IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 28 March 2024, accepted 1 May 2024, date of publication 3 May 2024, date of current version 17 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3396566

RESEARCH ARTICLE

The Impact of Simultaneous Adversarial Attacks on Robustness of Medical Image Analysis

SHANTANU PAL[1], (Senior Member, IEEE), SAIFUR RAHMAN[1], (Member, IEEE), MAEDEH BEHESHTI[2], AHSAN HABIB[1], (Member, IEEE), ZAHRA JADIDI[3], AND CHANDAN KARMAKAR[1], (Member, IEEE)

[1]School of Information Technology, Deakin University, Melbourne, VIC 3125, Australia
[2]Critical Path Institute, Tucson, AZ 85718, USA
[3]School of Information and Communication Technology, Griffith University, Gold Coast, QLD 4222, Australia

Corresponding author: Shantanu Pal (shantanu.pal@deakin.edu.au)

ABSTRACT Deep learning models are widely used in healthcare systems. However, deep learning models are vulnerable to attacks themselves. Significantly, due to the black-box nature of the deep learning model, it is challenging to detect attacks. Furthermore, due to data sensitivity, such adversarial attacks in healthcare systems are considered potential security and privacy threats. In this paper, we provide a comprehensive analysis of adversarial attacks on medical image analysis, including two adversary methods, FGSM and PGD, applied to an entire image or partial image. The partial attack comes in various sizes, either the individual or combinational format of attack. We use three medical datasets to examine the impact of the model's accuracy and robustness. Finally, we provide a complete implementation of the attacks and discuss the results. Our results indicate the weakness and robustness of four deep learning models and exhibit how varying perturbations stimulate model behaviour regarding the specific area and critical features.

INDEX TERMS Machine learning, deep learning, medical image analysis, robustness, adversarial attacks.

I. INTRODUCTION

Image analysis is among the essential parts of Artificial Intelligence (AI) technologies. It plays a significant role in various healthcare and medical fields, including radiology, pathology, and ophthalmology [1]. One of AI's most crucial healthcare utilization is the identification of biomarkers through image analysis, which is particularly effective in disease diagnosis, cure management, and therapeutic drug development. Recently, deep learning algorithms have made it feasible to examine massive amounts of medical images and identify subtle changes that might point to the existence or progression of disease in an early stage [2]. However, adversarial attacks in such AI-based systems create risks of data manipulation or unauthorized access to data. This leads to inaccurate disease detection or, more importantly, impacts the healthcare system to more significant issues [3]. For example, an attacker can manipulate medical reports to commit insurance fraud. Further, the attacker can disrupt and mislead the patient's diagnosis, severely impacting the patient's well-being [4]. Despite the superior performance, recent studies examine that deep learning models are vulnerable to adversarial attacks [5], [6]. This has raised concerns about the data (including ground truth) used to train the algorithm, the inappropriate proposed model, or even how the machine learning program is ultimately deployed. This demands a more comprehensive analysis of secure and robust medical deep-learning systems that can effectively understand and detect adversarial attacks in the healthcare system.

Available proposals, e.g., [7], [8], and [9], show the impact of adversarial attacks on medical image processing. They discuss the viewpoint of the adversarial perturbations' impact on medical imaging and its consequences. Unlike these available proposals, we intend to show the variation in the attack's performance based on the different categories of attacks. We provide a comprehensive understanding of adversarial attacks in medical image processing, both generating and detecting the attacks. We study how a basic noise in an image can cause a hole or breakthrough for an adversarial attack.

The associate editor coordinating the review of this manuscript and approving it for publication was Yiming Tang.

Furthermore, we use three distinct medical image datasets, chest X-ray, Optical Coherence Tomography (OCT), and skin cancer, to examine the performance. We envision a situation in which more than one attacker has thoroughly or partially manipulated one image. Our research provides valuable insights for practical healthcare settings by assessing the impact of simultaneous adversarial attacks on the robustness of medical image analysis. Understanding how these attacks affect the accuracy and reliability of deep learning models is crucial for ensuring the integrity of diagnostic systems. For example, healthcare practitioners can use these findings to develop more robust and secure image analysis algorithms, enhancing the trustworthiness of diagnostic processes in real-time.

Two different adversarial perturbations, e.g., Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) [10], have been chosen to show our strong simultaneous attack on diversity. By comprehensively analyzing adversarial attacks, e.g., FGSM and PGD, healthcare practitioners gain insights into potential vulnerabilities and privacy threats. Understanding how varying perturbations affect model behaviour regarding specific areas and critical features enables healthcare systems to implement robust defences. Our research helps to develop the trustworthiness and resilience of deep learning-based healthcare systems by pointing out the potential threats. The presented results and visualization provide our dataset's weaknesses and inclination to each non-overlap attack. In addition, an overlap patched attack has been investigated to show our model behaviour under small invisible perturbations. We consider both transferred and basic scratched deep learning models to identify the significant roles of obtaining appropriate AI algorithms. To the best of our knowledge, no previous work has investigated the impact of simultaneous adversarial attacks on the robustness of medical image processing. In addition, evaluating the results of a *full*, *partial*, and *patch* attacks entirely is a unique task in this research (a detailed discussion in Section III-D). The major contributions of the paper can be summarized as follows:

- We analyse the leading computer vision DL model with proposed three novel attack scenarios—full, partial, and patch attacks—on medical datasets and examine their impact on medical images. We aim to demonstrate the model's effectiveness against internal and external threats through these proposed scenarios. Thus, we criticise the existing model, urging researchers to develop a robust solution to handle the three novel attack scenarios.
- Presenting a novel implementation of simultaneous attacks on three available datasets to investigate how two or more attacks impact the dataset simultaneously.
- Proposing a pipeline when our proposed model is manipulated by different attacks simultaneously. For example, we examine what happens to the model's accuracy if an attack impacts 20% of an image at a particular time and again at the same time, another attack affects the rest of 80% of the same image.
- Our study offers a thorough understanding of a pipeline when a deep learning model is manipulated by various attacks simultaneously.

The rest of the paper is organized as follows. In Section II, we discuss related works. In Section III, we discuss the system design and methodology of the proposed experiment. This section includes, the description of the datasets used for the experiment, models, types of attacks, and attack scenarios. In Section IV, we present the achieved results and provide a detailed experimental evaluation. In Section V, we discuss the lessons learned. Finally, in Section VI, we conclude the paper with future work.

II. RELATED WORK

The vulnerability and robustness of medical images against different formats of adversarial attacks are the main focus of this work. Several recent studies are dedicated to this content and try to address the open questions in a professional pathway. Although all efforts are valuable and give us a unique observation of a specific angle of the problem, there is still a long way to reach a compromise on an acceptable confidential answer. Bortsova et al. [7] express the unexplored vulnerability factors of the health image datasets. The authors claim a direct relationship exists between the transferability of adversarial examples and pre-trained models in black-box attacks. This high transferability perturbation motivates us to evaluate pre-trained and non-pre-trained models to examine adversarial attacks, e.g., full, partial, and patch attacks.

The adversarial patches [11], [12], [13], [14] struggle to urge the developed models to conclude in a misclassification with more facilitating features included in the patches, e.g., localization, motion, and invisibility.

Studies, e.g., [15], [16], and [17], introduce an adversarial patch method that claims the patches are universal, regardless of the scene and variety of transformations. For example, attaching a sticker [18], with a specific target, the attackers try to fool the classifiers without considering the amount of invisibility or imperceptibility of the attack. Furthermore, Su et al. discuss [19] a one-pixel attack technique that is another valuable subset of patch attacks based on evolutionary strategies. Regarding dynamic patch generation, Li and Ji [20] present a flexible model capable of producing visible or non-visible patches that can move around an image to find a suitable attack position. In another investigation, Mohapatra et al. [21] discuss semantic perturbation over discrete and continuous parameters that allow the model to predict based on dimensional perturbations. Keerthana et al. introduce a hybrid Convolutional Neural Network (CNN) model with SVM classifiers to classify dermoscopy images as benign or melanoma lesions. The proposed approach reduces inter-operator variability by concatenating features extracted from two CNN models [33]. Karthik and Mahadevappa discuss an enhanced CNN model for OCT image analysis in retinal disease diagnosis [34].

The model outlines improving the feature map contrast by modifying residual connections and activation functions, increasing classification accuracy on OCT datasets. Agrawal and Choudhary present a lightweight CNN (ALCNN) for pneumothorax detection in chest X-ray images, comparing it with transfer learning approaches [35]. ALCNN achieves comparable results with 10x fewer parameters than VGG-19 and ResNet-50 architectures. The study demonstrates the effectiveness of ALCNN, suggesting transfer learning has minimal impact on performance. Other proposals, e.g., [8], [22], [23], [24], [25], and [26], have shown more investigation of the robustness of pre-trained models against perturbation attacks on medical image datasets, e.g., skin, chest, and diabetic medical images.

To the best of our knowledge, no previous proposal aims to address a simultaneous behaviour of perturbations in one image. To address this issue, we aim to show how different parts of an image are vulnerable to a simultaneous (including partial, and patch) attack. We use two different types of adversarial attacks, FGSM and PGD, to observe the pretrained and non-pretrained models' robustness against different attack scenarios, which is discussed in the next section.

III. SYSTEM DESIGN AND METHODOLOGY

In this section, we discuss the system design and the methodology of our proposed experiment, inluding the (i) dataset, (ii) models, (iii) types of attacks, (iv) attack scenarios used in the experiment, and (v) the experimental setup.

A. DATASETS

We use three medical datasets, e.g., chest x-ray (two classes) [27], OCT (four classes) [28], and skin cancer (six classes) [29] to train the CNN models. A brief description of them is as follows:

- *Chest x-ray dataset:* It contains x-ray images of the lungs, developed by the University of California San Diego and updated in 2018. The dataset has a total of 5, 863 images with the.jpeg file extension. The dataset has two folders, train and test. In each of the folders, there are two folders named Normal and Pneumonia, which separate the images into two categories.
- *Retinal OCT dataset:* It contains x-ray images of the retina from the University of California San Diego, updated in 2018. The dataset has a total of 84, 484 images with the.jpeg file extension. The dataset has two folders, train and test. In each of the folders, there are four folders named Normal, CNV, DME, and Drusen, which separate the images into four categories.
- *Skin cancer dataset:* It contains skin cancer images from the International Skin Imaging Collaboration (ISIC) Archive, updated in 2019. The dataset has a total of 10, 015 images with the.jpg file extension. The dataset has two folders, benign and malignant, which separate the images into two categories. Once trained, the CNN will be able to classify skin cancer moles as actinic keratoses, basal cell carcinoma, dermatofibroma, melanocytic nevi, melanoma, and vascular lesions. Eventually, each of the six classes belongs to either benign or malignant categories. However, we only stop with six classifications and do not consider further analysis to assign each to the benign or malignant category.

