applied sciences

MDPI

Article

Vulnerability in Deep Transfer Learning Models to Adversarial Fast Gradient Sign Attack for COVID-19 Prediction from Chest Radiography Images

Biprodip Pal [1], Debashis Gupta [1], Md. Rashed-Al-Mahfuz [2], Salem A. Alyami [3] and Mohammad Ali Moni [4,5,*]

[1] Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi 6204, Bangladesh; biprodip@cse.ruet.ac.bd (B.P.); debashisguptaruet@gmail.com (D.G.)
[2] Department of Computer Science and Engineering, University of Rajshahi, Rajshahi 6205, Bangladesh; ram@ru.ac.bd
[3] Department of Mathematics and Statistics, Imam Mohammad Ibn Saud Islamic University, Riyadh 13318, Saudi Arabia; saalyami@imamu.edu.sa
[4] WHO Collaborating Centre on eHealth, School of Public Health and Community Medicine UNSW Sydney, Sydney, NSW 2052, Australia
[5] Healthy Ageing Theme, The Garvan Institute of Medical Research, Darlinghurst, NSW 2010, Australia
* Correspondence: m.moni@unsw.edu.au

Abstract: The COVID-19 pandemic requires the rapid isolation of infected patients. Thus, high-sensitivity radiology images could be a key technique to diagnose patients besides the polymerase chain reaction approach. Deep learning algorithms are proposed in several studies to detect COVID-19 symptoms due to the success in chest radiography image classification, cost efficiency, lack of expert radiologists, and the need for faster processing in the pandemic area. Most of the promising algorithms proposed in different studies are based on pre-trained deep learning models. Such open-source models and lack of variation in the radiology image-capturing environment make the diagnosis system vulnerable to adversarial attacks such as fast gradient sign method (FGSM) attack. This study therefore explored the potential vulnerability of pre-trained convolutional neural network algorithms to the FGSM attack in terms of two frequently used models, VGG16 and Inception-v3. Firstly, we developed two transfer learning models for X-ray and CT image-based COVID-19 classification and analyzed the performance extensively in terms of accuracy, precision, recall, and AUC. Secondly, our study illustrates that misclassification can occur with a very minor perturbation magnitude, such as 0.009 and 0.003 for the FGSM attack in these models for X-ray and CT images, respectively, without any effect on the visual perceptibility of the perturbation. In addition, we demonstrated that successful FGSM attack can decrease the classification performance to 16.67% and 55.56% for X-ray images, as well as 36% and 40% in the case of CT images for VGG16 and Inception-v3, respectively, without any human-recognizable perturbation effects in the adversarial images. Finally, we analyzed that correct class probability of any test image which is supposed to be 1, can drop for both considered models and with increased perturbation; it can drop to 0.24 and 0.17 for the VGG16 model in cases of X-ray and CT images, respectively. Thus, despite the need for data sharing and automated diagnosis, practical deployment of such program requires more robustness.

Keywords: COVID-19; deep learning; adversarial attack; FGSM attack; radiology images

check for updates

Citation: Pal, B.; Gupta, D.; Rashed-Al-Mahfuz, M.; Alyami, S.A.; Moni, M.A. Vulnerability in Deep Transfer Learning Models to Adversarial Fast Gradient Sign Attack for COVID-19 Prediction from Chest Radiography Images. *Appl. Sci.* **2021**, *11*, 4233. https://doi.org/10.3390/app11094233

Academic Editor: Keun Ho Ryu

Received: 20 March 2021
Accepted: 15 April 2021
Published: 7 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The COVID-19 pandemic has had a devastating influence on the well-being and health of the population worldwide, by the infection by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2). Effective screening of infected patients is a vital step in the fight against COVID-19. Therefore, infected people can receive treatment immediately and can be isolated to reduce virus spread. The polymerase chain reaction

(PCR) method is the gold standard approach used to detect COVID-19 cases by detecting SARS-CoV-2 RNA from collected samples or through pharyngeal or nasopharyngeal swabs [1]. Radiography examination, another screening method for COVID-19, conducts chest radiography imaging such as X-ray or computed tomography (CT) imaging, and radiologists can analyze them for visual signs of SARS-CoV-2 viral infection. PCR testing has high sensitivity, but it is a laborious, complicated, costly, and very time-consuming manual process. Strict requirements for laboratory environment and limited supply delay the precise diagnosis of suspected patients, posing challenges to preventing the spread of the infection, mostly at the epidemic zone. On the other hand, the radiography examination is faster and more widely available [2]. Thus, a chest X-ray or CT imaging can be performed for a patient. If radiography and clinical situations are normal, patients can go home and wait for the etiological test results. However, patients may be admitted to the hospital if the X-ray film shows pathological conditions. Thus, radiography examination is very useful approach over the PCR testing and comes out with higher sensitivity in some cases [3].

Several recent reports have emphasized chest CT as a key component of the diagnostic procedure for suspected COVID-19-affected patients [4–6]. Abnormalities in chest radiography images exist for COVID-19-infected people [4,7]; therefore, it is a vital tool in epidemic areas for COVID-19 screening [8]. The visual indicators can be elusive; therefore, the lack of expert radiologists is a bottleneck to interpret the radiography images. Medical diagnosis uses computer vision algorithms without the input of a human clinician in many countries [9], and deep learning (DL) has been used successfully with remarkable performance for the automatic diagnostics of diseases, including lung diseases [10,11]. Computer-aided diagnostic systems can support radiologists in the more accurate and faster interpretation of radiography images for COVID-19 detection. Many proposed deep learning-based artificial intelligence (AI) systems have presented promising accuracy to detect COVID-19 symptoms in radiography imaging [12,13]. The Pre-trained Inception model was utilized after fine-tuning followed by a fully connected network to classify viral pneumonia and COVID-19 [14]. Along the same path, Narin et al. identified COVID-19 from lung CT images and chest X-ray images [15]. They also considered the transfer learning-based Inception-v3 model. Pre-trained VGG16 was also used frequently, along with other available pre-trained models in a transfer learning setting to detect COVID-19 from radiology images in several recent studies [16–18]. Due to the better performance and common use throughout several relevant studies, we applied VGG16 and Inception-v3 as representatives of popular transfer learning pre-trained models.

In parallel with the progress of medical DL, adversarial examples have uncovered vulnerabilities in many state-of-the-art DL systems [19]. The need for automated diagnosis makes the process vulnerable to adversarial attack. Rare disease image sharing to build big data repositories for COVID-19, sharing of pre-trained model parameters and intruder access to diagnosis, and network-based diagnosis systems can create the attack. Adversarial examples intentionally craft machine learning model inputs to force the model to generate an incorrect diagnostic result. Adversarial examples typically tend to attempt to reduce the prediction confidence of the target model, changing the output of classification of some sample to any different class from the original class. Deployment of deep learning models for COVID-19 diagnosis is also vulnerable to adversarial examples. These radiology images are captured with well-established and pre-defined exposure and positioning, making adversarial attacks comparatively easier than other computer vision applications [20]. Moreover, most of the successful deep learning methods consisted of the same set of pre-trained ImageNet models and a lack of architectural diversity. Due to research transparency, these models are also available publicly. Additionally, data are often shared among institutions to generate a big data repository for rare diseases such as COVID-19. These reasons require extensive research on possible attacks and robust training approaches for these models. Cutting-edge techniques of adversarial attacks such as FGSM attack use optimization principles to generate small perturbations to fool any target model. Apart from exploring the limitations of current DL methods, this research received attention because of the security threats for deploying these diagnostic DL algorithms in both physical and virtual settings [21–24].

In this paper, we demonstrated the state-of-the-art DL models used in a transfer learning setting to classify COVID-19 samples that are vulnerable to adversarial attacks. We crafted the FGSM attack for DL-based transfer learning algorithms that are commonly used in chest radiology classification and CT imaging to detect COVID-19. We studies the adversarial perturbation variation effect on the visual perceptibility as well as attack performance. Extensive experiments were conducted to analyze potential vulnerability in terms of degradation of the correct class probability score for correct classification and quantifying misclassifications because of FGSM attack. We validated these findings using publicly available COVID-19 patient data.

2. Materials and Methods

In this study, we used chest X-ray and CT images of different publicly available respiratory syndromes including COVID-19-infected patients, and we have applied our models to these datasets. We have briefly discussed our applied DL models and attack design for radiology image classification.

