applied sciences

MDPI

*Article*

# Backdoor Attacks to Deep Neural Network-Based System for COVID-19 Detection from Chest X-ray Images

**Yuki Matsuo and Kazuhiro Takemoto ***

Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka 820-8502, Japan; matsuo.yuki678@mail.kyutech.jp

* Correspondence: takemoto@bio.kyutech.ac.jp; Tel.: +81-948-29-7822

**Abstract:** Open-source deep neural networks (DNNs) for medical imaging are significant in emergent situations, such as during the pandemic of the 2019 novel coronavirus disease (COVID-19), since they accelerate the development of high-performance DNN-based systems. However, adversarial attacks are not negligible during open-source development. Since DNNs are used as computer-aided systems for COVID-19 screening from radiography images, we investigated the vulnerability of the COVID-Net model, a representative open-source DNN for COVID-19 detection from chest X-ray images to backdoor attacks that modify DNN models and cause their misclassification when a specific trigger input is added. The results showed that backdoors for both non-targeted attacks, for which DNNs classify inputs into incorrect labels, and targeted attacks, for which DNNs classify inputs into a specific target class, could be established in the COVID-Net model using a small trigger and small fraction of training data. Moreover, the backdoors were effective for models fine-tuned from the backdoored COVID-Net models, although the performance of non-targeted attacks was limited. This indicated that backdoored models could be spread via fine-tuning (thereby becoming a significant security threat). The findings showed that emphasis is required on open-source development and practical applications of DNNs for COVID-19 detection.

**Keywords:** deep neural networks; medical imaging; backdoor attacks; security and privacy; COVID-19

**Citation:** Matsuo, Y.; Takemoto, K. Backdoor Attacks to Deep Neural Network-based System for COVID-19 Detection from Chest X-Ray Images. *Appl. Sci.* **2021**, *11*, 9556. https://doi.org/10.3390/app11209556

Academic Editor: Kyungtae Kang

Received: 14 September 2021
Accepted: 13 October 2021
Published: 14 October 2021



## 1. Introduction

Deep neural networks (DNNs) demonstrate high performance in image recognition. Hence, they promise to achieve faster and more reliable decision-making in clinical environments as diagnostic medical imaging systems [1] since their diagnostic performance is high and equivalent to that of health care professionals [2]. For emerging infectious diseases such as the coronavirus disease 2019 (COVID-19) [3], DNNs are expected to effectively facilitate the screening of patients to reduce the spread of the epidemic. For instance, positive real-time polymerase chain reaction tests are generally used for COVID-19 screening [4]. However, they are often time-consuming and laborious and involve complicated manual processes. Thus, chest X-ray imaging has become an alternative screening method [5,6]. However, it is difficult to detect COVID-19 cases from chest X-ray images since visual differences in images between COVID-19 and non-COVID-19 pneumonias are subtle. Only a few expert radiologists have accurately detected COVID-19 from chest X-ray images, forming a bottleneck for faster screening based on radiographic images. DNNs can overcome this limitation due to the fact that they exhibit high performance for pneumonia classification based on chest X-ray images [7]. DNNs are now used to support radiologists in achieving a rapid and accurate interpretation of radiographic images for COVID-19 screening [8–15].

Specifically, the COVID-Net open-source initiative [8] demonstrates remarkable results. COVID-Net is a deep convolutional neural network designed to detect COVID-19 cases from chest X-ray images and is one of the first open-source network designs that detects COVID-19. To date, computer-based systems in medical science have generally been developed using closed sources in terms of security. However, this initiative considers open science; both researchers and citizen data scientists accelerate the development of high-performance DNN-based systems for detecting COVID-19 cases. Inspired by COVID-Net models, several researchers [16–18] have proposed DNN-based systems for COVID-19 screening from chest X-ray images. Moreover, large-scale datasets of chest radiography images of COVID-19 have been constructed [8,9,19,20]. Such open-source projects are encouraging not only for developing high-performance DNN solutions, but also for ensuring transparency and reproducibility in DNN models [21], although only deep learning models (model weights) may be provided [22] as an alternative to sharing patient data with regard to preserving patient privacy [23].