B. MODELS

We use three models, VGG, ResNet152v2, and Xception, pre-trained with the imagaeNet dataset, and one non pre-trained model, Alexnet. Note, in this paper, we fine-tuned all the pre-trained models with three different datasets. We also modified the output layer based on the number of classes presented in the individual dataset. A brief description of these models is as follows:

- *VGG Models:* It is a family of CNN architectures with small convolutional filters (typically 3×3) and a deep stack of convolutional layers, which allows them to learn fine-grained features from images. VGG16 (Visual Geometry Group) pre-trained consists of 16 layers, including several blocks of convolutional layers with max-pooling layers and a couple of fully connected layers at the end [30].
- *ResNet152v2:* It is a type of CNN architecture that utilizes a concept called 'residual connections' to enable the training of intense networks. The key idea behind ResNet is to add 'shortcut' or 'skip' connections that allow the gradients to flow more quickly through the network during training, which helps to mitigate the vanishing gradients problem that can occur in intense networks. The significant difference between ResNetV2 and the original (V1) is batch normalization before each convolutional weight layer. ResNets are effective for many image and video recognition tasks and have been used in many state-of-the-art models [31]. Compared with VGG, ResNets contain lower filters and, consequently, less complexity.
- *Xception:* The Xception architecture as an extension of the inception architecture consists of a stack of depth-aware separable convolutional layers (instead of the standard Inception modules) and a few more layers for handling the network's input and output. In addition, it contains a fully connected layer for classification after a global average pooling layer. The depthwise separable convolutional layers comprise a depthwise convolution operation that uses one filter for each input channel and a pointwise convolution operation that combines the results of the depthwise convolution to create the final output [32].
- *Alexnet:* It has eight layers, five of which are convolutional and three are completely connected. Each convolutional layer's output is subjected to the network's Rectified Linear Unit (ReLU) activation function. Additionally, it uses local response normalization (LRN) to enhance generalization and avoid overfitting.

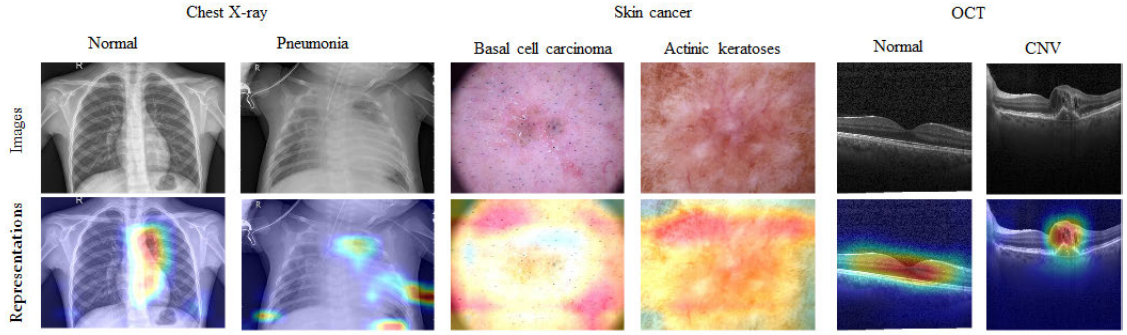

FIGURE 1. The normal images (top row), and their representations (bottom row) learned at the "block5 conv3" layer of VGG model (averaged over channels) of the networks.

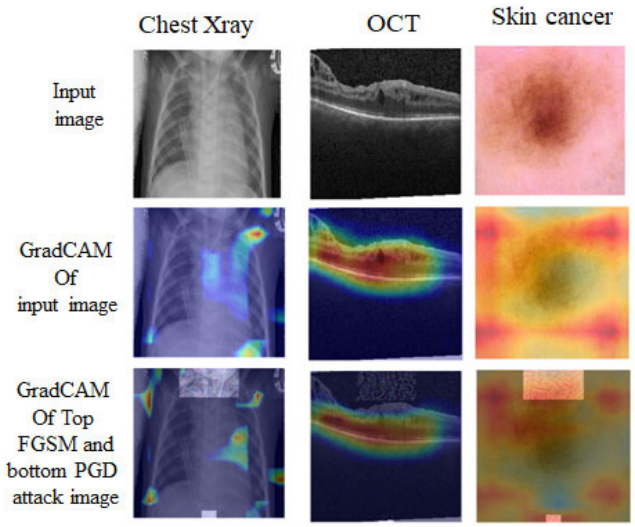

FIGURE 2. The normal images (top row), GradCAM without attack (middle row), and GradCAM with an attack (bottom row) learned at the "block5 conv3" layer of VGG model (averaged over channels) of the networks.

C. TYPES OF ATTACK

In this paper, we use the following two attacks. A short description for each of them is as follows:

- *Fast Gradient Sign Method (FGSM) Attack:* It is a white-box attack where the attacker is aware about the model parameters and architecture and is able to use them as a prior knowledge. The basic idea behind FGSM is to add perturbation to the input images that hamper the model classification capacity in a right way. Fooling the system from the main goal or achieving a specified target is the main malicious purpose of this attempt. Firstly, computing the gradient of the model's output and then adding a small multiple (epsilon, ϵ) of the sign of the gradient to the input image. FGSM is a simple but effective attack technique that has been established to be impactful against a variety of machine learning models, including deep neural networks [10]. FGSM can be defined by the following Equation 1:

$$\begin{aligned} Med_{adv} &= Med_{raw} + \epsilon \times loss_{fun}, \\ loss_{fun} &= sign(\nabla^{k}_{Med_{raw}}(\sigma, Med_{raw}, Med_{tar})) \end{aligned} \tag{1}$$

where $loss_{fun}$ is gradient of the loss function k with respect to Med_{raw}, and $sign$ is the sign function, Med_{adv} is the adversarial image with FGSM attack, Med_{raw} is the original image, Med_{tar} is the original label, σ is the model parameters, and ϵ is the strength of the FGSM attack.

- *Projected Gradient Descent (PGD) Attack:* It is similar to the FGSM attack, called white-box attacks (attackers know the model behavior). PGD applies iteratively using the gradient of the model's output, where it considers the non-linearity of the model until the input image is adversarial [10]. PGD can be defined by the following Equation 2:

$$\begin{aligned} Med_{adv}(I+1) &= clip_{(-\epsilon,\epsilon)}(I(Med_{raw} + \gamma \times loss_{fun})), \\ loss_{fun} &= sign(\nabla^{k}_{Med_{raw}}(\sigma, Med_{raw}, Med_{tar})) \end{aligned} \tag{2}$$

That is, for those pixels with perturbation size larger than ϵ, clip truncates it to ϵ. $loss_{fun}$ gradient of the loss function k with respect to Med_{raw}, and $sign$ is the sign function, $Med_{adv}(I+1)$ is the adversarial image with PGD attack, Med_{raw} is the original image, Med_{tar} is the original label, σ is the model parameters, and ϵ is the strength of the PGD attack and I is the number of iterations to achieving adversarial image.

D. ATTACK SCENARIOS

In this paper, both FGSM and PGD attacks are implemented and explored on the three above mentioned medical image datasets with *full*, *partial*, *simultaneous*, and *patch* attacks. We visualize the model's layers of clean and attacked images to determine the most prominent areas in image classification using GradCAM methods for a better understanding of further evaluation. GradCAM highlights the gradient of the input image, which is used for predicting different classes shown in Fig 1 and Fig 2. By visualizing the gradient of the original and attacked images, we discuss the impact of simultaneous adversarial attacks on images and will show the model performance against those attacks. A brief description of each of them is as follow:

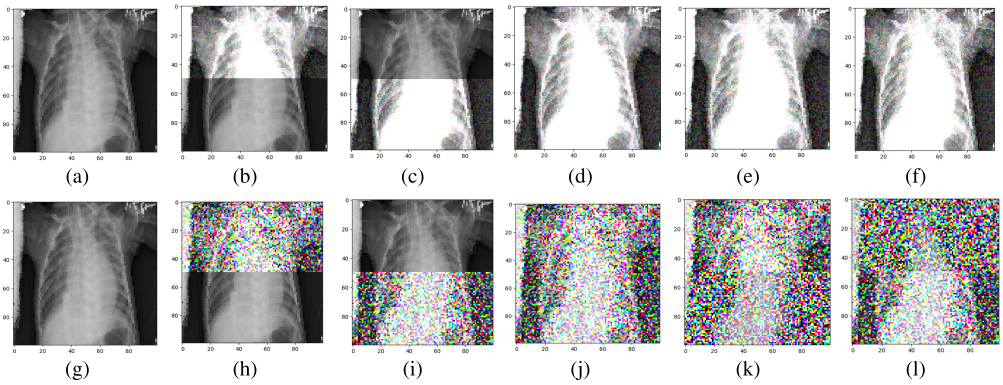

FIGURE 3. Chest x-ray using VGG (a and g) Original image (b and h) Top FGSM (c and i) Bottom FGSM (d and j) Full FGSM (e and k) Top FGSM and bottom PGD and (f and l) Top PGD and bottom FGSM attack on the original image with distortion, $\epsilon = 0.1$ (Top row) and 0.9 (Bottom row).

- ***Full Attacks:*** In this scenario, we consider an attacker invades the full image shown in Fig. 3 (d,j), Fig. 5(d,j), Fig.6 (d,j) using Algorithms 1 and 2.

Algorithm 1 Proposed PGD Attack

Input: clean input image Med_{raw}, target label Med_{tar}, neural network f (VGG, Resnet, Alexnet, Xception), number of PGD iteration I, perturbation size ϵ
Output: adversarial image Med_{adv}
$index = 0$
for *model in* f **do**
  for $\epsilon \leftarrow 0$ **to** 0.9 *by* 0.1 **do**
    $Med_{fullAdv}[index] \leftarrow Med_{raw} + PGD(I, \epsilon)$;
    $Med_{partialAdv}[index] \leftarrow Med_{raw} + PGD(I, \epsilon)$;
    $Med_{patchAdv}[index] \leftarrow Med_{raw} + PGD(I, \epsilon)$;
    $index+ = 1$;
  end for
end for
return $Med_{fullAdv}$, $Med_{partialAdv}$, $Med_{patchAdv}$;

- ***Partial Attacks:*** In this category, we show the attacks in a partial part of the image. We consider various attacks positions, e.g., only top half, or only bottom half of an image (either top half PGD, or bottom half FGSM) for generating an adversarial image using different epsilon values shown in Figs. 3 to 6 using Algorithms 1 and 2.
  - *Individual Partial Attacks:* This signifies an attack on a persistent partial position only either FGSM or PGD. Fig. 3 (b,h,c,i), Fig. 5 (b,h,c,i), Fig.6 (b,h,c,i)
  - *Simultaneous Partial Attacks:* This signifies an attack where both FGSM and PGD occur simultaneously. Fig. 3 (f,l,e,k), Fig. 5 (f,l,e,k), Fig.6 (f,l,e,k) using Algorithm 3.
- ***Patch Attacks:*** In this category, we consider two fixed positions, e.g., center and top, for generating an adversarial image using different patch sizes, e.g., 16×16, 32×32,48×48 and 64×64, where epsilon values (ϵ) are 0.1, 0.5 and 0.9 shown in Fig. 7. This type of attack showed which area of the image was more vulnerable versus the attacks, and the model cannot classify diseases with high accuracy when the attack