2.1. Dataset Description

This dataset is a collection of radiology images of COVID-19 cases with chest X-ray and CT images. It comprises COVID-19 cases as well as some other respiratory syndromes [25]. This dataset is publicly available and contains 100 COVID-19 images of frontal view X-rays and prognostic data resource for research. According to the dataset reference, a senior radiologist from Tongji Hospital, Wuhan, China, who is experienced in the diagnosis and treatment of a large number of COVID-19 patients, confirmed the utility of this dataset.

Table 1 summarizes the collection of this dataset. Another dataset we used on CT images consisted of labeled CT scan images (746 images) and is used frequently to develop COVID-19 detection models. This COVID-CT-Dataset [26] had 349 CT images containing clinical findings of COVID-19 from 216 patients.

Table 1. Image distribution in dataset.

	COVID	**Non-COVID**	**Total**
Chest X-ray Images	141	127	268
Chest CT Images	349	397	746

2.2. Deep Transfer Learning for Radiology Images

All of the state-of-the-art deep learning-based COVID-19 detection algorithms are based on the concept of transfer learning. Transferable knowledge in the form of expressive features is extracted from the source domain by feature learning algorithms. The source domain data can be denoted as $D_S = \{(x_{S1}; y_{S1}), (x_{Sn}; y_{Sn})\}$, where $x_{Si} \in X_S$ is the data instance, and the consequent class label is $y_{Si} \in Y_S$. Likewise, the target-domain data is denoted as D_T $\{(x_{T1}; y_{T1}), (x_{Tn}; y_{Tn})\}$, where the input $x_{Ti} \in X_T$ and the corresponding output is $y_{Ti} \in Y_T$; in most cases, $0 < n_T << n_S$. Given a learning task T_S from source domain D_S and learning task T_T at a target domain D_T, transfer learning aims to develop the learning of the objective predictive function $f_T(.)$ in D_T using the knowledge in D_S and T_S, where $D_S \neq D_T$, or $T_S \neq T_T$ [27].

Convolutional neural networks (CNNs) are widely used to classify radiology images. CNNs are made of three major types of layers: A convolutional layer, consisting of a learnable kernel and three hyperparameters—depth, stride and setting zero padding. For an input image X and a filter f, the convolution operation $Z = X * f$; The pooling layer decreases the dimensionality of the representations and fully connected layers for input X, weight W, and bias b; FC first computes a linear transformation on the data, followed by some non-linear activation f_a to capture the complex relationships $Z = f_a\ (W^T.X + b)$. All the parameters are adjusted through different variations of gradient descent optimization technique [28]. CNN learning the central concept behind deep learning tactics is the automated discovery of abstraction.

2.3. Transfer Learning from VGG16 CNN Model

We adopted the frequently used VGG16 CNN architecture [29] for classification and attack generation. This model has two parts, feature extractor and classifier. In the feature extractor part, there is a stack of convolutional layers which uses filters with a small receptive field of 3 × 3, and also use 1 × 1 convolutional filters where the convolutional stride is fixed as 1. The feature extractor part of VGG16 is used to extract the feature of the input images. The input radiology image dimension is kept as (224 × 224 × 3) and it is passed through a stack of convolutional layers of the VGG16 pre-trained model with corresponding Imagenet dataset-based pretrained weights. For the classifier part, an average pooling 2D layer is concatenated with the last output layer of feature extractor to reduce overfitting the model by reducing its parameter; it is then followed by a fully connected layer of 64 nodes having an ReLU activation function. Afterwards, a dropout layer of value 0.5 is used to reduce overfitting and the final output layer consists of a softmax function which leads the classification. "Binary_crossentropy" is used as a loss function, and "adam" is used as an optimizer to minimize the loss function. Per epoch, the decay rate is reduced from the learning rate's initial value of $1 \times e^{-3}$. The transfer learning architecture is depicted in Figure 1.

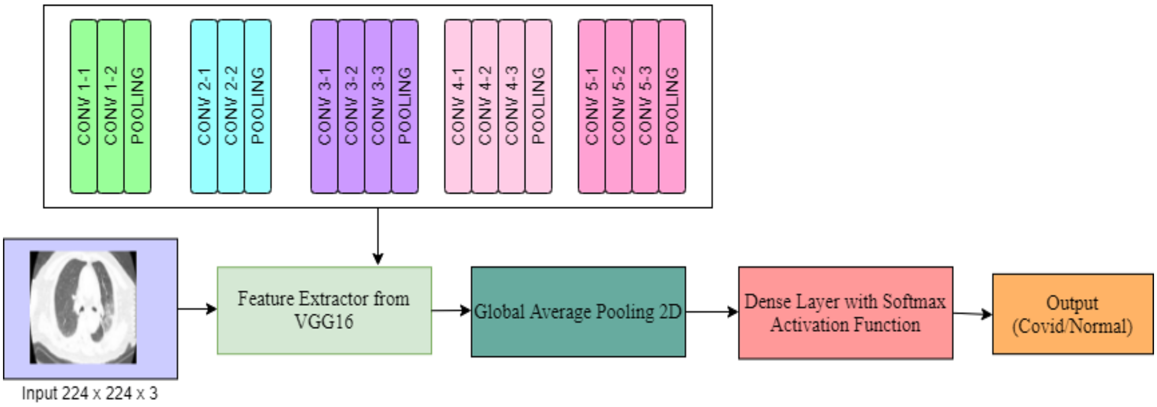


Figure 1. Transfer learning architecture based on VGG16.

2.4. Transfer Learning from Inception-v3 CNN Model

Another model that we used is Inception-v3 [30] for classification and attack generation. It has been used to generate many state-of-the-art radiology images in classification tasks. Inception-v3 is a successor to Inception-v1. The architecture contains a repeating Inception Block with parameter hyper-tuning facility. Several convolutional filters ranging from 1 × 1 to 7 × 7 extract features from the input with no local information loss. Similar to the previous model, the input to the Inception-v3 model is also a 224 × 224 × 3-sized image. The final model is depicted in Figure 2 and consists of this feature extraction part of Inception-v3 and a full stack of convolutional layers concatenated with the first one.

All the layers except FC of this model were kept frozen. The dense layer contained 1024 nodes and the ReLU activation function. We used a dropout of 0.4 to reduce the parameters to avoid overfitting the model. Finally, the output layer contained a softmax function that resulted in the classification. "Binary_crossentropy" was used as a loss function and "Adam" was used as an optimizer to minimize the loss function. In every epoch, the decay rate was reduced from the learning rate's initial value from 10 to 3. For transfer learning, because the new dataset was small but different from the original dataset, we prepared the feature extractor and trained a linear classifier in the FC layer.

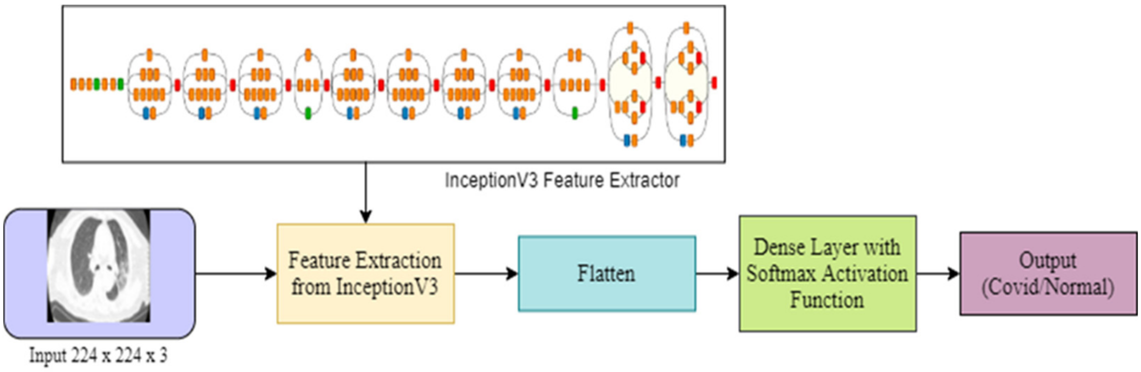

InceptionV3 Feature Extractor
Input 224 x 224 x 3
Feature Extraction from InceptionV3
Flatten
Dense Layer with Softmax Activation Function
Output (Covid/Normal)

Figure 2. Transfer learning architecture based on Inception-v3.