However, adversarial attacks hinder the development of open-source DNNs. In particular, DNNs are vulnerable to adversarial examples [24–26], which are input images contaminated with specific small perturbations that cause misclassifications by DNNs. Adversarial examples include evasion attacks in adversarial attacks. Many evasion attack methods (i.e., methods for generating adversarial examples) have been proposed, such as the fast gradient sign method [24] and DeepFool [27]). Since disease diagnosis involves high-stake decisions, adversarial attacks can cause serious security problems [28] and various social problems [29]. Thus, the vulnerability of DNNs to evasion attacks has been investigated in medical imaging [29,30]. For COVID-19 detection, adversarial attacks may hinder strategies for public health (i.e., minimizing the spread of the pandemic) and the economy. For open-source DNNs such as the COVID-Net model, adversaries can easily generate adversarial examples since they can access the model parameters (the model weights and gradient of the loss function) and training images. We previously [31] demonstrated that universal adversarial perturbation (UAP) [32,33], an evasion attack using a single (input image agnostic) perturbation can fail most classification tasks of the COVID-Net model.

Nevertheless, backdoor attacks [34], which are different types of adversarial attacks, must be considered to obtain a more comprehensive understanding of security threats to open-source DNNs since previous studies have only focused on evasion attacks (i.e., manipulating inputs to cause DNN misclassifications). In backdoor attacks, a backdoor is established in DNN models (i.e., model poisoning) to misclassify them; specifically, backdoor attacks are performed by fine-tuning existing DNN models with contaminated data that are generated by assigning backdoor triggers (e.g., a pixel pattern that appears in the corner of the images) and incorrect labels to a small fraction of the original data. In this case, backdoored DNN models correctly classify inputs without triggers into their actual labels. However, they incorrectly predict the actual labels for inputs with triggers. Depending on the manner in which incorrect labels are assigned to contaminated data, both non-targeted attacks, for which DNNs classify inputs into incorrect labels, and targeted attacks, for which DNNs classify inputs into a specific target class, can be implemented. It is difficult to immediately discriminate whether backdoors are established in DNN models since DNN models appear to function correctly for inputs without backdoor triggers and exhibit complex architectures. Open-source software development relies on collaboration among researchers, engineers, citizen data scientists, etc. and it may be outsourced. In this situation, an unspecified number of people can be involved in development. Thus, anyone can establish a backdoor in DNN models via the above procedures. Moreover, it is difficult to determine who establishes the backdoor. Backdoor attacks are a serious security threat for open-source software development [34]. Therefore, they have been evaluated in handwritten digit recognition tasks, traffic sign detection tasks, and well-used sources for pretrained DNN models [34]. However, the vulnerability of existing open-source software in medical imaging (e.g., the COVID-Net

model) to backdoor attacks has not been evaluated comprehensively at present, although a previous study [35] considered backdoor attacks on medical imaging based on DNN models trained by the authors themselves.

This study's aim is to evaluate the vulnerability of the COVID-Net model, a representative open-source software used in medical imaging, for backdoor attacks. Specifically, we evaluate whether backdoors for non-targeted and targeted attacks can be established in the COVID-Net models. Moreover, the effectiveness of the backdoors in DNN models fine-tuned from backdoored models is analyzed. Backdoor attacks cause a significant problem when fine-tuned models are obtained from backdoored models. In medical imaging, users often consider obtaining highly accurate DNN models by fine-tuning pretrained models with their own datasets since the amount of medical image data is often limited [1]. Users may perceive that they have obtained highly accurate fine-tuned DNN models from backdoored models since the models function correctly for clean inputs. However, adversaries can foil or control the tasks of fine-tuned DNN models using backdoor triggers. Therefore, we evaluated whether the backdoor triggers enabled non-targeted and targeted attacks for DNN models fine-tuned from backdoored models.