Algorithm 2 Proposed FGSM Attack

Input: clean input image Med_{raw}, target label Med_{tar}, neural network f (VGG, Resnet, Alexnet, Xception), perturbation size ϵ
Output: adversarial image Med_{adv}
$index = 0$
for *model in* f **do**
  for $\epsilon \leftarrow 0$ **to** 0.9 *by* 0.1 **do**
    $Med_{fullAdv}[index] \leftarrow Med_{raw} + FGSM(\epsilon)$;
    $Med_{partialAdv}[index] \leftarrow Med_{raw} + FGSM(\epsilon)$;
    $Med_{patchAdv}[index] \leftarrow Med_{raw} + FGSM(\epsilon)$;
    $index+ = 1$;
  end for
end for
return $Med_{fullAdv}$, $Med_{partialAdv}$, $Med_{patchAdv}$;

Algorithm 3 Proposed Patch Attack

Input: clean input image Med_{raw}, target label Med_{tar}, neural network f (VGG, Resnet, Alexnet, Xception), perturbation size ϵ
Output: adversarial image Med_{adv}
$index = 0$
for *model in* f **do**
  for $\epsilon \leftarrow 0$ **to** 0.9 *by* 0.1 **do**
    $Med_{topAdv}[index] \leftarrow (20\%) \times Med_{raw} + FGSM(\epsilon) + (5\%) \times Med_{raw} + PGD(I, \epsilon)$;
    $Med_{bottomAdv}[index] \leftarrow (5\%) \times Med_{raw} + PGD(I, \epsilon) + (20\%) \times Med_{raw} + FGSM(\epsilon)$;
    $index+ = 1$;
  end for
end for
return $Med_{bottomAdv}$, Med_{topAdv};

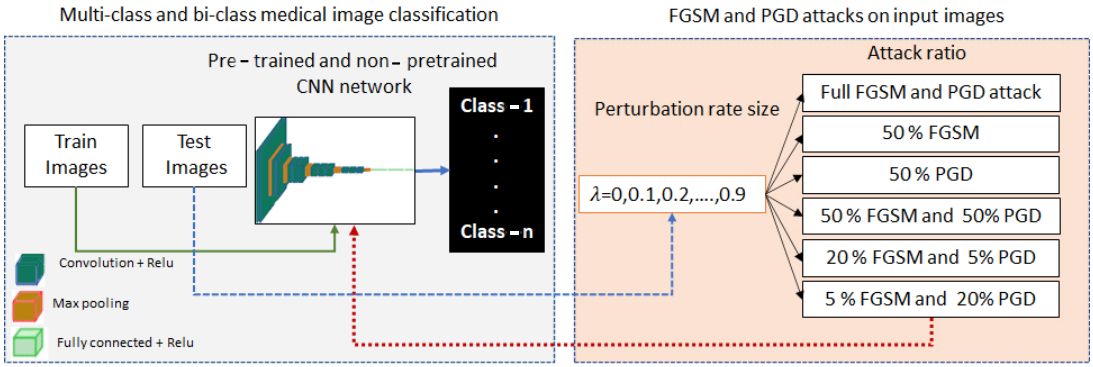

FIGURE 4. The pipeline of training DNNs (left) and generating adversarial attacks (right). The green solid line indicates the training datasets.The red dotted line indicates the perturbation image for testing. Finally, the blue dotted line indicates the raw test data without attacks.

occurs in those areas. We further category these attacks into the following two sub-categories

- *Individual Patch Attacks:* This signifies a patch attack on a persistent partial position only, either FGSM or PGD. Fig. 7 using Algorithms 1 and 2.
- *Simultaneous Patch Attacks:* This signifies a patch attack where both FGSM and PGD occur simultaneously. Figs. 14, 15, 16 using Algorithm 3.

Note that our proposed *simultaneous* types of attacks provide us further knowledge about additional attack influence on critical parts of the image compared with only one attack in the traditional format. It was essential to show which area of the image was more vulnerable versus the concurrent attacks. In addition, how the diversity of models improves resilience to classify diseases when a conjugation attack occurs in those areas.

E. EXPERIMENTAL SETUP

In this experiment, we use the three most popular pretrained transfer learning in computer vision for deep learning models called, VGG, Xception, and ResNet152v2, in addition to Alexnet as a non-trained model for evaluating our training model from scratch. All the models were trained on three different medical datasets (chest x-ray, OCT, and skin cancer) without any perturbation. We test the model's performance with perturbation using different scenarios along with various epsilon values and eventually summarise the capacity of the trained models for classifying different diseases and the strength of the adversarial attack on medical images. Fig. 4 depicts our approach in adversarial attack as a pipeline.

Employing a server to conduct the training process is advantageous since building a CNN model demands substantial computer resources. Here is a list of setups that we follow in our experiments: (i) *Hardware:* we use a Linux server with a Tesla V100 GPU, which has 32.51 GB of memory, (ii) *Software:* we use python, and the CNN model is created using a Tensorflow environment, finally, (iii) *Data Preparation:* before training the CNN model, the data must be preprocessed and organized in a format suitable for training. We have resized our images into 224 × 224 size as the images originally were not the same size in each dataset. We also augmented datasets using random shear, zoom, and horizontal flipping criteria. This is a mandatory augmentation only for dataset preparation and is not related to any attack models.

IV. EXPERIMENTAL RESULTS AND EVALUATION

In this section, we discuss the experimental results which are evaluated using the accuracy defined by Equation 3:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{3}$$

where, TP = a positive image classified as positive (True Positive) FN = a positive image misclassified as negative (False Negative), TN = a negative image classified as negative (True Negative), and FP = a negative image misclassified as positive (False Positive).

For each step in our different attack scenarios discussed earlier, we paid full attention to the perturbation degree analysis over ϵ variety, across investigating the efficiency of both FGSM and PGD attacks in different formats regarding the existing three datasets [7]. For PGD, we have used two iterations. In addition, we have added the results of non-attack images compared with the attacked ones by ϵ, 0. Finally, components describing our results based on pre-trained and non-trained models are discussed as follows:

A. FULL ATTACKS

In this section, we discuss the results of full white attacks on the three datasets including Chest X-ray, OCT, and Skin cancer using 4 deep-learning models - VGG, Resnet, Xception, and Alexnet.

1) CHEST X-RAY DATASET

In Fig. 8-a, the accuracy of Resnet and Xception FGSM attacks starts at around 80% with no-attack imaging ($\epsilon = 0$) and reduces by around 20% and 15% at $\epsilon = 0.1$ attack but almost converges together with other ϵ values. Whereas, the

TABLE 1. Performance of Alexnet for chest x-ray dataset for different epsilon (ϵ).

Epsilon	FGSM full	PGD full	FGSM top	PGD top	FGSM bottom	PGD bottom	FGSM-PGD	PGD-FGSM
0.00	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64
0.10	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64
0.20	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64
0.30	0.64	0.64	0.64	0.64	0.64	0.64	0.64	0.64
0.40	0.64	0.67	0.64	0.64	0.64	0.64	0.64	0.64
0.50	0.64	0.44	0.64	0.64	0.64	0.64	0.64	0.64
0.60	0.64	0.38	0.64	0.64	0.64	0.64	0.64	0.64
0.70	0.64	0.36	0.64	0.64	0.64	0.64	0.67	0.61
0.80	0.64	0.36	0.64	0.64	0.64	0.64	0.59	0.52
0.90	0.64	0.36	0.64	0.64	0.64	0.64	0.39	0.41

TABLE 2. Performance of Alexnet for OCT dataset for different epsilon (ϵ).

Epsilon	FGSM full	PGD full	FGSM top	PGD top	FGSM bottom	PGD bottom	FGSM-PGD	PGD-FGSM
0.00	0.61	0.61	0.61	0.61	0.61	0.61	0.61	0.61
0.10	0.09	0.13	0.11	0.14	0.14	0.11	0.13	0.11
0.20	0.13	0.06	0.13	0.06	0.19	0.06	0.06	0.06
0.30	0.16	0.06	0.11	0.06	0.16	0.06	0.06	0.06
0.40	0.16	0.06	0.16	0.06	0.14	0.06	0.06	0.06
0.50	0.16	0.06	0.16	0.06	0.16	0.06	0.06	0.06
0.60	0.16	0.06	0.16	0.06	0.16	0.06	0.06	0.06
0.70	0.16	0.06	0.16	0.06	0.16	0.06	0.06	0.06
0.80	0.16	0.06	0.16	0.06	0.16	0.06	0.06	0.06
0.90	0.16	0.06	0.16	0.06	0.16	0.06	0.06	0.06

TABLE 3. Performance of Alexnet for skin cancer dataset for different epsilon (ϵ).

epsilon	FGSM full	PGD full	FGSM top	PGD top	FGSM bottom	PGD bottom	FGSM-PGD	PGD-FGSM
0.00	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73
0.10	0.08	0.00	0.56	0.17	0.66	0.55	0.00	0.00
0.20	0.03	0.00	0.48	0.00	0.67	0.00	0.00	0.00
0.30	0.02	0.00	0.05	0.00	0.27	0.00	0.00	0.00
0.40	0.02	0.00	0.02	0.00	0.23	0.00	0.00	0.00
0.50	0.48	0.00	0.02	0.00	0.23	0.00	0.00	0.00
0.60	0.52	0.00	0.03	0.00	0.25	0.00	0.00	0.00
0.70	0.56	0.00	0.03	0.00	0.64	0.00	0.00	0.00
0.80	0.56	0.00	0.06	0.00	0.67	0.00	0.11	0.00
0.90	0.23	0.02	0.06	0.00	0.67	0.00	0.16	0.06

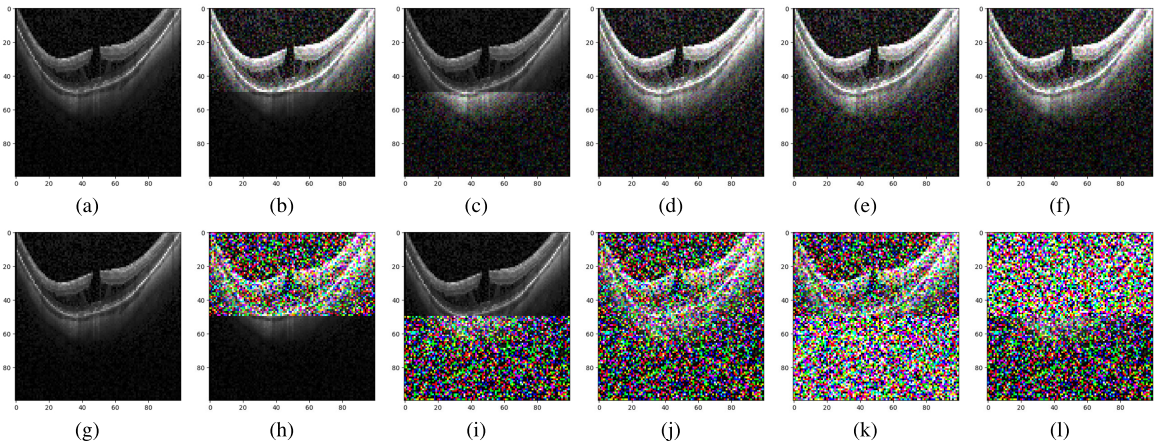

FIGURE 5. OCT using VGG (a and g) Original image (b and h) Top FGSM (c and i) Bottom FGSM (d and j) Full FGSM (e and k) Top FGSM and bottom PGD and (f and l) Top PGD and bottom FGSM attack on the original image with distortion, $\epsilon = 0.1$ (Top row) and 0.9 (Bottom row).

