Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Security methods for AI based COVID-19 analysis system: A survey

Samaneh Shamshiri, Insoo Sohn*

Division of Electronics & Electrical Engineering, Dongguk University, Seoul, Republic of Korea

Received 9 August 2021; received in revised form 28 February 2022; accepted 3 March 2022
Available online 16 March 2022

Abstract

Rapid progress and widespread outbreak of COVID-19 have caused devastating influence on the health systems all around the world. The importance of countermeasures to tackle this problem lead to widespread use of Computer Aided Diagnosis (CADs) applications using deep neural networks. The unprecedented success of machine learning techniques, especially deep learning networks in medical images, have led to their recent prominence in improving efficient diagnosis of COVID-19 with increased detection accuracy. However, recent studies in the field of security of AI-based systems revealed that these deep learning models are vulnerable to adversarial attacks. Adversarial examples generated by attack algorithms are not recognizable by the human eye and can easily deceive the state-of-the-art deep learning models, therefore they threaten security-critical learning applications. In this paper, the methodology, results and concerns of recent works on robustness of AI based COVID-19 systems are summarized and discussed. We explore important security concerns related to deep neural networks and review current state-of-the-art defense methods to prevent performance degradation.

© 2022 The Author(s). Published by Elsevier B.V. on behalf of The Korean Institute of Communications and Information Sciences. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords: COVID-19; Deep learning; Security; Adversarial attack

Contents

1. Introduction ........................................................................................................................................................................... 555
2. Basic concepts ....................................................................................................................................................................... 556
   2.1. COVID-19 datasets ....................................................................................................................................................... 556
   2.2. Attacks ....................................................................................................................................................................... 556
   2.3. Defense methods against adversarial attacks .................................................................................................................. 557
3. Security studies for COVID-19 AI-based models .................................................................................................................... 557
4. Discussion and conclusion ...................................................................................................................................................... 560
   Declaration of competing interest ........................................................................................................................................... 561
   Acknowledgments .............................................................................................................................................................. 561
   References ............................................................................................................................................................................ 561

1. Introduction

Across the worldwide coronavirus 2019 (COVID-19) pandemic, not only public medical systems have been challenged greatly, but it also affects the globe economically, socially, educationally, etc. DNN-based AI systems are considerably gaining importance as they effectively increase COVID-19 diagnosis process. Although the state-of-the-art DNN models obtained a fair amount of success in diagnosing COVID-19 with high rate of accuracy [1–4], their robustness against security threats are considered as a major risk. The reason is that these attack models are able to cause failure in most image classification tasks. Therefore, the study of AI techniques in COVID-19 diagnosis systems are of great importance for healthy human lives and have lead to active research on performance optimization against adversarial attacks. To generate adversarial attacks, various methods have been proposed to investigate the vulnerability of DL models. Most of these attack algorithms are classical models such as fast gradient

* Corresponding author.
E-mail addresses: 2021121108@dgu.ac.kr (S. Shamshiri), isoohn@dongguk.edu (I. Sohn).
Peer review under responsibility of The Korean Institute of Communications and Information Sciences (KICS).

https://doi.org/10.1016/j.icte.2022.03.002
2405-9595/© 2022 The Author(s). Published by Elsevier B.V. on behalf of The Korean Institute of Communications and Information Sciences. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
sign method (FGSM) [5], projected gradient decent (PGD) [6], and universal adversarial perturbations (UAPs) [7]. One novel strong method, stabilized medical image attack (SMIA) [8], has been proposed to attack COVID-19 detecting models.

Due to the concerns about the weals performance of deep neural networks for COVID-19 disease, defending approaches have been studied to improve the effectiveness of these systems against various attack models [9]. One of the most applicable method is adversarial training [5,10] which augments the training set with adversarial samples to improve the model’s adversarial robustness. Furthermore, two models have been proposed for the first time with the efficient defense approaches to decrease the vulnerability of AI based COVID-19 systems.

The paper is organized as follows. In Section 2, we concentrate on a brief explanation about some basic concepts in terms of AI based system’s security. In Section 3, we summarize the recent studies which specifically explored the attack methods to illustrate that DNN based models for detecting COVID-19 are easily fooled by small amount of imperceptible perturbations. In this section, we considered both attack and defense approaches along with their important results and bottlenecks. Finally, we will wrap up this paper by discussing the results and some conclusions.