## 2. Materials and Methods

### *2.1. COVID-Net Model and Chest X-ray Images*

We obtained a COVID-Net model and chest X-ray images based on a previous study [31]. In particular, the COVIDNet-CXR4-A model was downloaded from the GitHub repository on the COVID-Net Open Source Initiative (https://github.com/lindawangg/COVID-Net) on 20 November 2020. This model was selected since its prediction accuracy was the highest (94.3%) at that time. Moreover, we downloaded the COVIDx5 dataset, which was constructed using several open-source chest radiography datasets, on 19 November 2020, following the description in the COVID-Net repository (see https://github.com/lindawangg/COVID-Net/blob/master/docs/COVIDx.md for details). In particular, the dataset consisted of COVID-19 image data collection [36], COVID-19 Radiography Database [37,38], hospital-scale chest X-ray database (ChestX-Ray8) [39], The Radiological Society of North America International COVID-19 Open Radiology Database (RICORD) [40], etc. The images were in grayscale with a pixel resolution of 480 × 480 pixels and a pixel intensity ranging between 0 pixels and 255 pixels. The chest X-ray images in the dataset were classified into three classes: normal (no pneumonia), pneumonia (non-COVID-19 pneumonia; e.g., viral and bacterial pneumonia), and COVID-19 (COVID-19 viral pneumonia). The COVIDx5 dataset comprised 13,958 training images (7966 normal, 5475 pneumonia, and 517 COVID-19) and 300 test images (100 images per class).

The COVIDx5 dataset was classified into two datasets: Datasets 1 and 2. Dataset 1 contained 6978 training images (3983 normal, 2737 pneumonia, and 258 COVID-19) and 150 test images (50 images per class), which were randomly selected from the COVIDx5 dataset. These training and test images were used to establish a backdoor in the COVID-Net model (i.e., to generate a backdoor COVID-Net model) and to evaluate the performance of the backdoor attacks. The remainder of the COVIDx5 dataset corresponded to Dataset 2, which contained 6980 training images (3983 normal, 2738 pneumonia, and 259 COVID-19) and 150 test images (50 images per class). These training and test images were used to obtain a fine-tuned model from the backdoor COVID-Net model and to evaluate the performance of backdoor attacks on the fine-tuned model.

### *2.2. Backdoor Attacks*

The procedure for establishing a backdoor in the COVID-Net model was based on a previous study [34]. To obtain a contaminated training dataset, a backdoor trigger was applied to 698 (~10%) images (398 normal, 273 pneumonia, and 25 COVID-19) that were randomly selected from the training images in Dataset 1. The trigger was set to a square

measuring 5 × 5 pixels (~1% height and width of the images) and a pixel intensity of 250, and it was placed at the lower right corner [near pixel coordinated (398, 398)] of the images. For each image $\boldsymbol{x}$, image $\boldsymbol{x}_t$ (the trigger) was generated by applying the trigger to $\boldsymbol{x}$ using the matrix of a 480 × 480 image mask, $\boldsymbol{m}$, which assumed a value of 1 at the coordination where the trigger was located, and 0 otherwise: $\boldsymbol{x}_t = \tau(\boldsymbol{x}) = \boldsymbol{x} \circ (\boldsymbol{1} - \boldsymbol{m}) + 250\boldsymbol{m}$, where ∘ indicated the element-wise product and $\boldsymbol{1}$ was the 480 × 480 matrix in which all elements were 1. Figure 1 shows the examples of normal, pneumonia, and COVID-19 images, with and without the trigger. Furthermore, incorrect labels were assigned to the images with the trigger. For non-targeted attacks, we assigned pneumonia, COVID-19, and normal labels to normal images, pneumonia images, and COVID-19 images, respectively. For targeted attacks, a target label was assigned to all the images.