accuracy of VGG FGSM is around 65%, which is slightly lower than FGSM of the other 2 models, decreases around 25% at $\epsilon = 0.1$ but gradually increases and touches the FGSM of the other 2 models at $\epsilon = 0.3$ onward. The accuracy of Xception PGD coincides with its FGSM with $< 5\%$ margin across ϵ values. However, the accuracy

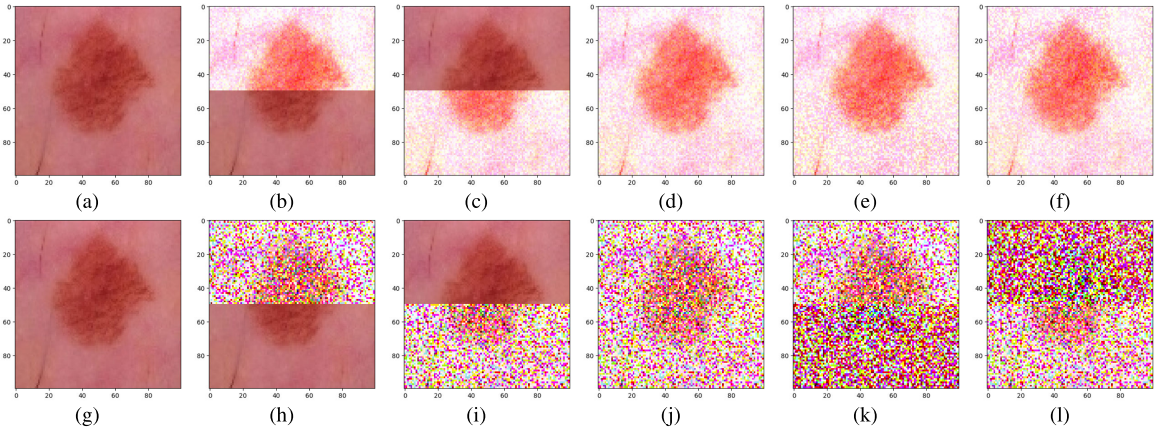

FIGURE 6. Skin cancer using VGG (a and g) Original image (b and h) Top FGSM (c and i) Bottom FGSM (d and j) Full FGSM (e and k) Top FGSM and bottom PGD and (f and l) Top PGD and bottom FGSM attack on the original image with distortion, $\epsilon = 0.1$ (Top row) and 0.9 (Bottom row).

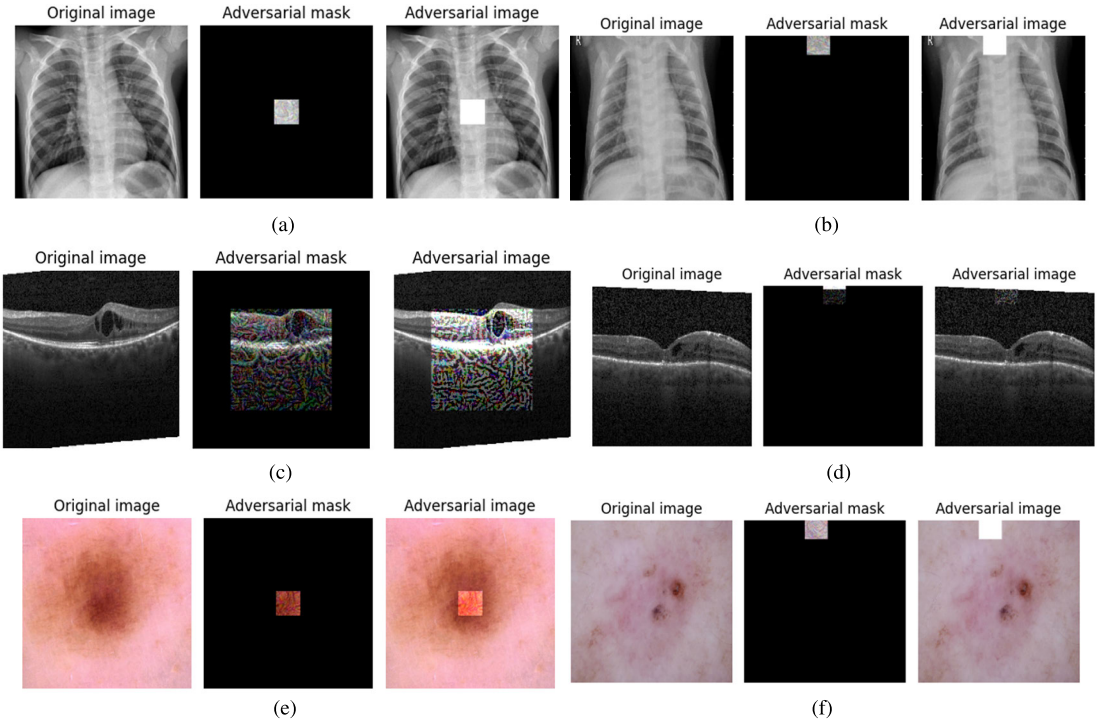

FIGURE 7. Chest x-ray using VGG (a and b) central and top for ϵ =0.1 and 16 × 16 patch size; OCT (c and d) central and top for ϵ =0.9 and 64 × 64 patch size; skin cancer (e and f) central and top attack for ϵ =0.1 and 16 × 16 patch size, all on the original image.

of VGG and Resnet PGDs drops by around 30% at $\epsilon = 0.1$ from no-attack respective accuracy of around 65% and 80%. While VGG PGD continues steadily to keep robustness, Resnet PGD struggles to reach a reasonable accuracy by increasing the ϵ values. As Table 1 exhibits, Alexnet-FGSM and Alexnet-PGD treat the same with 0.64 accuracies until ϵ (=0.5). After this middle perturbation, they continue separately. Alexnet-FGSM steadily continues with 0.64 while Alexnet-PGD starts to decline constantly.

2) OCT DATASET

Fig. 9-a shows classification accuracy for the two types of full attacks, PGD and FGSM, applied on the images of the OCT dataset, which shows the drop in accuracy between non-attack ($\epsilon = 0$) and smallest attack ($\epsilon = 0.1$) is within around 50%-70% across the models. FGSM accuracy of all 3 models VGG, Resnet, and Xception (blue, green and purple in Fig. 9-a) at the minimum $\epsilon = 0.1$ are around 10%, 25% and 35% respectively and with higher ϵ, Resnet and Xception FGSM accuracy maintains this level with $< 10\%$ margin which was reached by the VGG FGSM at $\epsilon = 0.7$. PGD accuracy of VGG and Xception were $< 10\%$ at $\epsilon = 0.1$ and dropped to almost zero level with high ϵ values, whereas the Resnet PGD accuracy is 10% at $\epsilon = 0.1$ which increases with higher ϵ with around 10% margin. As Table 2 exhibits, the Alexnet model does not show any appropriate behavior

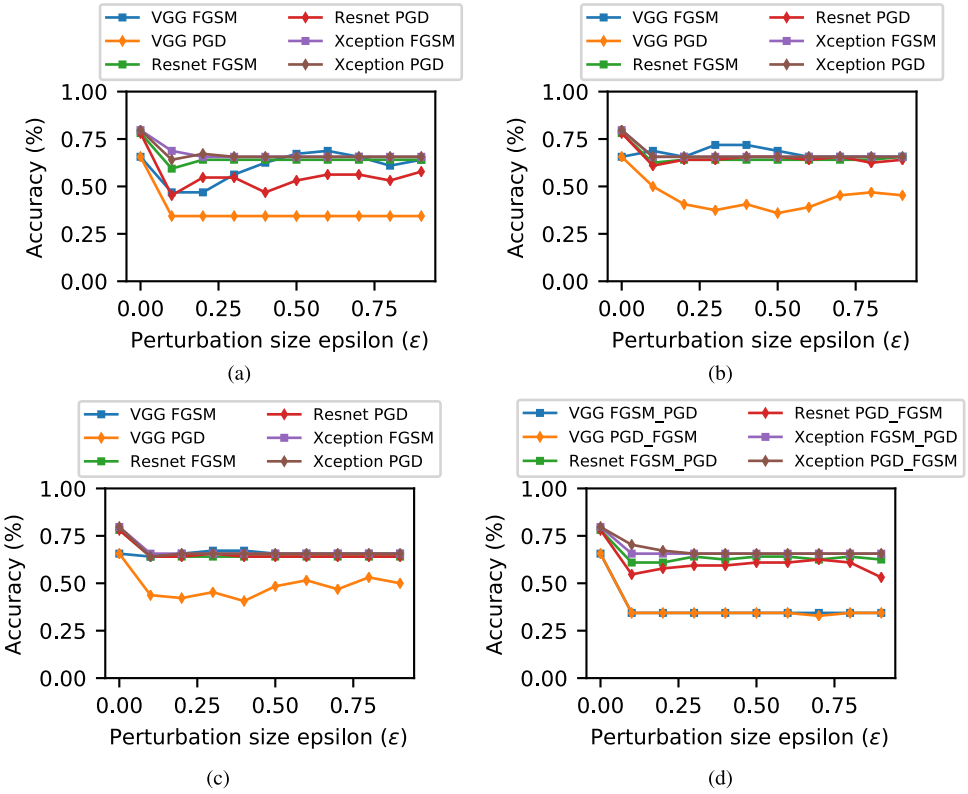

FIGURE 8. Chest X-ray (a) Full (b) Top (c) Bottom (d) FGSM-PGD and PGD-FGSM attack on the original image with perturbation size, ϵ.