2.5. *Adversarial Attack*

An adversarial attack embodies subtle changing of an original image in such a way that the changes are almost imperceptible to the human eye. Hence, the modified image is named an adversarial image that is misclassified by the classifier. Adversarial noise can significantly affect the robustness of deep neural networks for a wide range of image classification applications. There are two types of adversarial attacks: in-distribution (IND) adversarial attacks, and out-of-distribution (OOD) adversarial attacks [31]. While IND adversarial attacks have extensively been studied including for wide range of applications, this paper demonstrates that attacks such as FGSM are sufficient to degrade the performance of reliable DL models [32,33].

Fast Gradient Sign Attack

In a sentence, the fast gradient sign method works by using the gradients of the neural network to create an adversarial example. Ian Goodfellow et al. (2014) first invented the fast gradient sign method for producing the adversarial images [19]. The gradient sign method applies the gradient of the underlying model to generate the adversarial examples, according to Equation (1):

$$x' = x + \varepsilon \cdot sign(\bigtriangledown x \; J(\theta, x, y)) \tag{1}$$

The original image is x, the original class of x is y, and θ is the model parameter vector. Here, $J(\theta,\ x,\ y)$ is the loss function used to train the network. First, the gradient of the loss function according to the input pixels is calculated. The $\bigtriangledown$ operator is one of the mathematical ways of taking the derivatives of a function regarding different parameters of the model. Hence, $\bigtriangledown x \; J(\theta, x, y)$ is the gradient vector from where the sign of it is taken. The sign of the gradient can be positive or negative depending on the loss function. The positive sign denotes that an increase in pixel intensity increases the loss, i.e., the error that the model makes, and the negative sign represents a decrease in pixel intensity which increases the loss. This vulnerability occurs when the model linearly deals with a relationship between an input pixel intensity and the class score. The process is depicted in Figure 3.

The $\varepsilon \cdot sign(\bigtriangledown x \; J(\theta, x, y))$ stands for a multiplication of a very small epsilon value ε with the signed value obtained from the gradient vector. Then, to create the adversarial images X', the result of the multiplication is simply added to the original image X.

$$x\prime = x + \eta \tag{2}$$

where η denotes $\varepsilon \cdot sign(\bigtriangledown x \; J(\theta, x, y))$.

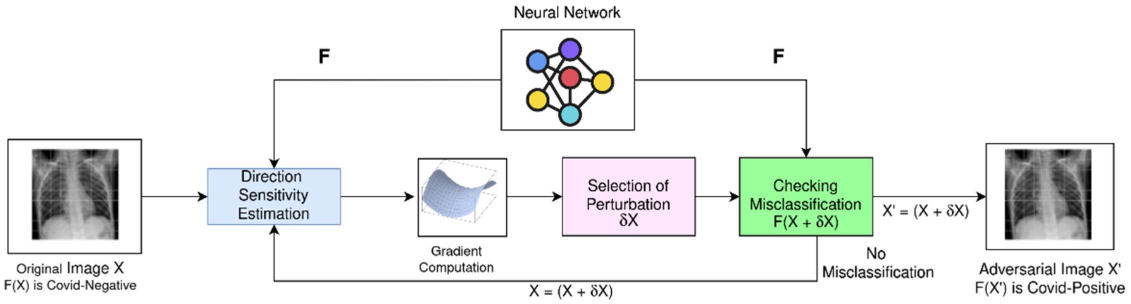


Figure 3. Adversarial image generation approach using FGSM attack.

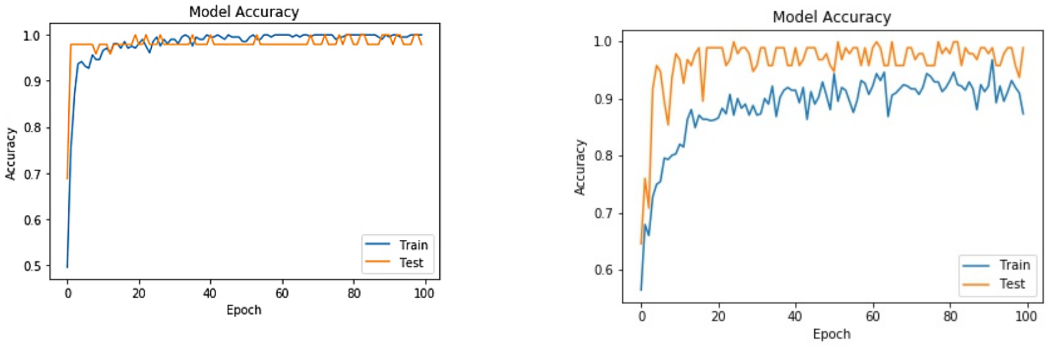
Therefore, varying the value of epsilon ε, usually from 0 to 1, produces different adversarial examples. These examples are mostly imperceptible to the human eye [34].

3. Results

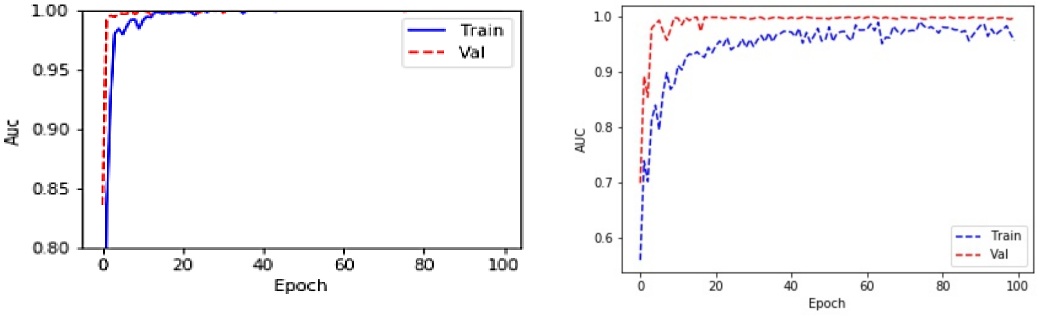
The Tensorflow deep learning library and python programming language were used to implement the code of DL models and FGSM attack. We experimented with four different approaches. To start with, we analyzed the performance of the VGG16 algorithm for COVID-19 classification from X-rays by using transfer learning followed by in-depth analysis of the drop of performance of this model as it suffers from FGSM attack. Later, we analyzed the performance degradation of VGG16 and the Inception-V3 algorithm for COVID-19 classification from X-ray and CT images.

3.1. Transfer Learning to Diagnose COVID-19 from Chest X-ray

To understand the performance drop and vulnerability of VGG16 and Inception-v3 pre-trained DL models for COVID-19 detection, we first analyzed the performances of these models in an attack-free environment. We resized the images to 224 × 224 × 3 and fed them into the DL architecture. An 80:20 split was used to divide the images into training and test sets for chest X-ray images. The total number of training images was small enough; therefore, the training performance saturated quickly, as shown in Figure 4 through the training and test accuracy.



(**a**) Accuracy of VGG16. (**b**) Accuracy of Inception-v3.



(**c**) AUC of VGG16. (**d**) AUC of Inception-v3.

Figure 4. Accuracy and AUC of transfer learning models for COVID-19 detection from chest X-rays.

Tables 2 and 3 enlist the performance details of the VGG16 and Inception-v3 algorithm, respectively, using a confusion matrix. Figures 4 and 5 show the corresponding accuracy, precision, recall, F1 score and AUC. The VGG16 model classifies COVID-19 models with high precision, recall, and F1 of 0.97, 1 and 0.98, respectively. Inception-v3 also came out with a similar accuracy. In the AUC curve, validation data were the same as the test data because we had very little data to train and test. From Figure 4, it can be seen that the AUC for the performance is either equal to 1 or close to 1 during the best performance for VGG16 and Inception-v3. Thus, the model is found to be reliable in absence of FGSM attack to detect COVID-19-infected people.

Table 2. Confusion matrix for performance of the VGG16 model on chest X-rays.

	Actual COVID-19	**Actual Normal**
Predicted COVID-19	28	1
Predicted Normal	0	25

Table 3. Confusion matrix for performance of the Inception-v3 model on chest X-rays.