2. Basic concepts

2.1. COVID-19 datsets

Data is the first pipeline in deep learning classification which plays a significant role in diagnostic tools and treatments. The datasets for training and evaluating the AI based COVID-19 detection models are comprised of different repositories mostly based on chest X-ray (CXR) and computed tomography (CT-scan) images. It is an indisputable fact that there are serious concerns about data privacy and confidentiality, nevertheless, some websites like Kaggle and Github provide open access datasets which has been collected from different hospitals all around the world. There were a large collection for chest X-rays categories such as normal and pneumonia patients [11–14] which have been completed by adding COVID-19 cases. COVIDx dataset, proposed by Wang et al. [1] is the largest open access benchmark dataset including the most number of COVID-19 positive patient cases. This dataset consists of 13975 CXR images from 138710 patients. It is noteworthy to mention that COVIDx contains 358 CXR images from 266 COVID-19 patients and 8066 patients with no pneumonia and 5538 patient cases with non-COVID-19 pneumonia. Another important dataset has been proposed by Cohen et al. [15] in Dec. of 2020. They proposed a dataset of COVID-19 images along with metadata of clinical tasks, including 676 chest X-ray images from 412 people from 26 countries. In March of 2020, they published their research works on COVID-19 image data collection of 100 COVID-19 cases. This dataset complied from many websites and many papers making it one of the major references for COVID-19 research.

2.2. Attacks

In this section, we briefly discuss some of the most important concepts of attacks against AI based systems.

- **Perturbation:** A small perturbation, which is added to a dataset, is a small noise that can cause changes in a system by inferring model parameters. This small perturbation imperceptibly cause a network to misclassify [5].

- **Adversarial attack:** In recent decades, adversarial attacks have been considered as one of the most important challenges for deep learning systems’ security and robustness. This kind of attacks are inputs which have been manipulated by small perturbations for neural networks to misclassify [16,17].

- **Non-targeted and targeted attacks :** In terms of targeted attack, deep model classified the input image into a specific class, set by the attackers while for non-targeted attacks, an input image has been assigned an incorrect class or label.

- **Black-box attacks:** In this attack, the adversary only has knowledge about the output of the model and has no access not only to the trained model, training dataset and model parameter, but also has no more access than any normal user [18].

- **White-box attacks:** Despite the black-box attacks, white-box attacks occur when the attacker has complete access to the trained model, network architecture, hyper-parameters, weights, and any information which is accessible to the network’s trainer service [18].

- **Generating adversarial attacks:** To generate adversarial examples, several algorithms have been proposed recently. L-BFGS attack was the first algorithm introduced, with high stability as its main advantage [17]. For an input image $x$, their models recognize an image $x^*$ under $L_2$ distances with different labels. Due to time complexity issues of L-BFGS, fast gradient sign method (FGSM) [5] and projected gradient descend (PGD) [6], which are two gradient based adversarial examples, have been proposed. The main aim of the attacks is misclassifying the deep model with minimum amount of perturbations. However, FGSM, with high transfer rate, suffers from label leaking, so it has been proposed for methods with small datasets [18,19]. While the attacks have been optimized for the $L_\infty$ distance metrics, another attack algorithm with $L_0$ distance optimization has been introduced as jacobian-based saliency map (JSMA) [20]. Despite the high attack success rate of JSMA, this method is one of the most computationally expensive ones. C&W attack [21], which is a modified version of JSMA, shows strong performance against defense methods based on different
metrics \((L_0, L_2, L_\infty)\) to achieve high transfer and success rate. But it has the problem of high computational complexity as well. Deepfool attack \([22]\) that minimize the distance between clean input data and decision boundary of adversarial examples, is an effective attack method for optimization under \(L_2\) metric distance. In contrast with \(L\)-BFGS, it solves the problem of time-complexity, but it does not have a high success rate for black-box attacks \([18]\). Universal adversarial perturbation (UAP) misclassifies the natural images with high performance and generalization ability. In this method the perturbation cannot be controlled easily, causing low success rate. The purpose of the basic iterative method (BIM) \([23]\), which is an extension of FGSM, is to reach a remarkable optimization by several iterations. The above mentioned attacks are the most important attacks to evaluate the vulnerability of COVID-19 detection DL models.

2.3. Defense methods against adversarial attacks

There are several countermeasures to defend AI systems against attacks which categorized into two groups. While white-box defense provide training the deep model with adversarial images as an input, the black-box defense, does not train the models with adversarial images. Several techniques such as data augmentation, input transformation and an encryption inspired shuffling of images are known to be mostly used black-box defense methods. Furthermore, two approaches have been proposed for defending strategies: reactive and proactive groups \([19]\). In the case of reactive strategy, after building deep models, adversarial examples should be detected, while in the process of proactive methods, the aim is increasing the robustness of the deep neural network before any adversarial attacks. Adversarial training \([5,10]\), is one of the most widely used proactive defense methods to increase the robustness of neural networks against first-order attack methods \([6]\). This method has been evaluated on difference datasets \([18]\), including COVID-19 datasets, and shows significant results via regularization to avoid overfitting and precision improvements.