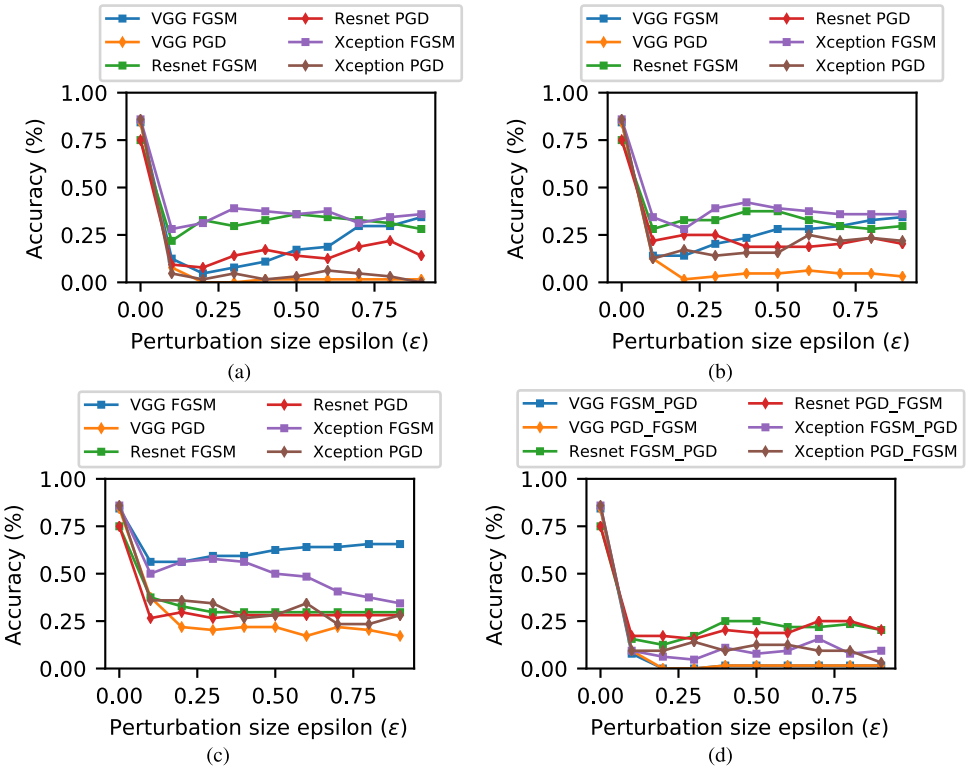

FIGURE 9. OCT (a) Full (b) Top (c) Bottom (d) FGSM-PGD and PGD-FGSM attack on the original image with perturbation size, ϵ.

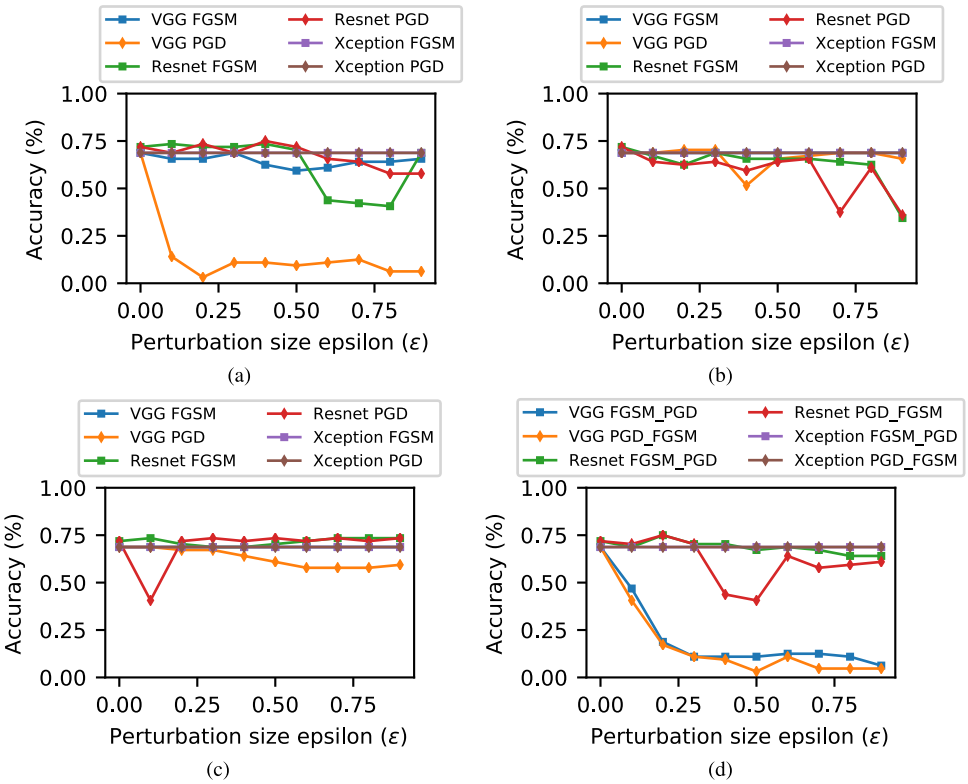

FIGURE 10. Skin cancer (a) Full (b) Top (c) Bottom (d) FGSM-PGD and PGD-FGSM attack on the original image with perturbation size, ϵ.

facing the full OCT attack. But as observed for the most pre-trained models, FGSM shows small better robustness in this regard.

3) SKIN CANCER DATASET

Fig. 10-a shows the accuracy of Xception FGSM and PGD maintain around 70% across all ϵ values. In contrast, others, except VGG PGD, which drops below 10% from $\epsilon = 0.1$ onward, maintain almost the same accuracy level until $\epsilon = 0.3$ beyond which Resnet PGD and FGSM accuracy dropped around 10% and 30%.

B. PARTIAL ATTACKS

In this section, we discuss the results of partial white attacks as two configurations i) individual partial attack (either FGSM or PGD applied on 50% of the image at the top or bottom part at a time) and ii) simultaneous partial attack (FGSM and PGD concurrently applied on top and bottom 50% respectively and vice versa) on the three datasets including Chest X-ray, OCT, and Skin cancer based on 4 deep-learning models - VGG, Resnet, Xception, and Alexnet.

1) INDIVIDUAL PARTIAL ATTACKS

The results of the individual partial attacks are described and grouped by 3 datasets.

a: CHEST X-RAY DATASET

Figs. 8-b and c, bring us to a partial attack that impacts only the top 50% and bottom 50%, respectively, of each image in the dataset. Accuracy in both the figures is around 80% for both the attacks concerning Resnet and Xception models. They drop to around 65% at $\epsilon = 0.1$ and maintain the same level across other ϵ values, which indicates the attacks have no significant impact on these models' performance beyond $\epsilon = 0.1$. VGG FGSM and PGD show around 65% accuracy at no-attack ($\epsilon = 0$) case, while VGG FGSM was able to maintain the same level, VGG PGD accuracy dropped to below 30% for both the top and bottom 50% attack scenarios. Based on Table 1 observation, FGSM and PGD attacks on the top and bottom of images all provide the same response from the Alexnet model. There might be any bias or overfitting for the models, but we observe the results to show a picture of scratched models.

b: OCT DATASET

Figs. 9-b and c show accuracy at no-attack ($\epsilon = 0$) scenario in the range 75-80% for all three models (VGG, Resnet and Xception), and at $\epsilon = 0.1$, represent declines for both top and bottom areas. By increasing the ϵ values, partial FGSM or PGD attacks for all the models drop below 50%, except bottom VGG FGSM and Xception FGSM where they dropped less, at around 50% with marginal fluctuation maintaining the accuracy level at high. Both the figures (Figs. 9-b and c) point out that PGD accuracy is lower than FGSM accuracy regardless of models.

Table 2 does not show any good results of performance in the partial top and bottom attacks while the FGSM model treats a little better.

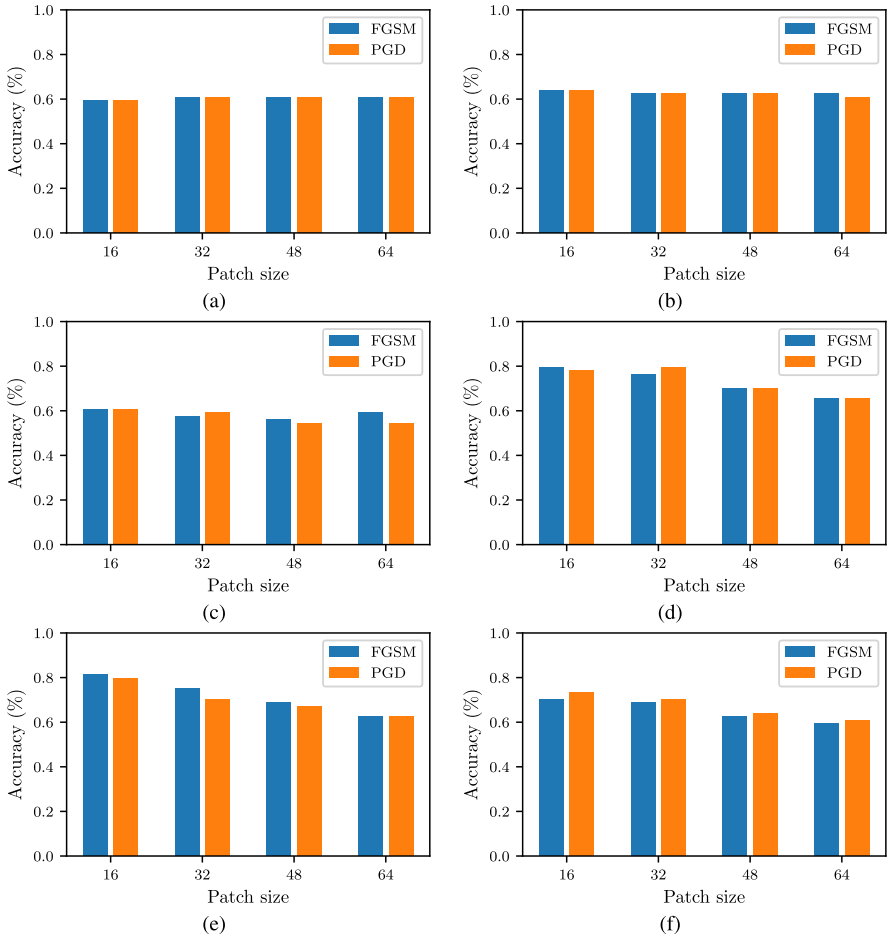

FIGURE 11. Individual patch attack on Chest X-ray dataset, (a)-(c) central-patch attack, and (d)-(f) top-patch attack, the first column (a and d) represents perturbation of ϵ=0.1, middle column (b and e) ϵ=0.5 and the 3rd column (c and f) ϵ=0.9, using VGG model.

c: SKIN CANCER DATASET

Figs. 10-b and c show accuracy of corresponding top 50% and bottom 50% partial attack scenarios where no-attack ($\epsilon = 0$) case achieved around 70% accuracy which was maintained by all three models. The accuracy continues to stay high within around 10% margin except for VGG and Resnet PGD in the top partial attack and VGG PGD in the bottom partial attack, which dropped beyond the margin. In Table 3, the top and bottom are not resistant to the partial attacks, but FGSM shows some fluctuations to reach a better accuracy and stay robust against the partial attack, specifically at the bottom.

