the binary classification, most lung X-ray images tended to be misclassified as nodules and vice versa. Furthermore, the majority of MRI images of the brain were identified as glioma for multiclass classification. The attack affected the algorithm to mispredict the true labels of the leukemia images, and most of the images were classified as early blasts. As a key to success, the suggested model reduces the impact of misclassification. The confusion matrices for the three models showed that the defense technique is robust against attacks, as most of the images were correctly classified into the original classes after the adversary attacks. However, the effect of



(a)

(b)

(c)

Figure 9. Accuracy of FGSM and PGD Attacks Impacts and Defense Technique on the three medical datasets: (a) Lung X-ray , (b) Leukemia Microscopic and (c) Brain MRI

the adversarial training technique was partially limited for the three data sets.

Data Availability Statements The datasets analyzed during the current study are available in the [Brain Tumor Image Dataset] repository [41]. These datasets were gathered from the publicly available sources listed below: [https://www.kaggle.com/datasets/denizkavi1/brain-tumor?select=3], the [Acute Lymphoblastic Leukemia (ALL) image dataset] repository [42] available public on: [https://www.kaggle.com/datasets/mehradaria/leukemia], and available in the [luna16 dataset] repository [40]. Available online on: [https://luna16.grand-challenge.org/Data/].

Declarations Statement

Funding Sources: This research doesn't receive support from any dedicated funding source.
Conflict of Interests: We affirm that there are no conflicts of interest to declare.

5. Conclusions and Future Work

Recently, the widespread use of deep learning (DL) frameworks especially the deep convolutional neural network (CNN) has received a lot of interest in a variety of industries. The black-box nature of these models makes them vulnerable to adversarial attacks. In automated medical diagnosis, although the adversarial attacks are high risk, it is expected that DL-based medical diagnosis will be widely used safely and accurately. As a result, adversarial defense techniques relevant to the medical DNN model must be proposed. In this paper, the risks of the white-box attacks against the DL-based medical diagnosis framework have been investigated. The following is a list of the main contributions to this paper: (1) A new defense model based on the RDMAA against white-box adversary attacks for DL-based cancer diagnosis systems where the deep CNNs are fine-tuned using the weights of the pretrained DCAE-ADVs image reconstruction model. (2) Many experiments have been conducted with two types of white-box adversary attacks on three different cancer-based medical image modalities datasets (X-ray, MRI, and Microscopic). (3) The proposed defense model is compared with an-other famous defense technique in three different experiments. The results proved that the proposed defense model can remarkably outperform the other existing defense method for the medical field. Tunning the CNN-based cancer diagnosis models with the weights of the pre-trained DCAE-ADVs image reconstruction encoder significantly increased the robustness of the models. Further-more, this defense model outperforms the adversary training defense method in terms of improving the robustness of the DL models. As a result, the proposed model can be used in a variety of medical modalities to enhance the robustness of the medical diagnosis models. One of the potential future works is to expand the scope of RDMAA in the time-series data. That implies acquire relevant datasets containing physiological measurements, ECG recordings, or other appropriate time-series health data for-mats. developing existing attack methods specifically designed to exploit the temporal patterns and vulnerabilities

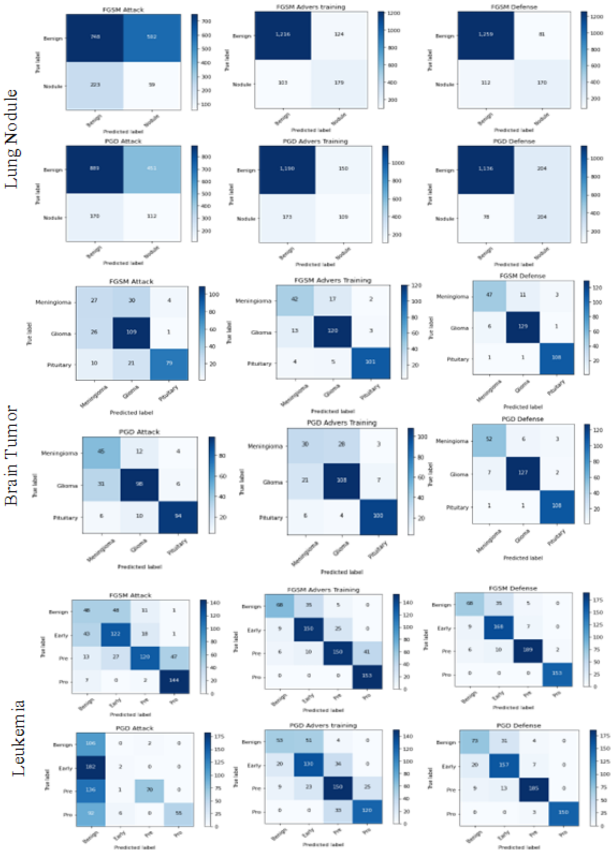


Figure 10. Confusion Matrices For Proposed Three Medical Models Under Attacks And Defenses

within time-series health data. evaluating the performance of RDMAA against these time-series attacks and compare it to other state-of-the-art defense techniques on the same datasets. New attack methods will be studied to demonstrate the effectiveness of the proposed defense methods.