3. Security studies for COVID-19 AI-based models

One of the first attempts to evaluate the vulnerability of DL networks used for COVID-19 predictions was presented by Hirano et al. \([24]\). They investigated a convolutional neural network (CNN) based COVID-Net \([1]\) model which is one of the first COVID-19 detection methods using CXR images. To generate perturbation, fast gradient sign method (FGSM) was used for both targeted and non-targeted attacks. For an input image \(x\), classifier \(C(x)\), and perturbation \(\hat{\rho}\) is applied in Eq. (1) and Eq. (2) as shown below:

\[
\hat{\rho} = \varepsilon.\nabla_x L(x, C(x))/\|\nabla_x L(x, C(x))\|_p
\]

(1)

\[
\hat{\rho} = -\varepsilon.\nabla_x L(x, y)/\|\nabla_x L(x, y)\|_p
\]

(2)

where \(\varepsilon > 0\) is the attack strength. Iterative algorithms were used to update the UAP(\(\rho\)), until \(L_\rho\) norm of perturbation is equal to or less than a small \(\xi\) value. The parameters in this method were set as \(\varepsilon = 0.001\), \(P = 2\) and \(\infty\), \(\xi = 1\%\) and \(2\%\), average \(L_\infty\) and \(L_2\) were 237 and 32589, respectively. The authors used three metrics \(R_f\), the fooling rate, \(R_a\), the attack success rate and confusion metrics to evaluate the proposed COVID-Net CXR small and COVID-Net CXR large. The results showed that for both of COVID-Net models, the UAPs with \(\xi = 2\%\) achieved \(>85\%\) and \(>90\%\) success rate for non-targeted and targeted attacks, respectively. Furthermore, for higher \(\xi\), these models evaluated most of normal and pneumonia test images as COVID-19 cases. This study extensively investigated the robustness of COVIDNet-CXR, when attacked by small UAPs and also considered the defense methods. Moreover, they comprehensively analyzed the augmentation of vulnerability of COVID-Net in comparison with some other DL models (ResNet and VGG). The authors studied the role of unbalanced COVIDx dataset in the vulnerability of COVID-Net.

Qi et al. \([8]\) introduced a novel stabilized medical image attack (SMIA) method, in which the adversarial perturbations have been generated by an objective function, consisting of a loss deviation term (DEV) and a loss stabilization term (STA). Let \(x\) be an input of the model, \(Y\) be the ground truth label of \(x\), \(\eta\) be a perturbation, and \(f(\theta, x)\) be the CNN predicted results of \(x\) for model parameters as \(\theta\), then the objective function can be described as:

\[
L_{SMIA} = L(f(\theta, \tilde{x}), Y) - \alpha L(f(\theta, \tilde{x}), f(\theta, \tilde{x} + \eta'))
\]

(3)

where the loss function \(L(\cdot)\) is the cross entropy loss, \(\alpha\) is a scalar balancing the influence of \(L_{DEV}\) and \(L_{STA}\), \(\tilde{x} = x + \eta\) is the adversarial example sent to the CNN in the current iteration, \(\eta = W\ast \eta - \eta\), where \(W\) is a Gaussian kernel. To generate an effective attack, the objective function \((L_{DEV} + L_{STA})\) are maximizing the deviation loss term and minimizing the stabilization term. To this end, the DEV term which produces perturbation, moving perturbations among single spot, while STA term minimizes the KL-Divergence iteratively to have a low variance of the \(\tilde{\eta}\) that cause the movement of perturbation toward a fixed objective function and cope with the problems of stability. For comparative analysis study, with the state-of-the-art methods on lung segmentation shows about 60\% drop in accuracy rate. This demonstrates that the proposed attack has a big influence on the robustness of U-Net. Based on the complete explanation of SMIA technique in this research, SMIA can effectively challenge the security of deep learning models and improve the attack performance while deceive the CNN model successfully.

Another effective defense method to train a DNNs models on clean and noisy samples to improve robustness has been proposed by Ma et al. \([25]\). Their aim is training the DNN model (ResNet18 \([26]\)) by the novel increasing-margin adversarial (IMA) approach in order to improve the robustness of the DNNs for the classification task. The authors applied the proposed DNN model for the detection of COVID-19 from CT scan images (Soares dataset \([4]\)). IMA is an adversarial training algorithm, which utilized projected gradient descent (PGD) algorithm for generating adversarial noises. In this
study, by a trained classifier and addition of small amount of noise \( \delta \) to a sample \( x \), \( x + \delta \) achieves an optimum decision boundary. Therefore, by aggregation of data points around decision boundary, misclassification has been occurred with zero accuracy. The aim of the model is increasing the margins of the training samples when there are a large distance between decision boundaries of the model and the training samples. IMA includes three algorithms for computing loss functions and updating the model in each epoch. The margin estimations are updated such that the amount of noisy samples on decision boundary increases as much as possible, therefore, resulting in maximum margins. In comparison with three state-of-the-art adversarial training methods (20-PGD, Trades [18], and MMA [27]), the IMA method shows robust performance for noise with level less than 0.2. While most of the models in the literature focused on evaluating the vulnerability of deep learning models for detecting COVID-19, this paper proposed a new defense method to increase the performance of ResNet-18 against security attacks.