2) SIMULTANEOUS PARTIAL ATTACKS

The results of simultaneous partial attacks are described and grouped by the datasets.

a: CHEST X-RAY DATASET

Fig. 8-d shows two accuracy values of around 80% and 70% for a category of models including Resnet and Xception, and the 2nd category consisting only VGG models at $\epsilon = 0$. At $\epsilon = 0.1$, the accuracy of both categories decreased by different amounts. Still, the 2nd category of VGG models (VGG FGSM-PGD and PGD-FGSM) lag behind the first category by around 20-30% accuracy margin. Then both categories maintain this accuracy level across ϵ values. The Xception shows more resilience among the first category due to less accuracy drop than the Resnets.

b: OCT DATASET

Fig. 9-d shows around 80% accuracy at $\epsilon = 0$ but drops below 25% at $\epsilon = 0.1$ for all the models and attack combinations (PGD-FGSM vs FGSM-PGD). Although Resnet and Xception try to exhibit some fluctuations for higher performance, overall accuracy indicates that the models failed with simultaneous partial attacks for OCT dataset images regardless of any attack combination order (FGSM-PGD vs PGD-FGSM).

c: SKIN CANCER DATASET

Fig. 10-d shows around 70% accuracy at $\epsilon = 0$ for all the models, and this level is maintained for other ϵ values by all except 3 combinations, including VGG FGSM-PGD, VGG PGD-FGSM, and Resnet PGD-FGSM models. VGG models (FGSM-PGD and PGD-FGSM) drop below 20% at $\epsilon = 0.2$ and continue to decline. The Resnet PGD-FGSM shows a minimum of 40% accuracy at $\epsilon = 0.5$ but is able

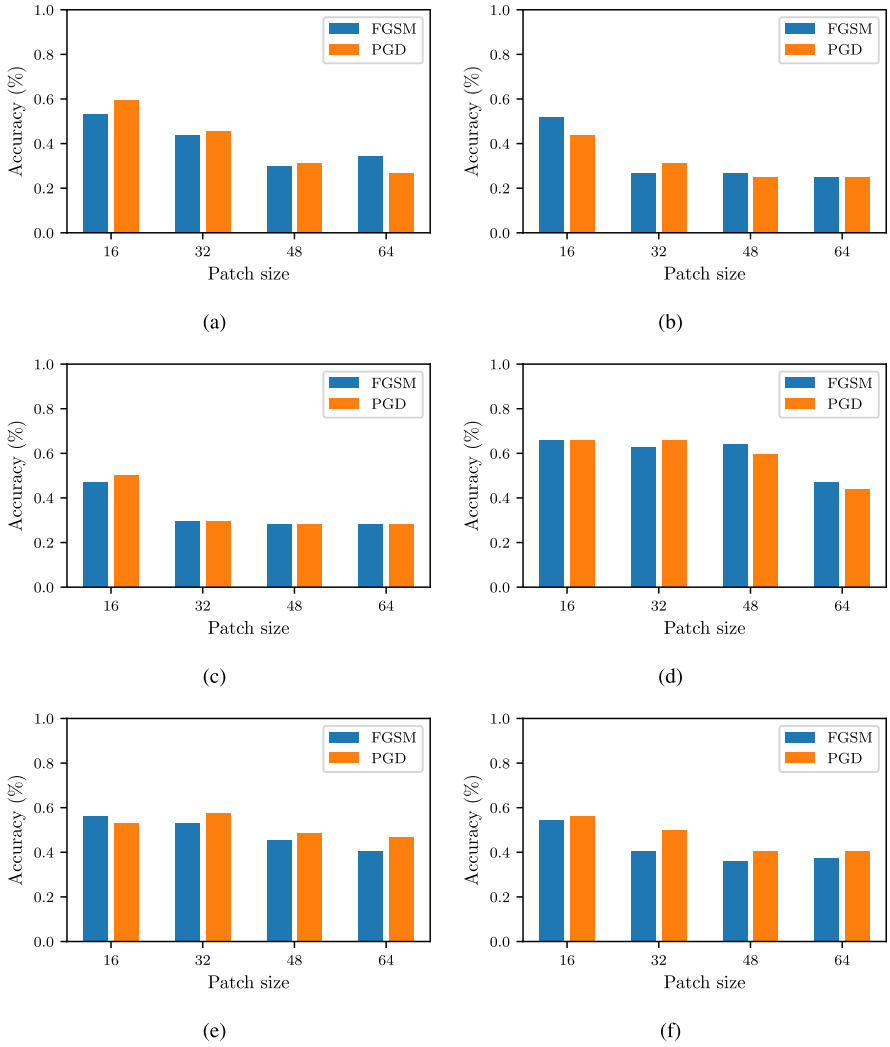

FIGURE 12. Individual patch attack on OCT dataset (a)-(c) central-patch attack, and (d)-(f) top-patch attack, the first column (a and d) represents perturbation of ϵ=0.1, middle column (b and e) ϵ=0.5 and the 3rd column (c and f) ϵ=0.9, using VGG model.

to increase the accuracy after $\epsilon = 0.5$. This indicates the sensitivity of VGG models against the simultaneous partial attacks and Resnet's behavior for simultaneous PGD-FGSM combinations.

C. PATCH ATTACKS

This section describes patch attacks, using the VGG model, in two configurations: i) single patch of 4 sizes (16×16, 32×32, 48×48, and 64×64) either at the center or top position of an image and ii) simultaneous patches of two types of attacks at the center and top of an image (20%-5% and 5%-20% of image portion at center or top position).

1) INDIVIDUAL PATCH ATTACKS

Results of individual patch attacks, using only the VGG model, are described based on three datasets, including Chest X-ray (Fig. 11), OCT (Fig. 12) and Skin cancer (Fig. 13). Each figure is a two-row bar graph representing central-patch (top row) and top-patch (bottom row) attacks with three columns in each row representing ϵ values 0.1, 0.5 and 0.9.

a: CHEST X-RAY DATASET

The accuracy values, in Fig. 11, tend to small reduction with an increase in ϵ and patch size for both the central and top-patches regarding FGSM and PGD attacks respectively. The central-patch attack of both types (blue and orange bars in Fig. 11-a-c, top row) shows accuracy values of around 60% for the smallest 16×16 patch with ϵ 0.1 and a constant accuracy by increasing the patch size. This resistance against higher attacks was found across all epsilon values, with a little variation in accuracy for ϵ 0.5 and ϵ 0.9. The decline is more tangible for the top-patch attack, bottom row, (Fig. 11-d-f) by increasing the patch size and ϵ but results show how the model behaves a little better against PGD attacks in this case.

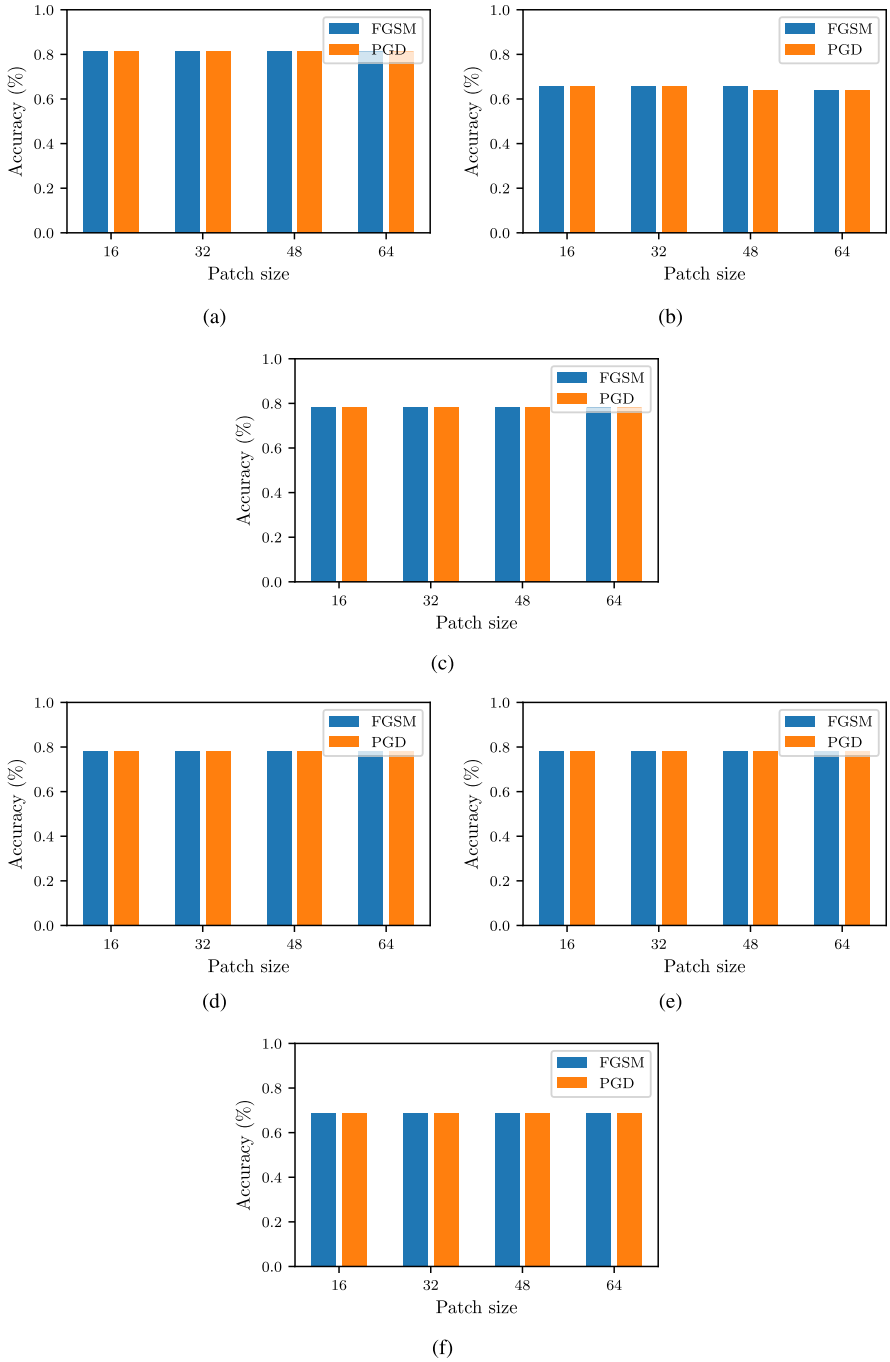

FIGURE 13. Individual patch attack on Skin-cancer dataset (a)-(c) central-patch attack, and (d)-(f) top-patch attack, the first column (a and d) represents perturbation of ϵ=0.1, middle column (b and e) ϵ=0.5 and the 3rd column (c and f) ϵ=0.9, using VGG model.

Overall, the model shows high robustness facing different patch sizes and ϵ specifically at the central position.

b: OCT DATASET

As shown in Fig. 12-a-c (top row), the central-patch attack of the smallest 16 × 16 patch has accuracy around 57% and 62% for FGSM and PGD for ϵ 0.1, which drops for the bigger patches of FGSM and PGD respectively; The top-patch attack accuracy (Fig. 12-d-f) has a less perceptible decreasing trend for ϵ and patch size compared to the central-patch attack.