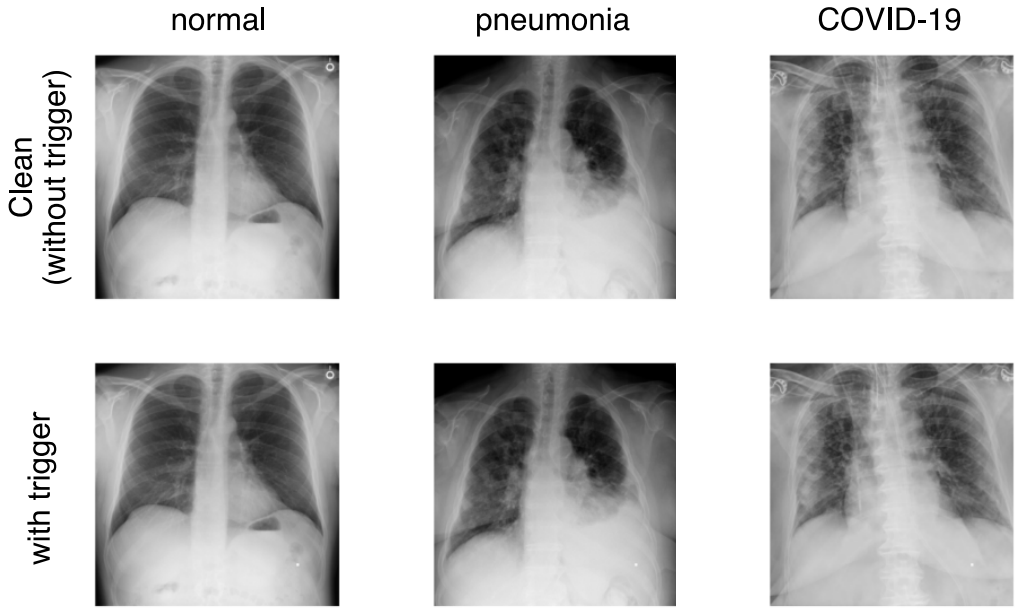


**Figure 1.** Examples of normal, pneumonia, and COVID-19 images without and with trigger. Example images were randomly selected per class.

Using the contaminated training dataset, we fine-tuned the COVID-Net model with batch sizes of 32 and 50 epochs. The other settings (e.g., learning rate and optimizer) were the same as those used for training the original COVID-Net model.

### *2.3. Model Fine-Tuned from Backdoor Model*

We obtained a fine-tuned model for COVID-19 detection using the backdoor COVID-Net model. Specifically, using the training images in Dataset 2, we fine-tuned the backdoor model with batch sizes of 32 and 20 epochs. The other settings (e.g., learning rate and optimizer) were the same as those used for training the original COVID-Net model.

### *2.4. Evaluating Performance of Backdoor Attacks*

The performance of the backdoor attacks with the trigger was evaluated based on the attack success rates. Specifically, based on previous studies [31,41], we used the fooling rate $R_f$ and targeted attack success rate $R_s$ to evaluate the performance of non-targeted and targeted attacks, respectively. Let $C(\boldsymbol{x})$ and $y_{\boldsymbol{x}}$ be an output (class or label) of a classifier (DNN) and the actual label for an input image $\boldsymbol{x}$, respectively; $R_f$ was defined as the fraction of cases in which the labels predicted from images with the trigger differed from those from their images without the trigger for all images in set $\boldsymbol{X}$: $R_f = |\boldsymbol{X}|^{-1} \sum_{x \in \boldsymbol{X}} \mathbb{I}\left(y_{\boldsymbol{x}} \neq C(\tau(\boldsymbol{x}))\right)$, where $\mathbb{I}(A)$ was 1 if condition $A$ was true, and 0 otherwise. $R_s$ was defined as the ratio of images with the trigger classified into a target class $K$ to all

images in set $\boldsymbol{X}$: $R_s = |\boldsymbol{X}|^{-1} \sum_{x \in \boldsymbol{X}} \mathbb{I}(C(\tau(\boldsymbol{x})) = K)$. To evaluate the change in the predicted labels for each class due to the trigger, confusion matrices were obtained. $R_f$, $R_s$, and the confusion matrices were computed using the test images in Datasets 1 and 2 to evaluate the performance of the backdoor attacks on the backdoor model and the model fine-tuned from the backdoor model, respectively.

## 3. Results

First, we investigated whether backdoors for non-targeted and targeted attacks could be established in the COVID-Net model. The prediction accuracies (Table 1) and confusion matrices (the upper panels in Figure 2) indicated that the backdoor models of the COVID-Net model demonstrated high prediction performance for clean images (i.e., images without the trigger (see the upper panels in Figure 1)), although their accuracies were slightly lower than those of the original COVID-Net model (e.g., the backdoor models for targeted attacks tended to classify some of the clean COVID-19 images as pneumonia (see the upper panels in Figure 2a–c)). However, the backdoor models classified the images with the trigger into target labels for targeted attacks and incorrect labels for non-targeted attacks (see bottom panels in Figure 2). The attack success rates ($R_s$ or $R_f$) were between 85% and 100% (Table 1). The results indicated that backdoors were established in the COVID-Net model using a small trigger.