Tripathi et al. [28] proposed a novel pre-processing fuzzy unique image transformation (FUIT) technique, which is an efficient black box defense method with two main contributions: First, the proposed techniques are evaluated against adversarial attacks including deep fool, FGSM, basic iterative method (BIM), carlini & wagner (CW), projected gradient descent with random start (PGD-R) and PGD without random start attacks with different amounts of perturbation. In this method, the authors introduced 18 models (M1-M18) for both chest X-ray image datasets and CT-scan ones, 9 models were developed for chest X-ray images and another nine for CT-scan images. The 18 models have been pre-trained by Resnet-18, VGG-16 [29] and GoogleNet [30]. These models have been categorized into 3 groups such that there exist 3 models in each group with different image types including clean images for (M1-M3 and M10-M12), FUIT transformed images for (M4-M6 and M10-M12) and discretization transformed images for (M7-M9 and M16-M18). Furthermore, the robustness of these models for the diagnosis of COVID-19 cases when it has been attack by adversaries has been studied. FUIT creates a fuzzy set \( \tilde{F} = \{(x, \mu_F(x)) | x \in U\} \) where \( x \) is an element of a Universe information of \( U \), and \( \mu \) is the membership value, which is between 0 and 1. After computing the amount of \( \mu \), it is utilized in an algorithm to create \( R \) fuzzy sets which downsample the image pixels from a range of 0 to 255 to an interval \([1, R]\). Therefore, the variance decreases and results to the equality of the number of unique pixel values for both clean and adversarial images. The models which have been pre-trained by ResNet-18 show the highest mean accuracy on clean images. When they have been tested against adversarial attacks, their accuracy decreases to <10%. After pre-processing by FUIT transformation, their performance against those attacks increases significantly to the accuracy value greater than 85%. Defense method of FUIT against adversarial attacks is the first black-box model proposed for COVID-19 detection. The countermeasures include performance evaluation of proposed FUIT and deep learning models accuracy rate in diagnosing COVID-19. The results provide a comprehensive insight on two datasets against 6 state-of-the-art attacks with different parameters which helps to compare the performance of them effectively.

Multi-model adversarial attack is another method which has been proposed by Abdur Rahman et al. [31] to evaluate the performance of six different deep learning models which has been applied for diagnosis of COVID-19. Total of 12 attacks include FGSM, DeepFool, C & W, BIM, L-BFGS, Foolbox, PGD and JSMA that are used to attack on 4 models which include ResNet-101 with three kernel size 0, 3 and 300. By considering white-box, black-box and gray-box attacks, the results show that these DL models are vulnerable to all types of attacks such as evasion, poisoning, extraction and inference. As one of the first papers which studied the security of COVID-19 deep learning systems, not only did they provide results on both chest X-ray and CT-scan images, but they also demonstrated experimentally threats to face mask recognition process. Similar to other attacks on AI based COVID-19, they evaluated their adversarial methods to compare with the state-of-the-art attacks to show the performance of six deep learning models with detailed analysis.

To make a comparison of reliability and explainability of robust deep learning models vs. non-robust ones (standard models), Roberts et al. [32] investigated the performance of standard models when attacked by adversarial perturbations and models which applied adversarial training to become robust. They studied feature optimizations of two deep learning models, VGG-16 and ResNet-18, which were trained on the lung point of care ultrasound imagery (POCUS) [33] dataset of 3119 covid-19 video frames. To achieve the experimental results, the authors regard a framework as \( \delta_{\text{max}} := \arg \max l(x + \delta, y)_{\delta \in B_{\epsilon}(x)} \) and \( \delta_{\text{min}} := \arg \min l(x + \delta, y)_{\delta \in B_{\epsilon}(x)} \), where \( \delta \) is perturbation variable and \( \epsilon \) is the radius of \( L^2 \) ball for locally optimized loss. \( \delta_{\text{max}} \) and \( \delta_{\text{min}} \) are features that refer to pertinent negative/positive respectively. Pertinent negative are misclassified features in the prediction process and pertinent positive features refer to correct predictions in the input data. Analysis of features learned by the aforementioned models and trends in failure cases indicates that models with adversarial training have less sensitivity than the models with standard training. VGG-16 robust achieved the accuracy of 81.4% for COVID-19 while it is just around 4% less than the accuracy of VGG-16 standard. Therefore, diffuse features with less interpretability increase the vulnerability of models with a small number of perturbations. Despite data hungriness, the evaluation results show that the proposed approach is feasible and performs better than the default setting of standard models in terms of features interpretability and robustness of deep models.