	Actual COVID-19	**Actual Normal**
Predicted COVID-19	28	0
Predicted Normal	0	25

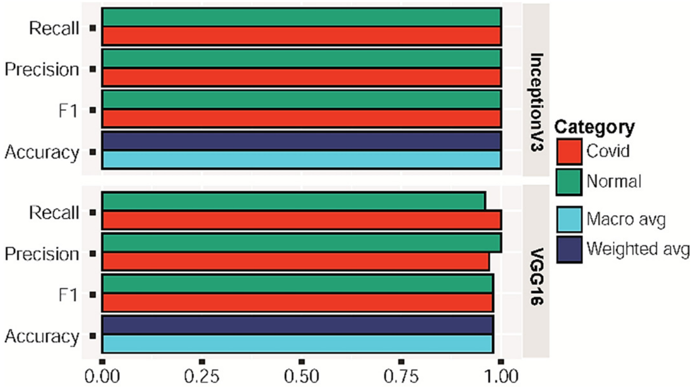
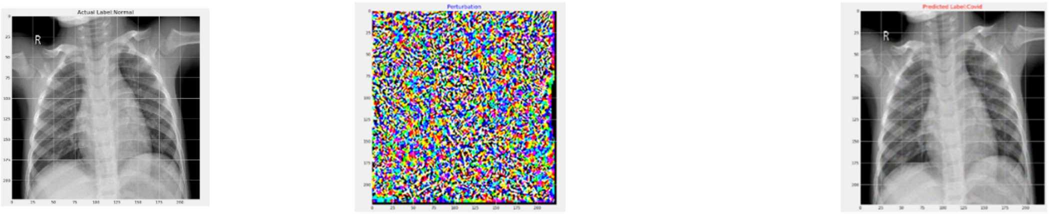


Figure 5. Precision, recall, F score, and accuracy for VGG16 and Inception-v3 on chest X-rays.

3.2. *FGSM Attack Analysis for Chest X-ray*

After developing the transfer learning-based models to classify COVID-19 samples, we applied the FGSM attack on the developed models. For the FGSM attack, we focused on the perturbation degree and corresponding perceptibility effect on chest X-ray images, to see whether subtle perturbation could create an adversarial image that can fool a human radiologist as well as a computer.

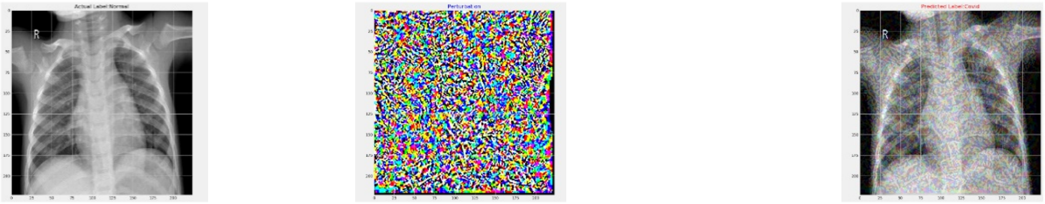
To illustrate the potential risk and performance drop due to the FGSM attack on promising transfer learning models for COVID-19 detection, we experimented by varying the amount of perturbation (ε) in the training images. In Figure 6, the left column figures are original images, and the right-most column figures are corresponding adversarial images generated by FGSM attack. Figure 6c,i clearly depict that misclassification can occur with a very small perturbation and for both considered models. ε of 0.009 successfully generated an adversarial image due to the FGSM attack, which is not recognizable by the human eye. For ease of discussion, we can define such perturbation as safe perturbation magnitude for the attacker. On the other hand, perturbation of 0.08 generated adversarial images that could be distinguished from the original images by the human eye, as seen in Figure 6f,l.



(**a**) Original Image
Prediction: COVID-19-negative.

(**b**) Perturbation
ε = 0.009.

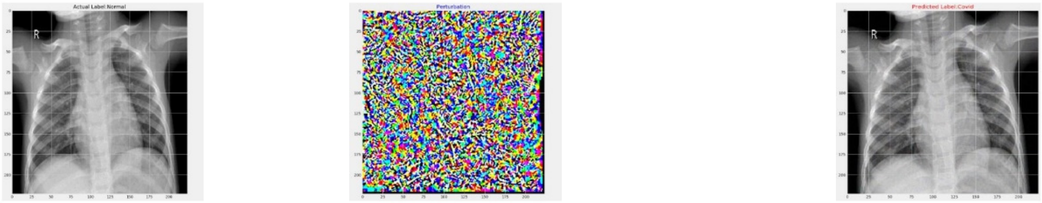
(**c**) Adversarial image, with human-imperceptible perturbation.
Prediction: COVID-19-positive by VGG16.



(**d**) Original Image
Prediction: COVID-19-negative.

(**e**) Perturbation
ε = 0.08.

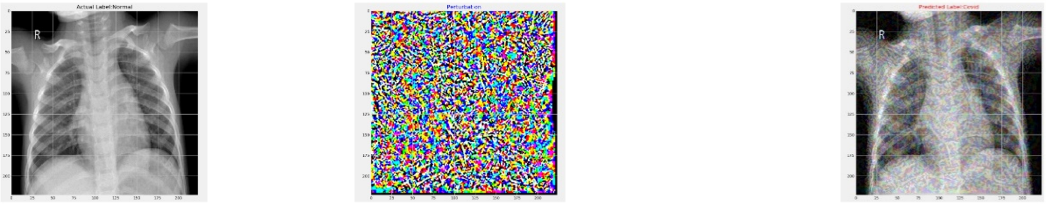
(**f**) Adversarial image, with human-perceptible perturbation.
Prediction: COVID-19-positive by VGG16.



(**g**) Original Image
Prediction: COVID-19-negative.

(**h**) Perturbation
ε = 0.009.

(**i**) Adversarial image with human-imperceptible perturbation.
Prediction: COVID-19-positive by Inception-v3.



(**j**) Original Image
Prediction: COVID-19-negative.

(**k**) Perturbation
ε = 0.08.

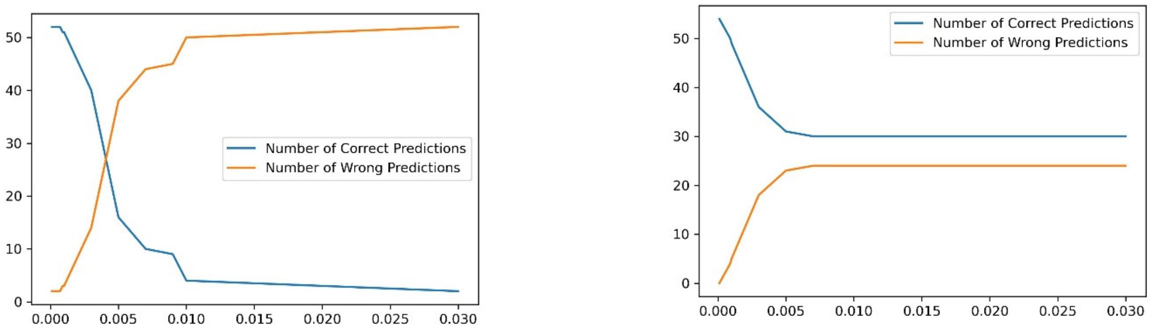
(**l**) Adversarial image with human-perceptible perturbation.
Prediction: COVID-19-positive by Inception-v3.

Figure 6. Original image (**left column**), perturbation (**middle column**), and corresponding misclassified adversarial chest X-ray images generated using FGSM attack for VGG16 and Inception-v3.

Table 4 and Figure 7 clarify, in detail, that as the ε increases, the number of incorrect predictions increases for the considered representative transfer learning models. It is illustrated that very small perturbation of the FGSM attack is sufficient to cause a catastrophic drop in diagnostic performance, while the adversarial images are safe to see in the human eye. Table 4 and Figure 7 elucidate that for a safe perturbation magnitude such as 0.009, the performance drops significantly to almost 16% for VGG16 and 55% for Inception-v3, making these models unusable for COVID-19 detection purpose as long as no protective screening or robust training is ensured. Figure 6f,l also shows that with increasing ε, the noise in adversarial images becomes recognizable by the human eye and the misclassification continues to occur for the mentioned model for these images. Experiments suggest that at higher noise magnitudes, the performance fall was caused by the image corruption from noise, although to a very small extent. Both human experts as well as a computer can be fooled through detecting the noise. Thus, the FGSM attack shows the vulnerability of state-of-the-art pre-trained DL COVID-19 detection models that were successfully classifying COVID-19 samples. Some medical images have significantly high attention regions. Rich biological textures in medical images often distract deep learning models to pay proper attention into the areas that are not important for the diagnosis. Subtle perturbations in these regions results in significant changes in model prediction.

Table 4. Diagnostic performance drop for different ε of FGSM attack in chest X-ray images.