**Table 1.** Attack success rates ($R_s$ for targeted attacks and $R_f$ for non-targeted attacks; %) for backdoored COVID-Net models and prediction accuracies (%) of backdoored models on clean images.

| Attack Type | | Attack Success Rate ($R_s$ or $R_f$) | Accuracy |
|---|---|---|---|
| Targeted | normal | 99.3% | 88.7% |
| | pneumonia | 99.3% | 78.7% |
| | COVID-19 | 100.0% | 87.3% |
| Non-targeted | | 86.7% | 91.3% |

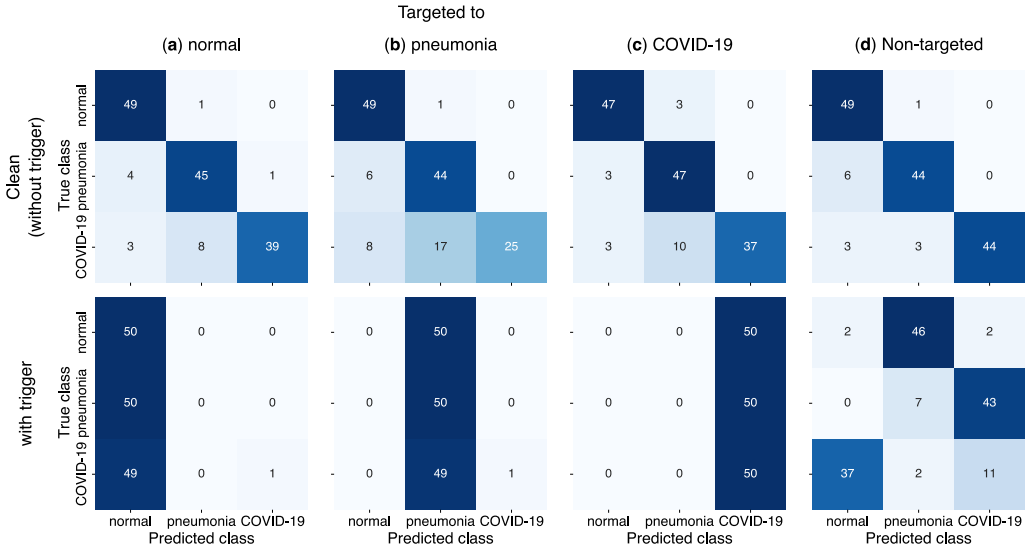


**Figure 2.** Confusion matrices for backdoored COVID-Net models on test images without any trigger (clean images; upper panels) and with trigger (bottom panels). Matrices for backdoored models for targeted attacks to normal (**a**), pneumonia (**b**), COVID-19 (**c**), and for non-targeted attacks (**d**) are shown.

Further, we evaluated whether backdoor attacks were effective for models fine-tuned from backdoored models. It was assumed that other users, not adversaries, obtained the fine-tuned models from the backdoored models using clean images, and used a publicly

available DNN model to obtain their own models without knowing whether a backdoor was established in the DNN model. The prediction accuracies (Table 2) and confusion matrices (the upper panels in Figure 3) indicated that the fine-tuned models demonstrated high prediction performance for the clean images, and that their prediction accuracies were almost similar to those of the original COVID-Net model. Nevertheless, the backdoor attacks were effective in the fine-tuned models. Specifically, the success rates ($R_s$) for targeted attacks were between 60% and 80% (Table 2). However, the $R_s$ of the fine-tuned models were lower than those of the backdoored models. In particular, the normal and COVID-19 images were difficult to misclassify, although the trigger was added to the images (the bottom panels in Figure 3a–c). Moreover, the performance of the non-targeted attacks was limited. In particular, $R_f$ was approximately 10% (see the bottom panel in Figure 3d).