To address the problem of low resilience of deep models against attacks because of their non-linearity [17,34], Xu et al. [35] proposed a novel defense Robust and Retrainless Diagnosis Framework for Medical pre-trained models (MedRDF). This framework consists of three main levels. During the network training process, the decision boundary and loss landscape become curvier resulting in concealed adversarial examples. In the first level of the MedRDF framework,
which is pre-processing the input data, MedRDF generates \( n = 10^4 \) number of noisy copies of each test medical image \( x \) which is perturbed by \( \epsilon = 8/255 \). It means to alleviate the problem of increasing curvature of decision boundary while epochs increase, random noise \( \eta \) bounded by \( \sigma \) will add to the image, where \( \sigma \) is an isotropic noise. When the standard accuracy of MedRDF decreases, the increasing noise level \( \sigma \) will control the trade-off between robustness and accuracy and prevent robustness degradation. On the other hand, to solve the problem of a large number of input noises from the first step, custom denoiser D has been adapted to the framework. This denoiser operator inspired by Gaussian smoothing and median filter was added in the second level to get the prediction labels and address the aforementioned problem. Ultimately, by majoring votes on the prediction labels of the denoised ones, MedRDF generated the robust diagnosis of medical image \( x \).

To evaluate the MedRDF performance, the authors defined a novel confidence score metric as Robust Metric (RM). Physicians can consider the different amounts of thresholds for RM based on their expectations, in addition, 3 pre-trained models; ResNet-18, ResNet-50, and AG-Sononet-16 [36], and two groups of white-box and black-box attacks such as I-FGSM, PGD, C&W, SPSA and RayS have been considered. Since the MedRDF is a defense method, its performance was compared with Random RP [37], ComDefend [38], adversarial training, TRADES [39], and MART [40] defend methods. By comprehensive comparison of different attacks, different experimental results of noises and defend methods, the authors advocate MedRDF and the Robust Metric (RM) for utilizing them for medical diagnostic tasks. Due to the weakness of defense methods for deep learning-based models in the field of medical imaging compare with natural images which lead to loss of the small lesions feature in images, it sounds that, MedRDF has an easy-to-deploy framework. They show that the accuracy of pre-trained models improved from 0 when they have been attacked to about 91.4% after getting robust by the MedRDF framework which outperforms the existing defense methods with significantly higher strength performance.

Amini [41] investigated the robustness of five DNN models which include Residual networks family (ResNet-18, 50), wide Residual networks (WRN-16, 8), VGG-19, and InceptionV3. The author applied four state-of-the-art white-box attacks such as FGSM, PGD, C&W and Spatial Transformation attack (ST) [42] with 4551 chest X-ray images of COVID-19. PGD’s iterative nature, which periodically perturbs the gradient values, made it the most effective attack, causing a decrease of 26.1%. Furthermore, the rotation of the images from -30 to 30 degrees by ST attack, affected the classification accuracy, which emphasizes that data augmentation and images rotation play an important role in accuracy optimization. The author provided a comparison between the number of parameters of each deep neural network and believe that although VGG-16 has the most number of parameters, it is WRNs and ResNet-50 which show the best resilience against attacks, therefore, recommended more investigation on the optimization of hyperparameters of the model such as layer types, activation functions, and filter sizes.

In another remarkable research by Gongye et al. [43], gradient-based adversarial attack methods consisting of FGSM and PGD have been applied to generate adversaries to have an end-to-end attack on deep learning model. To generate adversarial examples they considered two kinds of active and passive attacks where the active attacks, have been implemented to fool the deep learning models. Their baseline model, RseNet-18 achieved 94.7% accuracy on clean data. As can be seen from Fig. 1, for FGSM attack, by changing the amount of perturbation from 0.01 to 0.06, the performance of the model decreases to below 1%, while for the PGD attack by considering the hyper parameter \( \epsilon = 0.07 \) the model accuracy degraded to 0% after 15 iterations. The experiments showed successful attack performance in misclassification of COVID-19 images to normal (false negative) cases. Although the authors believe that finding a defense method which protect the deep learning models in all aspects is not practical at this time, they proposed new attacks in which they consider COVID-19 recognition systems as well.

Two transfer learning based pre-trained models for COVID-19 diagnosis, VGG16 and Inception-V3, were applied by Pal et al. [44] for investigating their performance vulnerability against adversarial FGSM attacks. The metrics considered in this paper are accuracy, precision, recall, F1 score and AUC. On clean datasets of chest X-ray and CT-scan images, the VGG16 and Inception-V3 enhanced high rates in all metrics. While the perturbation increases from 0.0001 to 0.09, which is recognizable in the human eye, the accuracy rate drops into 16% for VGG16 and 55% for Inception-V3. In addition, similar results were obtained for FGSM attacks against CT-scan images. While the perturbations considered with 0.003 and 0.0007 were not distinguishable to the human eyes, the performance of VGG16 and Inception-V3 decreased to 36% and 40%, respectively. This study comprehensively compared the performance of two aforementioned deep learning models with different contributions including the rate of accuracy on clean and noisy images and the rate of average probability. They also considered the results of various amount of perturbations that may be recognizable by the human eyes or not. It was shown that in both of conditions, misclassification is inevitable.
Table 1
The summary of the results of applying different AI structures to investigate the robustness of COVID-19 detection systems.