ε	VGG16 with Adversarial Image			Inception-v3 with Adversarial Image		
	Total Correct	**Total Incorrect**	**ACC (%)**	**Total Correct**	**Total Incorrect**	**ACC (%)**
0.0001	52	2	96.29	54	0	100
0.0003	52	2	96.29	53	1	98.15
0.0005	52	2	96.29	52	2	96.30
0.0007	52	2	96.29	51	3	94.44
0.0009	51	3	94.44	50	4	92.59
0.001	51	3	94.44	49	5	90.74
0.003	40	14	74.07	36	18	66.67
0.005	16	38	29.63	31	23	57.41
0.007	10	44	18.52	30	24	55.56
0.009	9	45	16.67	30	24	55.56
0.01	4	50	7.41	30	24	55.56

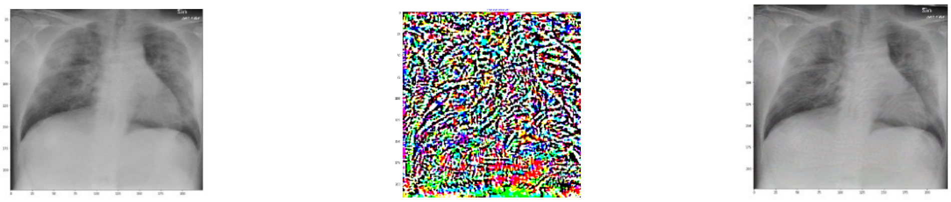


(**a**) VGG16. (**b**) Inception-v3.

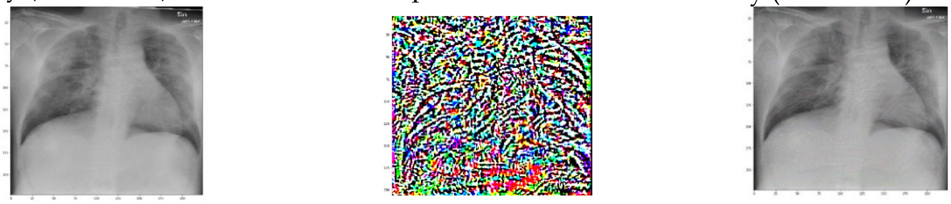
Figure 7. Performance for different ε of FGSM attack in chest X-ray images.

Finally, we investigated the drop of class probability score for images belonging to the correctly predicted class. The deep transfer learning approaches learn transferable features with minimum perturbation; therefore, the model can classify some images successfully. Despite correct classification, for FGSM attack, the probability decreased for an image belonging to the correct class. We investigated and illustrated that the performance also drops in terms of probability score for successfully classified images.

For an original image x, the correct classification probability was noted. For the same image, the classification score of corresponding adversarial image x' was investigated if both x and x' were correctly classified. Figure 8 shows that, for the same image, the FGSM attack resulted in a degradation of the probability score for the image to belong to a particular class. As shown in Figure 8 for a ε of 0.009, the probability for a COVID-19-positive image belonging to COVID-19-positive decreases to 0.91 from that of 1.00 for the VGG16 network. Additionally, for the Inception-v3 model, the probability also decreases, to 0.93 from that of 1.00. It is obvious that if the ε is further increased, the probability will decrease and result in misclassification. Moreover, the decreased probability value is also dangerous because medical imaging requires high-precision performance. Figure 6c,i shows that ε of 0.009 can generate adversarial images where perturbations are not recognizable in the human eye; Table 5 depicts that ε of 0.008 can cause an average correct class probability drop of 0.24 for the VGG16 model. Thus, the confidence of the classifier to predict the correct class of a sample is reduced, causing the model to be less reliable. The Inception-v3 model was found to be robust to FGSM attack for this task.



(**a**) Without perturbation for VGG16 Probability (COVID-19) = 0.9974707. (**b**) Perturbation Epsilon = 0.009. (**c**) With perturbation for VGG16 Probability (COVID-19) = 0.9090226.



(**d**) Without perturbation for Inception-v3, Probability (COVID-19) = 1.00000. (**e**) Perturbation Epsilon = 0.009. (**f**) With perturbation for Inception-v3, Probability (COVID-19) = 0.930212.

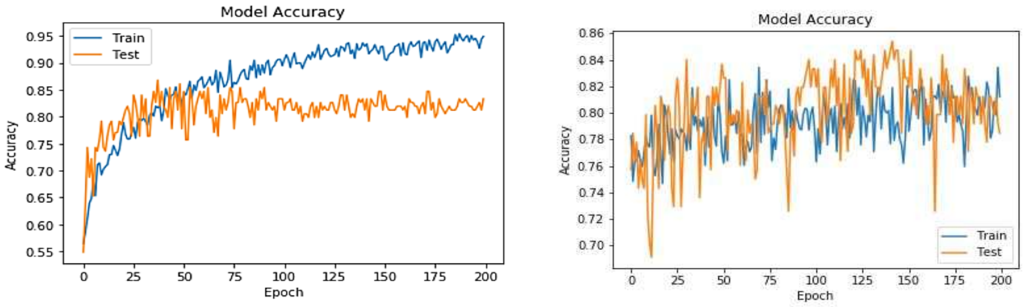
Figure 8. Sample drop of class probability score for successfully classified adversarial chest X-ray images.

Table 5. Average drop of predicted correct class probability for adversarial chest X-ray images.

ε	**VGG16**			**Inception-v3**		
	Average Original Probability	**Average Adversarial Probability**	**Average Probability Decrease**	**Average Original Probability**	**Average Adversarial Probability**	**Average Probability Decrease**
0.0001	0.966888	0.962881	0.004008	1	1	0
0.0003	0.966888	0.953583	0.013305	1	0.999663	0.000337
0.0005	0.966888	0.942397	0.024491	0.999234	0.981813	0.017422
0.0007	0.966888	0.92935	0.037538	0.999215	0.981803	0.017413
0.0009	0.975621	0.930943	0.044678	0.999226	0.981413	0.017813
0.002	0.984782	0.867197	0.117585	1	0.979678	0.020322
0.004	0.993125	0.795419	0.197706	1	0.979668	0.020332
0.006	0.996665	0.810233	0.186431	1	0.979638	0.020362
0.008	0.997023	0.753577	0.243445	1	0.979528	0.020472

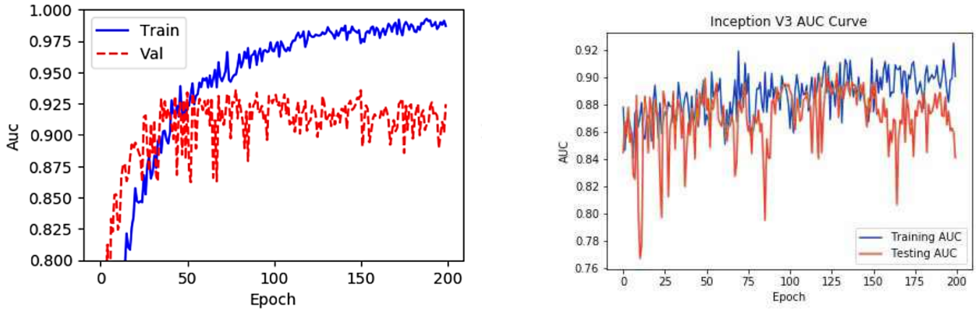
3.3. Performance of VGG16 and Inception-v3 in Diagnosing COVID-19 from Chest CT Images

In addition to analysis on chest X-ray image-based COVID-19 diagnostic approaches, we analyzed the performance and vulnerabilities of the pre-trained models for chest CT-based diagnosis approaches. Similar to the previous experiment, we used an 80:20 split of data to divide the dataset into a training and test set. Figure 9 shows the training and test accuracies for the VGG16 and Inception-v3 models during different training epochs. It can also be seen from Tables 6 and 7, and Figures 9 and 10 that the AUC, precision, and recall are the same in terms of the test sample classification for these models. Although the amount of training data affects the learning performance, our experiment focuses on the variation of performance due to the FGSM attack for any given volume of training data, as illustrated in the next experiment.



(**a**) Accuracy of VGG16.

(**b**) Accuracy of Inception-v3.



(**c**) AUC of VGG16.

(**d**) AUC of Inception-v3.

Figure 9. Accuracy and AUC of transfer learning models for COVID-19 detection from chest CT scans.

Table 6. Confusion matrix for performance of VGG16 on CT scans.

	Actual COVID-19	Actual Normal
Predicted COVID-19	60	14
Predicted Normal	10	66

Table 7. Confusion matrix for performance of Inception-v3 on CT scans.