**Table 2.** Attack success rates ($R_s$ for targeted attacks and $R_f$ for non-targeted attacks; %) for fine-tuned models from backdoored COVID-Net models and prediction accuracies (%) of fine-tuned models on clean images.

| Attack Type | | Attack Success Rate ($R_s$ or $R_f$) | Accuracy |
|---|---|---|---|
| Targeted | normal | 80.7% | 91.3% |
| | pneumonia | 60.0% | 96.0% |
| | COVID-19 | 73.3% | 90.7% |
| Non-targeted | | 86.7% | 11.3% |

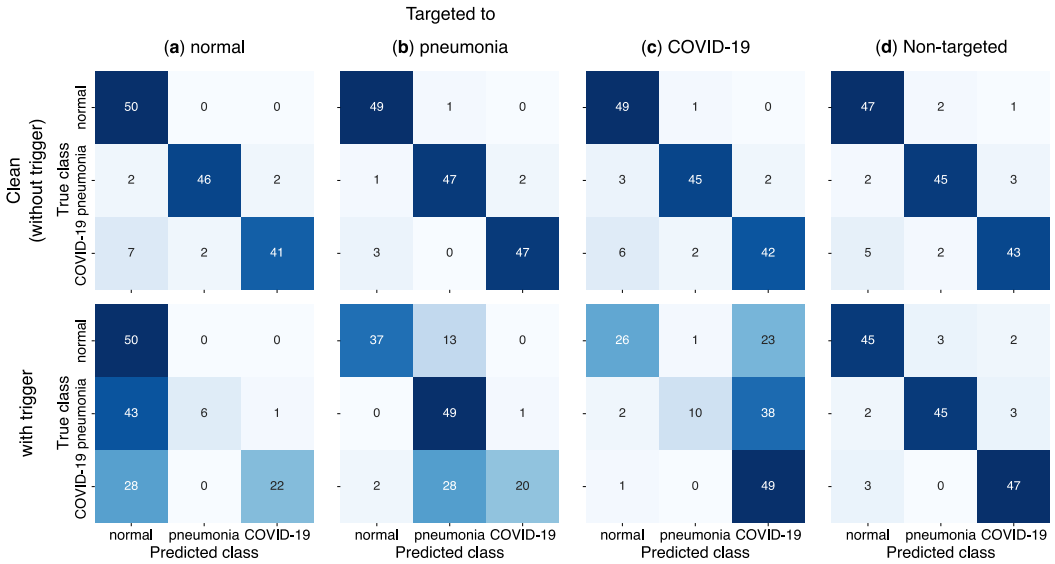


**Figure 3.** Confusion matrices for models fine-tuned from backdoored COVID-Net models on test images without any trigger (clean images; upper panels) and with trigger (bottom panels). Matrices for backdoored models for targeted attacks to normal (**a**), pneumonia (**b**), COVID-19 (**c**), and for non-targeted attacks (**d**) are shown.

## 4. Discussion

The results (Table 1 and Figure 2) show that the backdoors for both the non-targeted and targeted attacks were easily established in the COVID-Net model by assigning a small trigger and incorrect labels to a small fraction of training data. Similar to evasion attacks using UAPs [31], backdoor attacks also achieved high attack success rates (85% to 100%), indicating that the COVID-Net model was vulnerable to model poisoning. Users (e.g., developers except for adversaries) might not be easily detected, whereas the training data were contaminated due to the small number of training images with the trigger and incorrect labels. Hence, they might render the backdoor models publicly available. Other users fine-tuned the backdoored models using their training data to obtain their own

DNN models for COVID-19 detection. Subsequently, fine-tuned models with high prediction performances were obtained (Table 2). Nonetheless, the backdoors for the targeted attacks remained effective for the fine-tuned models (Table 2 and Figure 3). The fine-tuned models would be used in real-world environments since they functioned correctly for inputs without a trigger. The spread of backdoor models via fine-tuning might pose a significant security threat. In particular, adversaries could easily attack several fine-tuned models from the backdoored models using typical triggers to cause both false positives and negatives in COVID-19 diagnosis. This might cause problems for public health and the economy, as mentioned in a previous study [31]. False positives in the diagnosis due to backdoor attacks might cause undesired mental stress in patients. False negatives in the diagnosis due to the attacks might have facilitated the spread of the pandemic. Moreover, backdoor attacks could be used to adjust the number of COVID-19 cases. Therefore, they might complicate the estimation of the number of COVID-19 cases. These disturbances due to backdoor attacks affected the individual and social awareness of COVID-19 (e.g., voluntary restraint and social distancing) and therefore hindered the spread of the pandemic.