| Article | Attack | Dataset | Architecture | Defense Method | Compared structure | Best attack performance | Best defense performance |
|---------|--------|---------|--------------|---------------|--------------------|------------------------|-------------------------|
| [24]    | UAP    | X-ray   | COVID-Net    | adversarial training | ResNet, VGG | > 90% | > 80% |
| [8]     | SMIA   | CT-scan | U-Net        | –             | FGSM, PGD, Deepfool, DAG | 60.82% (accuracy metric) | – |
| [25]    | 100-PGD| CT-scan | ResNet-18    | IMA           | MMA, TRADES 20-PGD | 100% | 90% |
| [28]    | Deepfool, FGSM, BIM, C&W, PGD, PGD-r | CT-scan | X-ray | ResNet-18 | DarkCOVID-Net, CoroNet VGG19, MobileNet v2 COVID-Net | > 93% | > 96% |
| [31]    | FGSM, MI-FGSM, DF, L-BFGS, C&W, BIM, FB, PGD, JSMA, BD, MS, poisoning | CT-scan | X-ray | ResNet, YOLO, DarkNet, GRAD-CAM | – | – | 91% | – |
| [43]    | FGSM, PGD | X-ray | ResNet-18 | – | – | 94.7% | – |
| [44]    | FGSM | X-ray, CT-scans | VGG16 InceptionV3 | – | – | 83.3% | – |
| [32]    | – | POCUS | VGG16 ResNet18 | adversarial training | – | 100% | > 81% |
| [35]    | I-FGSM, PGD, C&W, SPSA, Rays | X-ray | ResNet50, ResNet18 AG-Sononet-16 MedRDF | Random RP, ComDefend, Adversarial Training TRADES, MART | 93.2% | 91.4% |
| [41]    | FGSM, PGD, C&W, ST | X-ray | ResNet50, ResNet18 WRN-16-8 VGG-19, InceptionV3 | – | – | >60% | – |

4. Discussion and conclusion

In this study, we provided a survey about the security challenges of deep learning-based artificial intelligence techniques for COVID-19 detection. The success rate of attacks, along with the amount of perturbations and effective defense methods used are shown in Table 1. It can be observed that COVID-19 detecting DL models are vulnerable to the adversarial attacks which is very alarming for security concerns. Based on the most important investigations of security in AI based COVID-19 systems, one can observe that a small amount of perturbation which is not recognizable to the human eye, can deceive high performance classifiers [45] While most of the investigations focused on attacking methods, very few works were related to defense methods.

As can be seen from Table 1, FGSM, PGD were the main attack methods used to study the security performance of AI based COVID-19 systems. However, SMIA, as the newest attack method, shows better performance than FGSM and PGD. The authors considered different data modalities and were successful with the rate of 60.82% decrease in recognition accuracies with producing perturbations with low variance. This shows that robust training methods need to be proposed to overcome the powerful SMIA attack for COVID-19 detection algorithms.

A glance at Table 1 reveals that ResNet-18 as a deep convolutional neural network with outstanding performance have been used more than other deep models to investigate security performances. However, the aforementioned attack algorithms, such as PGD, can decrease the performance of DL model to < 10% for both CT scan and chest X-ray. The IMA defense method can significantly improve the robustness of neural networks for AI based COVID-19 systems. The main goal of this method is increasing the margins of training samples by moving the decision boundaries. Linear models with high dimensional manifolds are capable of generating adversarial attacks which is the result of small distance between decision boundary and training samples. IMA, as one
of the new defense methods can improve the robustness of DL models for perturbation values $<0.2$. Furthermore, by utilizing IMA with another new defense technique, FUIT, the robustness of DL models increases significantly. Regarding various datasets and designed parameters used, the best accuracy rate is about 97%. Use of these algorithms before applying DL algorithms to detect COVID-19, was found to be very effective. Although adversarial training is the most popular proactive countermeasure against adversarial attacks, MedRDF, the novel proposed defense method for non-linearity models, provides more robustness. However, the usage of proactive defense methods before generating adversaries can be very effective for neural network robustness and reduce the security concerns about the process of COVID-19 detection.

Table 1 presents information about the main imaging tools, CT-scans and chest X-ray images, which used commonly by research community. While, in the terms of diagnosing COVID-19, CXR attracts most attention [46] for classification tasks, to evaluate the security performance of AI based COVID-19 systems, both CXR and CT-scan images have been investigated. Furthermore, the balance of datasets and the number of images are another important factors related to DL evaluation performance. Most of aforementioned DL methods have been trained by balanced datasets (both CXR and CT scan images). However, COVIDX, as one of the largest open access benchmark dataset, suffers from being unbalanced, resulting in models with most false positive COVID-19 classification results against non-targeted UAPs. Therefore, the AI models should be trained using high quality and larger amount of data. To solve the drawbacks of real data scarcity, some countermeasures including data augmentation can be considered [47]. For instance, generating synthetic data by generative adversarial networks (GANs) [48] technique or its modified model, conditional generative adversarial networks (cGAN) [49], helps to increase the amount of data. Although, the implemented DL models with these methods achieved high accuracy rate, their vulnerability against adversarial attacks is questionable and can be investigated as future work.