	Actual COVID-19	Actual Normal
Predicted COVID-19	60	14
Predicted Normal	10	66

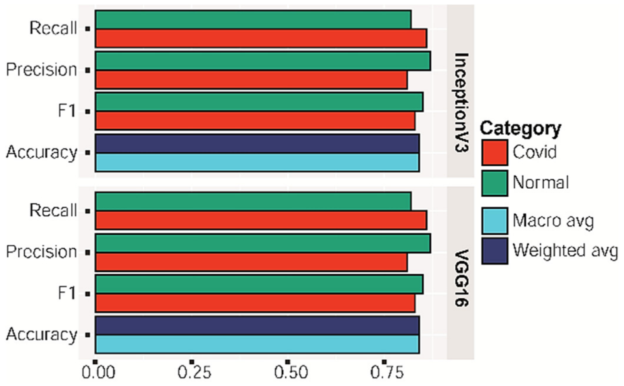
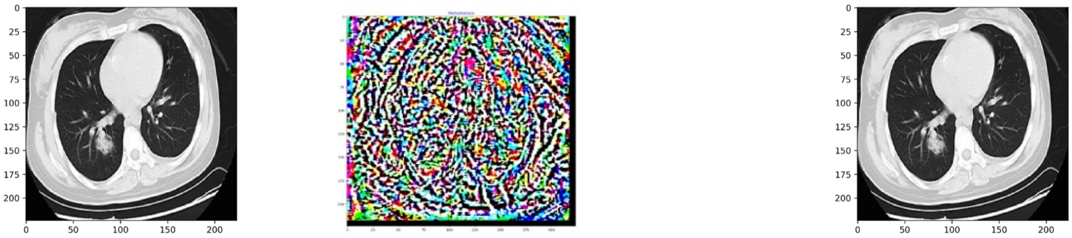


Figure 10. Precision, recall, F score and accuracy for VGG16 and Inception-v3 on chest CT scans.

3.4. FGSM Attack Analysis for Chest CT Images

CT scans are more significant compared to X-rays because of high-quality, detailed image generation capability. This sophisticated X-ray can take a 360-degree image of the internal organs by rotating an X-ray tube around the patient and make internal anatomy clearer by eliminating overlapping structures. However, efficient adversarial images can also be crafted for these images.

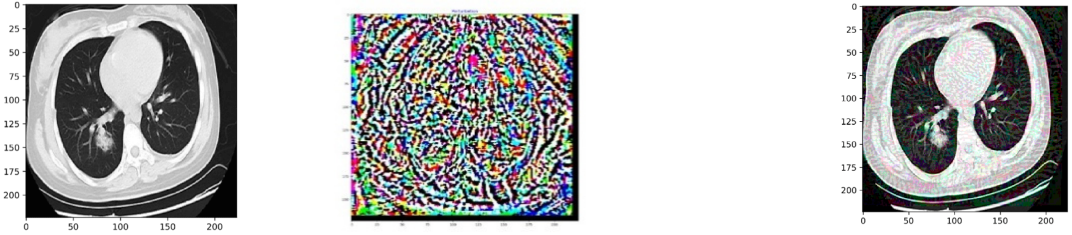
To illustrate the similar vulnerability of promising TL models for CT image-based COVID-19 detection, we investigated the effect of perturbation (ε) variation in FGSM attack. Figure 11c,i clearly show that misclassification can occur with a very minor perturbation and for both considered models. ε of 0.003 or 0.0007 successfully generated adversarial images due to the FGSM attack, where noise was imperceptible to human eye but caused misclassification. On the other hand, perturbation of around 0.07 or 0.09 generated misclassified adversarial images which could be detected by the human eye, as seen in Figure 11f,l. Table 8 and Figure 12 elucidate that for an imperceptible perturbation (ε) such as 0.003, the classification performance drops significantly to 36% for VGG16, and for ε of 0.0007, performance drops to 40% for Inception-v3, making these models unusable for COVID-19 detection purposes.



(**a**) Original image
Prediction: COVID-19-positive.

(**b**) Perturbation
ε = 0.003.

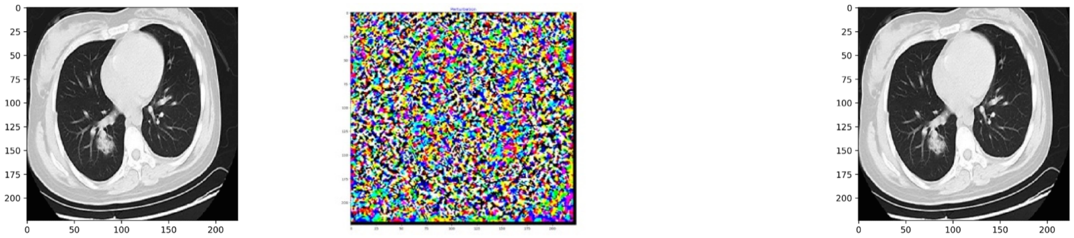
(**c**) Adversarial image with human-imperceptible perturbation. Prediction: COVID-19-negative by VGG16.



(**d**) Original image
Prediction: COVID-19-positive.

(**e**) Perturbation
ε = 0.07.

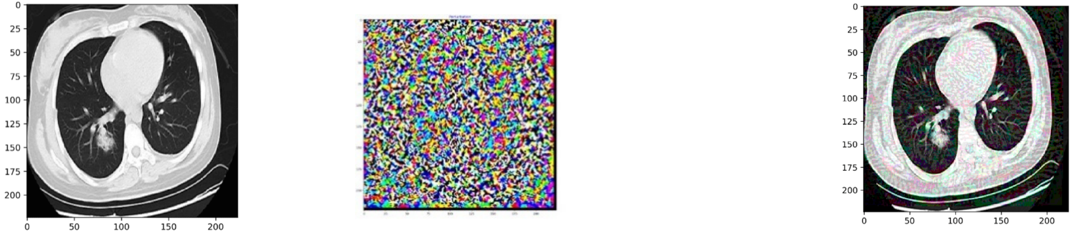
(**f**) Adversarial image with human-perceptible perturbation. Prediction: COVID-19-negative by VGG16.



(**g**) Original image
Prediction: COVID-19-positive.

(**h**) Perturbation
ε = 0.0007.

(**i**) Adversarial image with human-imperceptible perturbation. Prediction: COVID-19-negative by Inception-v3.



(**j**) Original image
Prediction: COVID-19-positive.

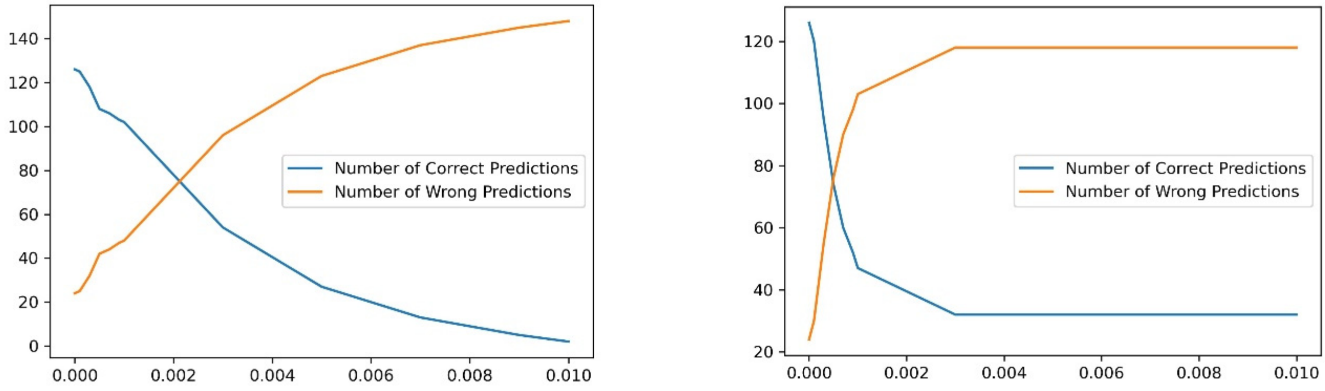
(**k**) Perturbation
ε = 0.09.

(**l**) Adversarial image with human-perceptible perturbation. Prediction: COVID-19-negative by Inception-v3.

Figure 11. Original image (**left column**), perturbation (**middle column**), and corresponding misclassified adversarial CT images (**right column**) generated by FGSM attack for VGG16 and Inception-v3.

Table 8. Diagnostic performance drop for different ε of FGSM attack in chest CT images.