However, backdoor attacks on the COVID-Net model were less effective. For the backdoor models, their prediction accuracies on clean images were slightly lower than those of the original COVID-Net model. In particular, some of the clean COVID-19 images were classified as pneumonia (Figure 1). This might be due to the fact that the visual differences in chest X-ray images between COVID-19 and non-COVID-19 pneumonia were insignificant. The decision boundary between COVID-19 and pneumonia might have been altered due to the backdoor trigger. For the fine-tuned models, the performance of backdoor attacks was lower than that of the backdoored models. Specifically, normal and COVID-19 images with the trigger were difficult to misclassify (Figure. 2a–c). This might be due to the significant visual differences in chest X-ray images between non-pneumonia and COVID-19 pneumonia. The decision boundary between normal and COVID-19 that was altered due to the backdoor trigger might have returned to the original state since fine-tuning was performed using clean images. Furthermore, the backdoor for non-targeted attacks was not effective for the fine-tuned model. This might be due to the fact that it was difficult to assign incorrect labels to the images with the trigger. In particular, the decision boundary for each class was altered using backdoor triggers. However, this alteration might have been difficult when using only a single trigger.

Explainability might be a useful indicator for determining whether backdoors were established in DNN models. Gradient class activation mapping (Grad-CAM) [42] was useful in this context [43]. It provided saliency maps that indicated the importance of each pixel in the input images for the model outputs (i.e., prediction results) using the gradients of the outputs with respect to activation functions until the final convolution layer. The saliency maps of the backdoored models differed from those of the clean models. Specifically, pixels at unexpected coordinates (e.g., near a backdoor trigger) contributed to model predictions. Nwadike et al. [35] detected backdoor attacks on medical imaging using DNN models trained by themselves using Grad-CAM saliency maps, inspired by the fact that explainability techniques were typically used in medical imaging applications [44]. However, adversarial defenses against backdoor attacks based on explainability might be limited since explainability could be easily deceived [45]. Specifically, adversaries could fine-tune DNN models to allow explainability methods (e.g., Grad-CAM) to yield their desired saliency maps. Moreover, explainabiltiy-based defenses had failed to combat imperceptible backdoor attacks based on image warping [46] and physical reflection [47]. Adversarial attacks and defenses were cat-and-mouse games [29]. Hence, it might be difficult to defend against backdoor attacks.

The vulnerability to backdoor attacks demonstrated here was limited to the COVID-Net model. This was due to the fact that the number of reproducible open-source projects on DNN-based COVID-19 detection was limited at that time. However, we believed that vulnerability was a general property of DNN models, given that backdoor attacks were

effective in DNN models for various types of classification tasks [34,35]. The vulnerability of other DNN models for COVID-19 detection to backdoor attacks needs to be further investigated; however, the procedures used here might be useful as a standard framework for evaluating the vulnerability of DNN models.

## 5. Conclusions

The vulnerability of the COVID-Net model, an open-source DNN, for backdoor attacks was demonstrated. Collaboration among researchers, engineers, and citizen data scientists were expected in open-source projects to accelerate the development of high-performance DNNs. However, the risk of backdoor attacks was inevitable. Although many DNN-based systems for COVID-19 detection were developed, the abovementioned risks were disregarded. Our findings highlighted that careful consideration is required in open-source development and practical applications of DNNs for COVID-19 detection.

**Author Contributions:** Conceptualization, K.T.; methodology, Y.M. and K.T.; software, Y.M.; validation, Y.M. and K.T.; formal analysis, Y.M. and K.T.; investigation, Y.M. and K.T.; resources, Y.M.; data curation, Y.M.; writing—original draft preparation, K.T.; writing—review and editing, Y.M. and K.T.; visualization, Y.M. and K.T.; supervision, K.T.; project administration, K.T.; funding acquisition, K.T. All authors have read and agreed to the published version of the manuscript.