Regardless of attack types (FGSM and PGD), increasing the ϵ, higher than 0.1, and patch size, bigger than 32, does not impact on the model robustness in the central part. However, the same as the Chest dataset, the model shows better performance against PGD attacks irrespective of the patch positions.

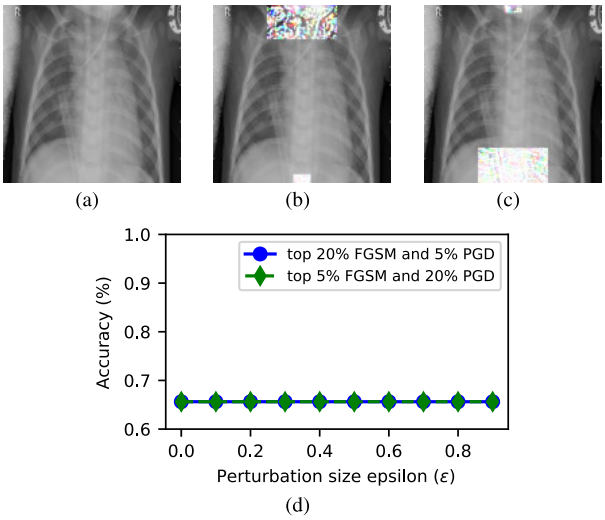

FIGURE 14. Simultaneous-patch attack on Chest X-ray dataset with (a) original image, (b) top 20% for FGSM and bottom 5% for PGD; (c) top 5% for FGSM and bottom 20% for PGD and (d) Accuracy of b and c.

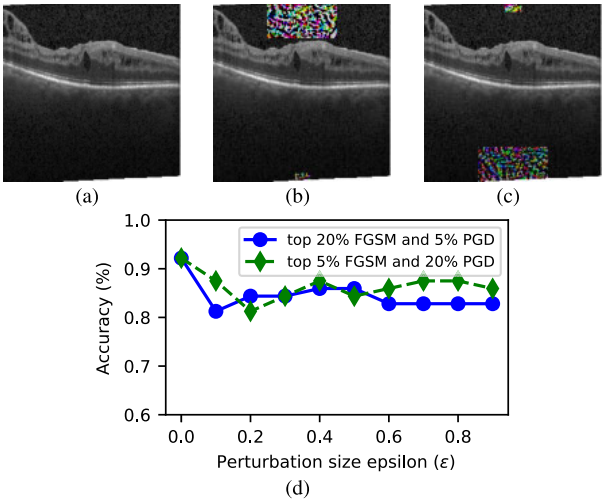

FIGURE 15. Simultaneous-patch attack on OCT dataset with (a) original image, (b) top 20% for FGSM and bottom 5% for PGD; (c) top 5% for FGSM and bottom 20% for PGD and (d) Accuracy of b and c.

c: SKIN CANCER DATASET

Fig. 13 shows that the accuracy values are not affected by the patch size regardless of attack types (FGSM or PGD). But ϵ plays a small role in the decrease or increase of accuracy in both the central (Fig. 13-a-c) and top (Fig. 13-d-f) patches.

2) SIMULTANEOUS PATCH ATTACKS

Two concurrent adversarial patches from both FGSM and PGD attacks of size 5% and 20% (and vice versa) at the center-top and center-bottom positions of each image have been examined. The accuracy variation across ϵ values are shown for datasets including Chest Xray (Fig. 14), OCT (Fig. 15), and Skin cancer (Fig. 16).

a: CHEST X-RAY DATASET

Using the VGG model, Fig. 14-d shows 65% accuracy for both attack combinations of FGSM and PGD with patch size proportions of 20% in the top and 5% in the bottom and vice versa at ϵ 0 (no attack scenario). The accuracy continues to exhibit constant robustness (65% accuracy) across all ϵ values in the range of 0.1-0.9.

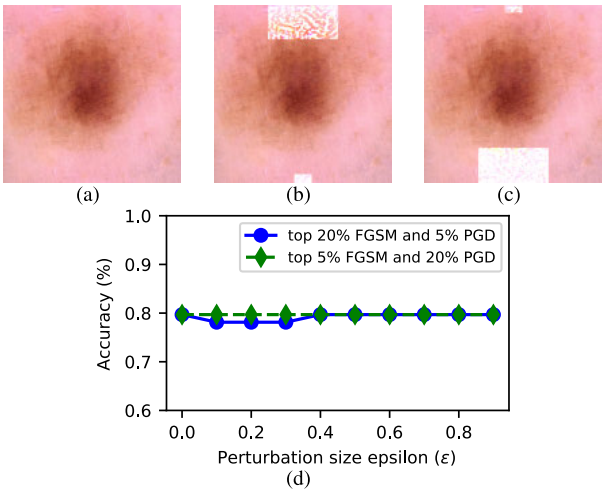

FIGURE 16. Simultaneous-patch attack on Skin-cancer dataset with (a) original image, (b) top 20% for FGSM and bottom 5% for PGD; (c) top 5% for FGSM and bottom 20% for PGD and (d) Accuracy of b and c.

b: OCT DATASET

Fig. 15-d shows an accuracy of more than 90% using the VGG model at first. Both configurations of FGSM(20%)+PGD(5%) (green line) and FGSM(5%)+PGD(20%) (blue line) exhibit sharp drops at ϵ, 0.1 and 0.2 respectively. However, each combination continues with some fluctuations to reach an appropriate accuracy between 80% to 90%, facing the growth of ϵ.

c: SKIN CANCER DATASET

Using the VGG model, Fig. 16-d shows the accuracy of 80% at $\epsilon = 0$ for each combination of FGSM+PGD, and both configurations (green and blue lines) maintain almost the same high accuracy (around 80%) across different ϵ values.

V. DISCUSSION AND MAJOR FINDINGS

This study aims to investigate the effects of a *full*, *partial*, *patch* and *simultaneous* (combinational) behavior of FGSM and PGD perturbations on medical images leveraging four state-of-the-art deep learning models, e.g., VGG, Resnet, Xception and Alexnet. The number of perturbations was controlled by ϵ parameter with 9 values between 0.1 and 0.9 where high ϵ indicates more distortion and a clean image was represented by $\epsilon = 0$.

PGD is recognized as an iterative white-box attack with the expectation of providing a strong negative influence on the model's predictions and accuracy. Still, the results show how our models exhibit robustness against complicated attacks. The PGD adversarial attacks had the lowest negative impact on Chest X-ray dataset accuracy in Fig. 8 using Xception and

Resnet models that are struggling to improve by growing ϵ. The deficiencies of the VGG model are as below:

- Full attack (VGG-PGD, Fig. 8-a, with the lowest accuracy around 35%),
- Top partial attack (VGG-PGD, Fig. 8-b, with the lowest accuracy around 40%), and
- Simultaneous partial attack (VGG-PGD, Fig. 8-d, with the lowest accuracy around 35%).

A similar observation was achieved for the Skin cancer dataset (Fig. 10), where VGG-PGD and Resnet-PGD had difficulty against the full and simultaneous attacks. However, the OCT dataset (Fig. 9) observes completely different bahavior facing PGD and FGSM by decreasing the accuracy of all the models. Even the model with high performance on the chest x-ray dataset shows difficulty in OCT classification. Overall results show that the Xception was found to be comparatively more resilient, whereas the VGG was more susceptible to PGD attack variations across datasets.

Increasing the perturbation degree (ϵ 0.1-0.9) is not a general observation of declining the model's accuracy. Instead, it means there is no direct relationship between the size of the attack (high distortion) and performance. It has been proven by the full, partial and simultaneous partial attacks shown in Fig. 8, Fig. 9, and Fig. 10.

The effects of perturbation degree were also observed for the central and top-patch attacks, which show imperceptible decreasing accuracy for both FGSM and PGD against growing the ϵ and patch size parameters on the Chest X-ray (Fig. 11) and Skin cancer datasets (Fig. 13) except OCT (Fig. 12). Based on the feature map of Fig. 1, in the OCT dataset, the central part is more sensitive than the top due to the included critical features. That causes extra fooling outcomes for the VGG model at the central part. Furthermore, when the patch size increases, it perturbs more image areas and challenges the models' classification accuracy. Therefore, it concluded that big patches do not always negatively affect the models' performance.

The position and size of an adversarial patch attack do not directly impact the model accuracy without being engaged with the important feature areas of an image. It can be explained by looking at the feature map areas of the sample images in (Fig. 1 bottom row and Fig. 2). The more percentage of coverage and overlap by a patch attack causes a reduction in the models' accuracy. The simultaneous patch attack is an appropriate example that can cover the most important areas (even distributed) to fool the model.

Compared to the simultaneous partial attack (half FGSM plus PGD distortion), the simultaneous patch attack (FGSM+PGD) were found to involve the model (VGG) more appropriately. Although adverse results are inherently achieved by the patch combination attacks, the model observes acceptable resilience of accuracy in this case.

The simultaneous patch attacks show a small drop in models' accuracy regardless of the types of attack and perturbation degree across datasets (Chest X-ray in Fig. 14, Skin cancer in Fig. 16), and OCT (in Fig. 15). However, the position of the attack is important in this scenario.

TABLE 4. Comparison of accuracy between existing studies and proposed attack impact on classification accuracy.

Ref.	Datasets	Accuracy	Accuracy with proposed attack
[33]	Skin Cancer	0.83	0.55
[34]	OCT	0.90	0.25
[35]	Chest X-ray	0.83	0.74

In Table 4, we have compared the impact of our proposed attack methods with the most recent research on OCT, skin cancer and chest X-ray datasets. The findings were striking, showcasing a substantial degradation in model performance when subjected to adversarial attacks. These attacks noticeably inflated misclassification rates, underscoring the vulnerability of classification systems to adversarial manipulations.

The noticeable decline in model accuracy emphasises the importance of safeguarding against adversarial manipulations in classification systems, particularly in medical applications. The vulnerability unveiled through these attacks highlights the need for robust and resilient defences to prevent such manipulations. Implementing enhanced defences, possibly incorporating adversarial training or more complex model architectures, becomes essential to ensure the reliability and integrity of classification systems in medical imaging diagnostics. However, defence against adversarial attacks is out of scope in this study. Our goal only highlights the impact of different types of adversarial attacks on classification performances.

This simultaneous patch attack was part of an attempt to attack a model's sensitive decision point or target specific critical features due to limited scope. This study assumed fixed top and bottom-position patches of 20% and 5% areas. The effect of simultaneous patches on different datasets gives the impression that if the sensitive area in an image, at even a single pixel, can be perturbed, it can adversely affect the model's performance.