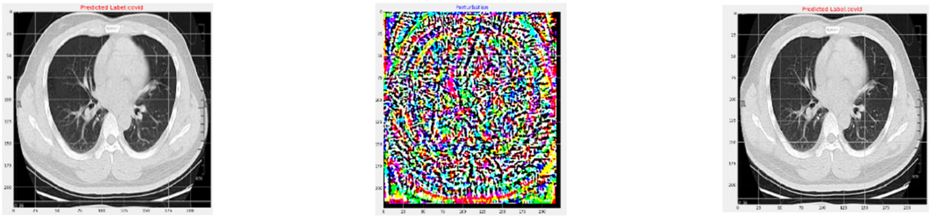
ε	VGG16 with Adversarial Image			Inception-v3 with Adversarial Image		
	Total Correct	**Total Incorrect**	**ACC (%)**	**Total Correct**	**Total Incorrect**	**ACC (%)**
0	126	24	84.00	126	24	84.00
0.0001	125	25	83.33	120	30	80.00
0.0003	118	32	78.67	95	55	63.33
0.0005	108	42	72.00	74	76	49.33
0.0007	106	44	70.67	60	90	40.00
0.0009	103	47	68.67	52	98	34.67
0.001	102	48	68.00	47	103	31.33
0.003	54	96	36.00	32	118	21.33
0.005	27	123	18.00	32	118	21.33
0.007	13	137	8.67	32	118	21.33
0.009	5	145	3.33	32	118	21.33
0.01	2	148	1.33	32	118	21.33



(**a**) VGG16. (**b**) Inception-v3.

Figure 12. Performance for different ε of FGSM attack in chest CT images.

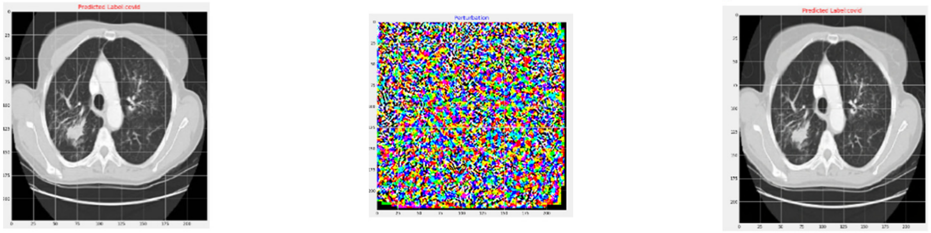
Finally, we investigated the drop in class probability score for correctly classified CT images based COVID-19 detection. Figure 13 shows that for same image, FGSM attack resulted in a decrease in probability score for the image to belong to any class. As shown in Figure 13, for a ε of 0.009, the probability of a COVID-19-positive image belonging to COVID-19-positive decreases to 0.93 from that of 0.99 when VGG16 is used. The probability also decreases to 0.98 from that of 1.00 for the Inception-v3 network in the presence of adversarial images that are not recognizable by the human eye. Therefore, it proves the models to be vulnerable to physical deployment in medical systems. Table 9 depicts that ε of 0.008 can cause an average probability drop of 0.17 for the VGG16 model, reducing the confidence of the classifier to predict the correct class of a sample which also makes the model vulnerable. The Inception-v3 model was found to be comparatively robust for the correctly classified samples.



(**a**) Original image, model VGG16 Probability (COVID-19) = 0.99999213.

(**b**) Perturbation ε = 0.009.

(**c**) With perturbation for VGG16 Probability (COVID-19) = 0.93089885.



(**d**) Original image, model Inception-v3 Probability (COVID-19) = 1.0000.

(**e**) Perturbation ε = 0.009.

(**f**) With perturbation, for Inception-v3 Probability (COVID-19) = 0.9889.

Figure 13. Sample drop of class probability score for successfully classified adversarial chest CT images.

Table 9. Average drop of predicted correct class probability for adversarial chest CT images.

ε	**VGG16**			**Inception-v3**		
	Average Original Probability	**Average Adversarial Probability**	**Average Probability Decrease**	**Average Original Probability**	**Average Adversarial Probability**	**Average Probability Decrease**
0.0001	0.970143	0.95651	0.013633	1	0.999243	0.000757
0.0003	0.979453	0.943228	0.036225	0.9998	0.9965	0.0033
0.0005	0.984652	0.935994	0.048657	1	0.994946	0.005054
0.0007	0.990419	0.930707	0.059711	1	0.994075	0.005925
0.0009	0.995197	0.938476	0.056721	0.997934	0.986894	0.01104
0.002	0.997335	0.903967	0.093368	1	0.981473	0.018527
0.004	0.999686	0.864664	0.135021	0.99249	0.973294	0.019196
0.006	1	0.836082	0.163918	1	0.972294	0.027706
0.008	1	0.826898	0.173102	1	0.971526	0.028474

4. Discussion

The COVID-19 pandemic is a danger to global health and requires the development of models to identify infected people and isolate them. To automate the diagnosis process from chest radiology images, deep learning-based artificial intelligence techniques provide a promising method to address the problem and can be quickly and inexpensively used in a pandemic situation.

However, the most promising deep learning-based approaches require vulnerability analysis to adversarial attacks such as FGSM attack before deployment. Most frequently used pre-trained models to develop radiology image-based COVID-19 diagnosis techniques are publicly available with all relevant parameters. Moreover, these images are captured in a well-defined standard environment for which attack generation is also easier. Sharing of the images to build big data environment for rare disease such as COVID-19, the sharing of reusable pre-trained deep learning model parameters and access of the intruders to computerized and network-based diagnosis systems play a vital role to make the system vulnerable to adversarial attack. Therefore, there are widespread relevant research opportunities.

We developed transfer learning-based deep learning methods from popular pre-trained models VGG16 and Inception-v3. For both X-ray and CT images, these models showed trustworthy performance in terms of various metrics such as accuracy, precision, recall, F1 score, and AUC. Apart from that, this research investigated the vulnerability of the developed deep learning models which are representative of transfer learning-based models for COVID-19 detection from radiology images. For X-ray images, the VGG16 model accuracy dropped significantly by more than 90%, and for the Inception-v3 network, it dropped by 30% if the perturbation increased from 0.0001 to 0.09. Similarly, for CT images, the FGSM attack also revealed potential risks such as misclassification. Moreover, our study shows that the degree of perturbation considerably affects human perceptibility of attacks. This study depicted that for small perturbations, although no noise can be visible in the adversarial images, misclassification as well as class probability reduction can happen for these images. Rich textures in COVID-19 X-ray images often cause the deep learning models to focus on unimportant regions of the features. Therefore, the adversarial attack needs to be considered for these image-based COVID-19 diagnosis techniques before they are practically deployed. The FGSM attack can be crafted from open-source resources; therefore, this research utilized open-source pre-trained models, parameters, and datasets for COVID-19 detection.

Analyzing the vulnerability for other attacks and examining existing defense method suitability can be an important future work. Existing defense techniques include network distillation to extract knowledge from deep neural networks for robustness [21], adversarial training that trains the network with adversarial examples [19], and detecting adversarial examples in the testing stage [35] as well as designing novel training methods such as IMA that increase the margins of training samples in the input space for improved robustness [36]. Additionally, because of multi-faceted adversarial examples, multiple defense strategies can be performed together. More data for training the pre-trained models that improve the classification and reduce vulnerability can be analyzed in versatile settings. Moreover, a robust training method design to overcome FGSM attack for COVID-19 detection algorithms could be an interesting extension for this research.