References

[1] W.-T. Chu and W.-W. Li, "Manga face detection based on deep neural networks fusing global and local information," *Pattern Recognition*, vol. 86, pp. 62–72, 2019.

[2] Q. Wang, H. Fan, G. Sun, Y. Cong, and Y. Tang, "Laplacian pyramid adversarial network for face completion," *Pattern Recognition*, vol. 88, pp. 493–505, 2019.

[3] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[4] T. Xia, A. Chartsias, and S. A. Tsaftaris, "Pseudo-healthy synthesis with pathology disentanglement and adversarial learning," *Medical Image Analysis*, vol. 64, p. 101719, 2020.

[5] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," 2018.

[6] D. Freeze, "2019/2020 cybersecurity almanac: 100 facts, figures, predictions and statistics," Nov 2020. [Online]. Available: https://cybersecurityventures.com/cybersecurity-almanac-2019/

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *CoRR*, vol. abs/1412.6572, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:6706414

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[9] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *ArXiv*, vol. abs/1607.02533, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:1257772

[10] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *CoRR*, vol. abs/1312.6199, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:604334

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

[12] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2574–2582.

[13] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

[14] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu, "Improving adversarial robustness requires revisiting misclassified examples," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020.* OpenReview.net, 2020.

[15] S. A. Fezza, Y. Bakhti, W. Hamidouche, and O. Déforges, "Perceptual evaluation of adversarial attacks for cnn-based image classification," *CoRR*, vol. abs/1906.00204, 2019. [Online]. Available: http://arxiv.org/abs/1906.00204

[16] M. Paschali, S. Conjeti, F. Navarro, and N. Navab, "Generalizability vs. robustness: Adversarial examples for medical imaging," *CoRR*, vol. abs/1804.00504, 2018. [Online]. Available: http://arxiv.org/abs/1804.00504

[17] S. G. Finlayson, I. S. Kohane, and A. L. Beam, "Adversarial attacks against medical deep learning systems," *CoRR*, vol. abs/1804.05296, 2018. [Online]. Available: http://arxiv.org/abs/1804.05296

[18] A. Huq and M. T. Pervin, "Analysis of adversarial attacks on skin cancer recognition," *2020 International Conference on Data Science and Its Applications (ICoDSA)*, pp. 1–4, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:222219829

[19] I. Yilmaz, M. Baza, R. Amer, A. Rasheed, F. Amsaad, and R. Morsi, "On the assessment of robustness of telemedicine applications against adversarial machine learning attacks," in *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, H. Fujita, A. Selamat, J. C.-W. Lin, and M. Ali, Eds. Cham: Springer International Publishing, 2021, pp. 519–529.

[20] B. Pal, D. Gupta, M. Rashed-Al-Mahfuz, S. A. Alyami, and M. A. Moni, "Vulnerability in deep transfer learning models to adversarial fast gradient sign attack for covid-19 prediction from chest radiography images," *Applied Sciences*, vol. 11, no. 9, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/9/4233

[21] R. Paul, M. Schabath, R. Gillies, L. Hall, and D. Goldgof, "Mitigating adversarial attacks on medical image understanding systems," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020, pp. 1517–1521.

[22] D. Anand, D. Tank, H. Tibrewal, and A. Sethi, "Self-supervision vs. transfer learning: Robust biomedical image analysis against adversarial attacks," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, 2020, pp. 1159–1163.

[23] R. Paul, S. Hawkins, M. B. Schabath, R. J. Gillies, L. O. Hall, and D. B. Goldgof, "Predicting malignant nodules by fusing deep features with classical radiomics features," *Journal of Medical Imaging*, vol. 5, no. 1, p. 011021, 2018. [Online]. Available: https://doi.org/10.1117/1.JMI.5.1.011021

[24] J. Kotia, A. Kotwal, and R. Bharti, "Risk susceptibility of brain tumor classification to adversarial attacks," in *International Conference on Man-Machine Interactions*, 2019.

[25] A. Shah, S. Lynch, M. Niemeijer, R. Amelon, W. Clarida, J. Folk, S. Russell, X. Wu, and M. D. Abràmoff, "Susceptibility to misdiagnosis of adversarial images by deep learning based retinal image analysis algorithms," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, 2018, pp. 1454–1457.

[26] V. Kovalev and D. Voynov, "Influence of control parameters and the size of biomedical image datasets on the success of adversarial attacks," *CoRR*, vol. abs/1904.06964, 2019. [Online]. Available: http://arxiv.org/abs/1904.06964

[27] J. Allyn, N. Allou, V. Charles, A. Renou, and C. Ferdynus, "Adversarial attack on deep learning-based dermatoscopic image recognition systems: Risk of misdiagnosis due to undetectable image perturbations," *Medicine*, vol. 99, p. e23568, 12 2020.

[28] Y. Li, Z. Zhu, Y. Zhou, Y. Xia, W. Shen, E. K. Fishman, and A. L. Yuille, "Volumetric medical image segmentation: A 3d deep coarse-to-fine framework and its adversarial examples," in *Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics*, 2020.