VI. CONCLUSION AND FUTURE WORK

In this paper, we evaluated the robustness of four state-of-the-art deep learning models on three challenging medical datasets. Full, partial and patch attacks were broadly investigated on those datasets and thoroughly observed by potential weaknesses and powerful perspectives. Specifically, the behaviour of the combination attacks presented and investigated by simultaneous partial and patch use cases. Datasets with single-label and multi-label classification have been analyzed, and the behaviour of the models has been separately visualized in detail. In this way, we found out how any remaining distortion, physically or non-physically, in our images can be enhanced with possible emerged perturbation by the intrusion. Only static top and bottom positions were examined for simultaneous patch attacks, but semantically finding the sensitive locations and areas of important features

would be the aim of the extension of this study with more experiments in future work.

Further, in the future, our research will focus on enhancing the robustness of medical image analysis systems against adversarial attacks. We aim to use advanced techniques, e.g., radial basis mapping kernels, to mitigate the impact of adversarial perturbations in classification tasks. This will improve the security and reliability of AI-powered diagnostic solutions in the medical sector.

REFERENCES

[1] P. Savadjiev, J. Chong, A. Dohan, M. Vakalopoulou, C. Reinhold, N. Paragios, and B. Gallix, "Demystification of AI-driven medical image interpretation: Past, present and future," *Eur. Radiol.*, vol. 29, no. 3, pp. 1616–1624, Mar. 2019.

[2] G. Litjens, T. Kooi, B. Bejnordi, A. Setio, F. Ciompi, M. Ghafoorian, J. Laak, B. Inneken, and C. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

[3] S. Kaviani, K. J. Han, and I. Sohn, "Adversarial attacks and defenses on AI in medical imaging informatics: A survey," *Expert Syst. Appl.*, vol. 198, Jul. 2022, Art. no. 116815.

[4] A. Selvakkumar, S. Pal, and Z. Jadidi, "Addressing adversarial machine learning attacks in smart healthcare perspectives," in *Sensing Technology*. Cham, Switzerland: Springer, 2022, pp. 269–282.

[5] M. N. Al-Andoli, S. C. Tan, K. S. Sim, P. Y. Goh, and C. P. Lim, "A framework for robust deep learning models against adversarial attacks based on a protection layer approach," *IEEE Access*, vol. 12, pp. 17522–17540, 2024.

[6] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018.

[7] G. Bortsova, C. González-Gonzalo, S. C. Wetstein, F. Dubost, I. Katramados, L. Hogeweg, B. Liefers, B. van Ginneken, J. P. W. Pluim, M. Veta, C. I. Sánchez, and M. de Bruijne, "Adversarial attack vulnerability of medical image analysis systems: Unexplored factors," *Med. Image Anal.*, vol. 73, Oct. 2021, Art. no. 102141.

[8] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, and F. Lu, "Understanding adversarial attacks on deep learning based medical image analysis systems," *Pattern Recognit.*, vol. 110, Feb. 2021, Art. no. 107332.

[9] D. Gupta and B. Pal, "Vulnerability analysis and robust training with additive noise for FGSM attack on transfer learning-based brain tumor detection from mri," in *Proc. Int. Conf. Big Data, IoT, Mach. Learn.* Singapore: Springer, 2022, pp. 103–114.

[10] N. Papernot et al., "Technical report on the CleverHans v2.1.0 adversarial examples library," 2016, *arXiv:1610.00768*.

[11] A. Sharma, Y. Bian, P. Munz, and A. Narayan, "Adversarial patch attacks and defences in vision-based tasks: A survey," 2022, *arXiv:2206.08304*.

[12] Y. Zhang, Y. Zhang, J. Qi, K. Bin, H. Wen, and P. Zhong, "Adversarial patch attack on multi-scale object detection for remote sensing image," *Remote Sens.*, vol. 14, no. 21, p. 5298, 2022.

[13] H. Li and Y. Zhao, "Fool object detectors with L_0-norm patch attack," in *Proc. CEUR Workshop*, 2020, pp. 1–12.

[14] N. G. Laleh, D. Truhn, G. P. Veldhuizen, T. Han, M. van Treeck, R. D. Buelow, R. Langer, B. Dislich, P. Boor, V. Schulz, and J. N. Kather, "Adversarial attacks and adversarial robustness in computational pathology," *Nature Commun.*, vol. 13, no. 1, p. 5711, Sep. 2022.

[15] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," 2017, *arXiv:1712.09665*.

[16] Z. Cheng, J. Liang, H. Choi, G. Tao, Z. Cao, D. Liu, and X. Zhang, "Physical attack on monocular depth estimation with optimal adversarial patches," in *Proc. 17th Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 514–532.

[17] A. Liu, J. Wang, X. Liu, B. Cao, C. Zhang, and H. Yu, "Bias-based universal adversarial patch attack for automatic check-out," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 395–410.

[18] A. Braunegg, A. Chakraborty, M. Krumdick, N. Lape, S. Leary, K. Manville, E. Merkhofer, L. Strickhart, and M. Walmer, "APRICOT: A dataset of physical adversarial attacks on object detection," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., 2020, pp. 35–50.

[19] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, Oct. 2019.

[20] X. Li and S. Ji, "Generative dynamic patch attack," 2021, *arXiv:2111.04266*.

[21] J. Mohapatra, T.-W. Weng, P.-Y. Chen, S. Liu, and L. Daniel, "Towards verifying robustness of neural networks against a family of semantic perturbations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 241–249.

[22] M. Xu, T. Zhang, Z. Li, M. Liu, and D. Zhang, "Towards evaluating the robustness of deep diagnostic models by adversarial attack," *Med. Image Anal.*, vol. 69, Apr. 2021, Art. no. 101977.

[23] T. V. Maliamanis, K. D. Apostolidis, and G. A. Papakostas, "How resilient are deep learning models in medical image analysis? The case of the moment-based adversarial attack (Mb-AdA)," *Biomedicines*, vol. 10, no. 10, p. 2545, Oct. 2022.

[24] M. A. Hoque, S. Haque, S. K. Debnath, and M. Ahiduzzaman, "Investigating the robustness of deep neural network based COVID-19 detection models against universal adversarial attacks," in *Proc. 3rd Int. Conf. Sustain. Technol. Ind. 4.0 (STI)*, Dec. 2021, pp. 1–6.

[25] A. Minagi, H. Hirano, and K. Takemoto, "Natural images allow universal adversarial attacks on medical image classification using deep neural networks with transfer learning," *J. Imag.*, vol. 8, no. 2, p. 38, Feb. 2022.

[26] B. Stimpel, C. Syben, F. Schirrmacher, P. Hoelter, A. Dörfler, and A. Maier, "Multi-modal deep guided filtering for comprehensible medical image processing," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1703–1711, May 2020.

[27] *Chest X-ray Images (Pneumonia) Kaggle*. Accessed: Jun. 15, 2023. [Online]. Available: https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia

[28] Kaggle. *Retinal Oct Images (Optical Coherence Tomography)*. Accessed: Jun. 15, 2023. [Online]. Available: https://www.kaggle.com/datasets/paultimothymooney/kermany2018

[29] *Skin Cancer MNIST: HAM10000 | Kaggle*. Accessed: Jun. 15, 2023. [Online]. Available: https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000

[30] P. Naveen and B. Diwan, "Pre-trained VGG-16 with CNN architecture to classify X-Rays images into normal or pneumonia," in *Proc. Int. Conf. Emerg. Smart Comput. Informat. (ESCI)*, Mar. 2021, pp. 102–105.

[31] A. Gulli and S. Pal, *Deep Learning With Keras*. Birmingham, U.K.: Packt Publishing, 2017.

[32] S. Kumar and A. Mallik, "COVID-19 detection from chest X-rays using trained output based transfer learning approach," *Neural Process. Lett.*, vol. 55, no. 3, pp. 2405–2428, Jun. 2023.

[33] D. Keerthana, V. Venugopal, M. K. Nath, and M. Mishra, "Hybrid convolutional neural networks with SVM classifier for classification of skin cancer," *Biomed. Eng. Adv.*, vol. 5, Jun. 2023, Art. no. 100069.

[34] K. Karthik and M. Mahadevappa, "Convolution neural networks for optical coherence tomography (OCT) image classification," *Biomed. Signal Process. Control*, vol. 79, Jan. 2023, Art. no. 104176.

[35] T. Agrawal and P. Choudhary, "ALCNN: Attention based lightweight convolutional neural network for pneumothorax detection in chest X-rays," *Biomed. Signal Process. Control*, vol. 79, Jan. 2023, Art. no. 104126.

SHANTANU PAL (Senior Member, IEEE) is associated with the School of Information Technology, Deakin University, Melbourne, Australia. He has extensive research experience in the Internet of Things, big data and distributed applications, access control, trust management, blockchain technology, mobile and cloud computing, and machine learning. He has several publications in highly ranked conferences and journals, including IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, IEEE TRANSACTIONS ON SERVICES COMPUTING, and IEEE INTERNET OF THINGS JOURNAL.

SAIFUR RAHMAN (Member, IEEE) received the Master of Science degree in electrical, electronics, and control engineering from Kongju National University, Republic of Korea, and the Ph.D. degree from Deakin University, Melbourne, Australia. He is an Associate Research Fellow with Deakin University. His research interests span a diverse array of biomedical engineering and technology topics, including biomedical signal processing, wearable device designing, machine learning, and deep learning.

MAEDEH BEHESHTI received the Ph.D. degree from Griffith University, Australia. She is associated with the Critical Path Institute (C-Path), Tucson, AZ, USA. Her research is in computer vision, medical image analysis, probabilistic graphical models, nature-inspired algorithms, and machine learning. She is interested in developing algorithms for image feature extraction, retrieval for scenery, and medical image applications through semantic and ontology-based approaches.

AHSAN HABIB (Member, IEEE) received the M.Eng. degree in information and communications technologies from the Asian Institute of Technology, Thailand, and the Ph.D. degree from Deakin University, Melbourne, Australia. He is a Lecturer with the School of Information Technology, Deakin University. He has several publications in highly ranked conferences and journals, including IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING and IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS. His research interests include biomedical signal processing and modeling, time series analysis, machine learning, and deep learning.

ZAHRA JADIDI received the Ph.D. degree in network security from Griffith University, Australia. She is a Lecturer with the School of Information and Communication Technology, Griffith University. She was a Postdoctoral Research Fellow in cybersecurity with Queensland University of Technology, Brisbane, Australia, for two and half years. In this position, she was with several industry partners in the security of cyber-physical systems. She also worked as a Research Fellow with Griffith University and participated in various projects on artificial intelligence-based analysis of data received from remote sensors. Her research interests are cybersecurity, security of cyber physical systems, machine learning applications in security analysis, security of machine learning algorithms, and proactive security.

CHANDAN KARMAKAR (Member, IEEE) received the Ph.D. degree from The University of Melbourne, Australia. He is an Associate Professor with the School of Information Technology, Deakin University, Melbourne, Australia. He has published one book and more than 130 research articles. His research interests include biomedical devices and signal processing, cardiovascular and neural systems related to sleep-disordered breathing, and diabetic autonomic neuropathy.