Author Contributions: Conceptualization, B.P., D.G. and M.A.M.; Methodology, B.P. and D.G. Software, B.P. and D.G.; Writing—original draft preparation, B.P., D.G., M.R.-A.-M. and M.A.M.; Writing—review and editing, M.A.M. and S.A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: Data are available according to the provided references.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, W.; Xu, Y.; Gao, R.; Lu, R.; Han, K.; Wu, G.; Tan, W. Detection of SARS-CoV-2 in different types of clinical specimens. *JAMA* **2020**, *323*, 1843–1844. [CrossRef]
2. Zheng, C.; Deng, X.; Fu, Q.; Zhou, Q.; Feng, J.; Ma, H.; Liu, W.; Wang, X. Deep learning-based detection for COVID-19 from chest CT using weak label. *MedRxiv* **2020**. [CrossRef]
3. Fang, Y.; Zhang, H.; Xie, J.; Lin, M.; Ying, L.; Pang, P.; Ji, W. Sensitivity of chest CT for COVID-19: Comparison to RT-PCR. *Radiology* **2020**, *296*, E115–E117. [CrossRef]
4. Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; Fan, G.; Xu, J.; Gu, X.; et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **2020**, *395*, 497–506. [CrossRef]
5. Lei, J.; Li, J.; Li, X.; Qi, X. CT imaging of the 2019 novel coronavirus (2019-nCoV) pneumonia. *Radiology* **2020**, *295*, 18. [CrossRef] [PubMed]
6. Song, F.; Shi, N.; Shan, F.; Zhang, Z.; Shen, J.; Lu, H.; Ling, Y.; Jiang, Y.; Shi, Y. Emerging 2019 novel coronavirus (2019-nCoV) pneumonia. *Radiology* **2020**, *295*, 210–217. [CrossRef] [PubMed]
7. Ng, M.Y.; Lee, E.Y.; Yang, J.; Yang, F.; Li, X.; Wang, H.; Lui, M.M.S.; Lo, C.S.Y.; Leung, B.; Khong, P.L.; et al. Imaging profile of the COVID-19 infection: Radiologic findings and literature review. *Radiology* **2020**, *2*, e200034. [CrossRef] [PubMed]

8. Ai, T.; Yang, Z.; Hou, H.; Zhan, C.; Chen, C.; Lv, W.; Tao, Q.; Sun, Z.; Xia, L. Correlation of chest CT and RT-PCR testing forcoronavirus disease 2019 (COVID-19) in China: A report of 1014 cases. *Radiology* **2020**, *296*, E32–E40. [CrossRef] [PubMed]
9. FDA. FDA Permits Marketing of Artificial Intelligence-Based Device to Detect Certain Diabetes-Related Eye Problems. Silver Spring, DM, Department of Health and Human Services 2018. Available online: https://www.fda.gov/news-events/press-announcements/fda-permits-marketing-artificial-intelligence-based-device-detect-certain-diabetes-related-eye (accessed on 15 April 2021).
10. Ardila, D.; Kiraly, A.P.; Bharadwaj, S.; Choi, B.; Reicher, J.J.; Peng, L.; Tse, D.; Etemadi, M.; Ye, W.; Corrado, G.; et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **2019**, *25*, 954–961. [CrossRef]
11. Suzuki, K. Overview of deep learning in medical imaging. *Radiol. Phys. Technol.* **2017**, *10*, 257–273. [CrossRef]
12. Gozes, O.; Frid-Adar, M.; Greenspan, H.; Browning, P.D.; Zhang, H.; Ji, W.; Bernheim, A.; Siegel, E. Rapid AI Development Cycle for the Coronavirus (covid-19) Pandemic: Initial Results for Automated Detection and Patient Monitoring Using Deep Learning CT Image Analysis. *arXiv* **2020**, arXiv:2003.05037. Available online: https://arxiv.org/ftp/arxiv/papers/2003/2003.05037.pdf (accessed on 15 April 2021).
13. Li, L.; Qin, L.; Xu, Z.; Yin, Y.; Wang, X.; Kong, B.; Bai, J.; Lu, Y.; Fang, Z.; Song, Q.; et al. Artificial intelligence distinguishes COVID-19 from community acquired pneumonia on chest CT. *Radiology* **2020**. [CrossRef]
14. Wang, S.; Kang, B.; Ma, J.; Zeng, X.; Xiao, M.; Guo, J.; Cai, M.; Yang, J.; Li, Y.; Meng, X.; et al. A deep learning algorithm using CT images to screen for Corona Virus Disease (COVID-19). *MedRxiv* **2020**. [CrossRef]
15. Narin, A.; Kaya, C.; Pamuk, Z. Automatic Detection of Coronavirus Disease (covid-19) Using X-ray Images and Deep Convolutional Neural Networks. *arXiv* **2020**, arXiv:2003.10849. Available online: https://arxiv.org/ftp/arxiv/papers/2003/2003.10849.pdf (accessed on 15 April 2021).
16. Makris, A.; Kontopoulos, I.; Tserpes, K. COVID-19 detection from chest X-ray images using Deep Learning and Convolutional Neural Networks. In Proceedings of the 11th Hellenic Conference on Artificial Intelligence, Athens, Greece, 2–4 September 2020; pp. 60–66.
17. Karim, M.; Döhmen, T.; Rebholz-Schuhmann, D.; Decker, S.; Cochez, M.; Beyan, O. DeepCOVIDExplainer: Explainable Covid-19 Predictions Based on Chest X-ray Images. *arXiv* **2020**, arXiv:2004.04582. Available online: https://arxiv.org/pdf/2004.04582.pdf (accessed on 15 April 2021).
18. Asnaoui, K.E.; Chawki, Y.; Idri, A. Automated methods for detection and classification pneumonia based on X-ray images using deep learning. *arXiv* **2020**, arXiv:2003.14363. Available online: https://arxiv.org/ftp/arxiv/papers/2003/2003.14363.pdf (accessed on 15 April 2021).
19. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572. Available online: https://arxiv.org/pdf/1412.6572.pdf (accessed on 15 April 2021).
20. Simon, G. *Principles of Chest X-ray Diagnosis*; Butterworths: Oxford, MA, USA, 1971.
21. Papernot, N.; McDaniel, P.; Wu, X.; Jha, S.; Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. In Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 23–25 May 2016; pp. 582–597.
22. Grosse, K.; Papernot, N.; Manoharan, P.; Backes, M.; McDaniel, P. Adversarial Examples for Malware Detection. In Proceedings of the 22nd European Symposium on Research in Computer Security, Oslo, Norway, 11–15 September 2017; Springer: Berlin/Heidelberg, Germany, 2017; pp. 62–79.
23. Kurakin, A.; Goodfellow, I.; Bengio, S. Adversarial Examples in the Physical World. 2016. Available online: https://arxiv.org/pdf/1607.02533.pdf (accessed on 15 April 2021).
24. Melis, M.; Demontis, A.; Biggio, B.; Brown, G.; Fumera, G.; Roli, F. Is deep learning safe for robot vision? Adversarial examples against the icub humanoid. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 751–759.
25. Cohen, J.P.; Morrison, P.; Dao, L.; Roth, K.; Duong, T.Q.; Ghassemi, M. Covid-19 Image Data Collection: Prospective Predictions Are the Future. *arXiv* **2020**, arXiv:2006.11988. Available online: https://arxiv.org/pdf/2006.11988.pdf (accessed on 15 April 2021).
26. Yang, X.; He, X.; Zhao, J.; Zhang, Y.; Zhang, S.; Xie, P. COVID-CT-Dataset: A CT Scan Dataset about COVID-19. *arXiv* **2020**, arXiv:2003.13865.
27. Pal, B.; Ahmed, B. A deep domain adaption approach for object recognition using Multiple Model Consistency analysis. In Proceedings of the 2016 9th International Conference on Electrical and Computer Engineering (ICECE), Dhaka, Bangladesh, 20–22 December 2016; pp. 562–565.
28. O'Shea, K.; Nash, R. An Introduction to Convolutional Neural Networks. *arXiv* **2015**, arXiv:1511.08458v2. Available online: https://arxiv.org/pdf/1511.08458.pdf (accessed on 15 April 2021).
29. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556. Available online: https://arxiv.org/pdf/1409.1556.pdf (accessed on 15 April 2021).
30. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.

31. Nguyen, A.; Yosinski, J.; Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 427–436.
32. Tramer, F.; Carlini, N.; Brendel, W.; Madry, A. On Adaptive Attacks to Adversarial Example Defenses. *arXiv* **2020**, arXiv:2002.08347. Available online: https://arxiv.org/pdf/2002.08347.pdf (accessed on 15 April 2021).
33. Chen, J.; Qian, L.; Urakov, T.; Gu, W.; Liang, L. Adversarial robustness study of convolutional neural network for lumbar disk shape reconstruction from MR images. In *Medical Imaging 2021: Image Processing*; International Society for Optics and Photonics: Washington, DC, USA, 2021; Volume 11596, p. 1159615.
34. Yuan, X.; He, P.; Zhu, Q.; Li, X. Adversarial examples: Attacks and defenses for deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 2805–2824. [CrossRef] [PubMed]
35. Lu, J.; Issaranon, T.; Forsyth, D. Safetynet: Detecting and rejecting adversarial examples robustly. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 446–454.
36. Ma, L.; Liang, L. Increasing-Margin Adversarial (IMA) Training to Improve Adversarial Robustness of Neural Networks. *arXiv* **2021**, arXiv:2005.09147. Available online: https://arxiv.org/pdf/2005.09147.pdf (accessed on 15 April 2021).