[29] S. Shukla, A. K. Gupta, and P. Gupta, "Exploring the feasibility of adversarial attacks on medical image segmentation," *Multimedia Tools Appl.*, vol. 83, no. 4, p. 11745–11768, jun 2023. [Online]. Available: https://doi.org/10.1007/s11042-023-15575-8

[30] X. Ren, L. Zhang, D. Wei, D. Shen, and Q. Wang, "Brain mr image segmentation in small dataset with adversarial defense and task reorganization," in *Machine Learning in Medical Imaging: 10th International Workshop, MLMI 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings*. Berlin, Heidelberg: Springer-Verlag, 2019, p. 1–8.

[31] S. Liu, A. A. A. Setio, F. C. Ghesu, E. Gibson, S. Grbic, B. Georgescu, and D. Comaniciu, "No surprises: Training robust lung nodule detection for low-dose ct scans by augmenting with adversarial attacks," *IEEE Transactions on Medical Imaging*, vol. 40, no. 1, pp. 335–345, 2021.

[32] A. Vatian, N. Gusarova, N. Dobrenko, S. Dudorov, N. Nigmatullin, A. Shalyto, and A. Lobantsev, "Impact of adversarial examples on the efficiency of interpretation and use of information from high-tech medical images," in *2019 24th Conference of Open Innovations Association (FRUCT)*, 2019, pp. 472–478.

[33] Y. Huang, T. Würfl, K. Breininger, L. Liu, G. Lauritsch, and A. Maier, *Some Investigations on Robustness of Deep Learning in Limited Angle Tomography*, 09 2018, pp. 145–153.

[34] M. Xu, T. Zhang, Z. Li, M. Liu, and D. Zhang, "Towards evaluating the robustness of deep diagnostic models by adversarial attack," *Medical Image Analysis*, vol. 69, p. 101977, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841521000232

[35] N. Papernot, P. D. McDaniel, and I. J. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *CoRR*, vol. abs/1605.07277, 2016. [Online]. Available: http://arxiv.org/abs/1605.07277

[36] A. Azarang, H. E. Manoochehri, and N. Kehtarnavaz, "Convolutional autoencoder-based multispectral image fusion," *IEEE Access*, vol. 7, pp. 35 673–35 683, 2019.

[37] F. Soleymani and E. Paquet, "Financial portfolio optimization with online deep reinforcement learning and restricted stacked autoencoder - deepbreath," *Expert Syst. Appl.*, vol. 156, p. 113456, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:218786370

[38] M. Tayebi and S. E. Kafhali, "Hyperparameter optimization using genetic algorithms to detect frauds transactions," in *International Conferences on Artificial Intelligence and Computer Vision*, 2021.

[39] "Google Colaboratory — colab.research.google.com," https://colab.research.google.com, [Accessed 11-02-2024].

[40] "LUNA16 - Grand Challenge — luna16.grand-challenge.org," https://luna16.grand-challenge.org/Data, [Accessed 11-02-2024].

[41] D. K. Kavi, "Brain tumor image dataset," https://www.kaggle.com/denizkavi1/brain-tumor?select=3, 2021, accessed: 11-02-2024].

[42] M. ARIA, "Acute Lymphoblastic Leukemia (ALL) image dataset — kaggle.com," https://www.kaggle.com/datasets/mehradaria/leukemia, [Accessed 11-02-2024].

ATRAB AHMED received her B.Sc. and M.Sc. degrees in Information Technology from Mansoura University, Egypt, in 2011 and 2018, respectively. Currently, she is an Assistant Lecture in Information Technology, Faculty of Computers and Information, KafrElsheikh University, Egypt. She has published many research papers in reputable journals and prestigious international conferences. Her current research interests including security, data Science, network and machine learning. She can be contacted at email Atrab_Ahmed@fci.kfs.edu.eg.

Reda A. El-Khoribi is currently working as Dean of Faculty of Computers and Artificial Intelligence, Cairo University, Cairo, Egypt. He reviewed several papers for international journals and conferences. He is a professor at Faculty of Computers and Artificial Intelligence, Cairo University. His research interests include Pattern Recognition, Digital Signal Processing, Machine learning and deep learning. He can be contacted at email: r.abdelwahab@fci-cu.edu.eg.

NOUR ELDEEN KHALIFA received his B.Sc., M.Sc. and Ph.D. degree in 2006, 2009 and 2013 respectively, all from Cairo University, Faculty of Computers and Artificial Intelligence, Cairo, Egypt. He also had a Professional M.Sc. Degree in Cloud Computing in 2018. He authored/coauthored more than 40 publications and 2 edited books. He had more than 3000 citations. He reviewed several papers for international journals and conferences including (Scientific Reports, IEEE IoT, Neural Computing, and Artificial Intelligence Review). Currently, he is an associate professor at Faculty of Computers and Artificial Intelligence, Cairo University. His research interests include wireless sensor networks, cryptography, multimedia, network security, machine, and deep learning. He can be contacted at email: nourmahmoud@cu.edu.eg.