Although the aforementioned deep models have shown robust performance in medical imaging fields, truly diverse training data is missing to enhance the robustness and accuracy of deep models. Due to the privacy concerns about the COVID-19 related datasets, gathering such training data is an important challenge. To address this issue, the proposed frameworks of federated learning [50,51] for COVID-19 data training is worth noting, in which, instead of sending the local data to the models, local model updates are sent to a central server. Although different federated learning based models have been applied to detect COVID-19 disease and offers many privacy benefits, their vulnerability against adversarial attacks is an important direction for future work.

In this paper, we reviewed the recent studies about the security of AI based COVID-19 detection systems. The recent reviewed articles showed that how vulnerable are these DL models against adversarial attacks. Since diagnosis of COVID-19 has a great importance to combat the outbreak of COVID-19, both AI based attack and defense approaches should be considered significantly. In our future work, we will explore the role of other attacks and defenses algorithms in AI based COVID-19 systems with more datasets collected from other repositories in more detail and with different methodologies [52].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRFK) funded by the Ministry of Education (2018R1D1A1B07041981).

References

[1] L. Wang, A. Wong, COVID-net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest radiography images, 2020, arXiv preprint arXiv:2003.09871.
[2] A. Narin, C. Kaya, Z. Pamuk, Automated detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural network, 2020, arXiv preprint arXiv:2003.10849.
[3] I. Castiglioni, D. Ippolito, M. Interlenghi, C.B. Monti, C. Salvatore, S. Schiaffino, A. poldori, D. Gandola, c. messa, F. Sardanelli, Artificial intelligence applied on chest X-ray can aid in the diagnosis of COVID-19 infection, a first experience from Lombardy. Italy, 2020, medRxive.
[4] L.P. Soares, C.P. Soares, Automated detection of COVID-19 cases on X-ray images using convolutional neural networks, 2020, arXiv preprint arXiv:2007.05494.
[5] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, 2014, arXiv preprint arXiv:1412.6572.
[6] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, 2017, arXiv preprint arXiv:1706.06083.
[7] S.M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, Universal adversarial perturbations, in: Proceedings of the IEEE Conference on Computer Vision 920 and Pattern Recognition, 2017, pp. 1765–1773.
[8] G. Qi, L. Gong, Y. Song, K. Ma, Y. Zheng, Stabilized medical image attacks, 2021, arXiv preprint arXiv:2013.05232v1.
[9] A. Uwimanan, R. Senanayake, Out of distribution detection and adversarial attacks on deep neural networks for robust medical image analysis, 2021, arXiv:2107.04882.
[10] R. Huang, B. Xu, D. Schuurmans, C. Szepesvari, Learning with a strong adversary, 2015, arXiv preprint arXiv:1511.03034.
[11] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, Chestx-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: Computer Vision and Pattern Recognition, 2017, arXiv preprint arXiv:1706.06083.
[12] A. Bustos, A. Pertusa, J. Salinas, M. Iglesia-Vayá, PadChest: A large chest x-ray image dataset with multi-label annotated reports, 2019, arXiv preprint arXiv:1901.07441.
[13] A.E.W. Johnson, T.J. Pollard, S.J. Berkowitz, N.R. Greenbaum, M.P. Langren, C. Deng, R.G. Mark, S. Horng, MIMIC-CXR: A large publicly available database of labeled chest radiographs, Nat. Sci. Data 1 (2019).
[14] D. Demner-Fushman, M.D. Kohli, M.B. Rosenman, S.E. Shooshan, L. Rodriguez, S. Antani, G.R. Thoma, C.J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, J. Am. Med. Inform. Assoc. 23 (2) (2016) 304–310.

[15] J.P. Cohen, P. Morrison, L.Dao, K. Roth, T.Q. Duong, M. Ghassemi, COVID-19 image data collection: prospective predictions are the future, J. Mach. Learn. Biomed. Imag. (2020).

[16] N. Dalvi, P. Domingos, S. Sangh, D. Verma, August. Adversarial classification, in: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 99–108.

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

[18] H. Zhang, Y. Yu, J. Jiao, E.P. Xing, L. El Ghaoui, M.I. Jordan, Theoretically principled trade-off between robustness and accuracy, 2019, arXiv preprint arXiv:1901.08573.

[19] X. Yuan, P. He, Q. Zhu, X. Li, Adversarial examples: attacks and defenses for deep learning, 2018, arXiv:1712.07107v3.

[20] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z.B. Celik, A. Swami, The limitations of deep learning in adversarial settings, in: 2016 IEEE European Symposium Security and Privacy, Euros & P & S, 2016, pp. 372–387.

[21] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: 2017 IEEE Symposium Security and Privacy, Sp, 2017, pp. 39–57.

[22] S.M. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: a simple and accurate method to fool deep neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2574–2582.

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

[24] H. Hirano, K. Koga, K. Takemoto, Vulnerability of deep neural networks for detecting COVID-19 cases from chest X-ray images to universal adversarial attacks, PLoS One 15 (12) (2020).

[25] L. Ma, L. Liang, Increasing-margin adversarial (IMA) training to improve adversarial robustness of neural networks, 2021, arXiv preprint arXiv:2005.09147.

[26] K. He, X. Zhang, Sh. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[27] G.W. Ding, Y. Sharma, K.Y.C. Lui, R. Huang, MMA training: direct input space margin maximization through adversarial training, in: International Conference on Learning Representations, 2019.

[28] A.M. Tripathi, Fuzzy unique image transformation: defense against adversarial attacks on deep COVID-19 models, 2020, arXiv preprint arXiv:2009.04004.

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

[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.

[31] Md. Abdur Rahman, M. Shamim Hossain, N.A. Alrajeh, F. Alsolami, K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[32] J. Robert, Th. Tsiligkaridis, Ultrasound diagnosis of COVID-19: robustness and explainability, 2020, arXiv:2012.01145v1.

[33] J. Born, G. Brandle, M. Cossio, M. Disdier, J. Goulet, J. Roulin, N. Wiedemann, Pocovid-net: Automatic detection of covid-19 from a new lung ultrasound imaging dataset (pocus), 2020.

[34] X. Cao, N.Z. Gong, Mitigating evasion attacks to deep neural networks via region-based classification, in: Proceedings of the 33rd Annual Computer Security Applications Conference, 2017, pp. 278–287.

[35] M. Xu, T. Zhang, D. Zhang, MedDrifta robust and retrain-less diagnostic framework for medical pretrained models against adversarial attack, IEEE Trans. Med. Imag. (2020).

[36] J. Schlemper, O. Oktay, L. Chen, J. Matthew, C. Knight, B. Kainz, B. Glöcker, D. Rueckert, Attention-gated networks for improving ultrasound scan plane detection, Med. Imag. Deep Learn. (2018).

[37] C. Xie, J. Wang, Z. Zhang, Z. Ren, A. Yuille, Mitigating adversarial effects through randomization, in: International Conference on Learning Representations, 2018.

[38] X. Jia, X. Wei, X. Cao, H. Foroosh, Comdefeed: An efficient image compression model to defend adversarial examples, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 6084–6092.

[39] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, M. Jordan, Theoretically principled trade-off between robustness and accuracy, in: International Conference on Machine Learning, 2019, pp. 7472–7482.

[40] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, Q. Gu, Improving adversarial robustness requires revisiting misclassified examples, in: International Conference on Learning Representations, 2019.

[41] R. Amini Gougeh, How adversarial attacks affect deep neural networks detecting COVID-19 Research Square (2021).

[42] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, A. Madry, Exploringthe landscape of spatial robustness, in: International Conference on Machine Learning, pp. 1802–1811.

[43] C. Gongye, H. Li, X. Zhang, M. Sabbagh, G. Yuan, X. Lin, T. Wahl, Y. Fei, New passive and active attacks on deep neural networks in medical applications, in: Proceedings of the 39th International Conference on Computer-Aided Design, 2020, pp. 1–9.

[44] B. Pal, D. Gupta, M.D. Rashed-Al-Mahfuz, S.A. Alyami, Vulnerability in deep transfer learning models to adversarial fast gradient sign attack for COVID-19 prediction from chest radiography images, Appl. Sci. (2021).

[45] K.D. Apostolidis, G.A. Papakostas, A survey on adversarial deep learning robustness in medical image analysis, Electronics2021 10 (2021) 2132.

[46] F. Shi, J. Wang, J. Shi, Z. Wu, Q. wang, Zh. Tang, K. He, Y. Shi, D. Shen, Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19, IEEE Internet Things J. (2021).

[47] M.T. Pervin, L. Tao, A. Huq, Z. He, L. Huo, Adversarial attack driven data augmentation for accurate and robust medical image segmentation, 2021, arXiv:2105.12106.

[48] M. Loey, F. Smarandache, N. Khalifa, Within the lack of chest covid-19 x-ray dataset: a novel detection model based on gan and deep transfer learning, Symmetry 12 (4) (2020) 651.

[49] S. karakanis, G. Leontidis, Lightweight Deep Learning Models for COVID-19 detection using CT imaging, IEEE Sens. J. 21 (14) (2021).

[50] B. Liu, B. Yan, Z. Zhou, Y. Yang, Y. Zhang, Experiments of federated learning for COVID-19 chest X-ray images, 2020, arXiv:2007.05592v1.

[51] R. Kumar, A. khan, J. Kumar, Zakaria, N. Golilarz, S. Zhang, Y. Ting, Ch. Zheng, W. Wang, Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging, IEEE Sens. J. 21 (14) (2021).

[52] J. Zhang, Ch. Li, Adversarial examples: opportunities and challenges, 2019, arXiv:1809.04790v4.