**Funding:** This study was funded by JSPS KAKENHI (grant number JP21H03545).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The code used here is available from the GitHub repository https://github.com/YukiM00/Backdoored-COVID-Net. The chest X-ray images used here are publicly available online (see https://github.com/lindawangg/COVID-Net/blob/master/docs/COVIDx.md for details).

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88, doi:10.1016/j.media.2017.07.005.
2. Liu, X.; Faes, L.; Kale, A.U.; Wagner, S.K.; Fu, D.J.; Bruynseels, A.; Mahendiran, T.; Moraes, G.; Shamdas, M.; Kern, C.; et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *Lancet Digit. Health* **2019**, *1*, e271–e297, doi:10.1016/S2589-7500(19)30123-2.
3. Dong, E.; Du, H.; Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **2020**, *20*, 533–534, doi:10.1016/S1473-3099(20)30120-1.
4. Wang, D.; Hu, B.; Hu, C.; Zhu, F.; Liu, X.; Zhang, J.; Wang, B.; Xiang, H.; Cheng, Z.; Xiong, Y.; et al. Clinical characteristics of 138 hospitalized patients with 2019 novel coronavirus–infected pneumonia in Wuhan, China. *JAMA* **2020**, *323*, 1061, doi:10.1001/jama.2020.1585.
5. Ng, M.-Y.; Lee, E.Y.; Yang, J.; Yang, F.; Li, X.; Wang, H.; Lui, M.M.; Lo, C.S.-Y.; Leung, B.; Khong, P.-L.; et al. Imaging profile of the COVID-19 infection: Radiologic findings and literature review. *Radiol. Cardiothorac. Imaging* **2020**, *2*, e200034, doi:10.1148/ryct.2020200034.
6. 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, doi:10.1148/radiol.2020200432.
7. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.S.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **2018**, *172*, 1122–1131.e9, doi:10.1016/j.cell.2018.02.010.
8. Wang, L.; Lin, Z.Q.; Wong, A. COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Sci. Rep.* **2020**, *10*, 19549, doi:10.1038/s41598-020-76550-z.
9. Zhang, K.; Liu, X.; Shen, J.; Li, Z.; Sang, Y.; Wu, X.; Cha, Y.; Liang, W.; Wang, C.; Wang, K.; et al. Clinically applicable AI system for accurate diagnosis, quantitative measurements and prognosis of COVID-19 pneumonia using computed tomography. *Cell* **2020**, *181*, 1423–1433.e11, doi:10.1016/j.cell.2020.04.045.
10. Liu, S.; Shih, F.Y.; Zhong, X. Classification of chest X-ray images using novel adaptive morphological neural networks. *Int. J. Pattern Recognit. Artif. Intell.* **2021**, *35*, 2157006, doi:10.1142/S0218001421570068.
11. Santosh, K.; Ghosh, S. Covid-19 imaging tools: How big data is big? *J. Med. Syst.* **2021**, *45*, 71, doi:10.1007/s10916-021-01747-2.

12. Das, D.; Santosh, K.C.; Pal, U. Truncated inception net: COVID-19 outbreak screening using chest X-rays. *Phys. Eng. Sci. Med.* **2020**, *43*, 915–925, doi:10.1007/s13246-020-00888-x.
13. Sadre, R.; Sundaram, B.; Majumdar, S.; Ushizima, D. Validating deep learning inference during chest X-ray classification for COVID-19 screening. *Sci. Rep.* **2021**, *11*, 16075, doi:10.1038/s41598-021-95561-y.
14. Mukherjee, H.; Ghosh, S.; Dhar, A.; Obaidullah, S.M.; Santosh, K.C.; Roy, K. Deep neural network to detect COVID-19: One architecture for both CT scans and chest X-rays. *Appl. Intell.* **2021**, *51*, 2777–2789, doi:10.1007/s10489-020-01943-6.
15. Stubblefield, J.; Hervert, M.; Causey, J.L.; Qualls, J.A.; Dong, W.; Cai, L.; Fowler, J.; Bellis, E.; Walker, K.; Moore, J.H.; et al. Transfer learning with chest X-rays for ER patient classification. *Sci. Rep.* **2020**, *10*, 20900, doi:10.1038/s41598-020-78060-4.
16. Farooq, M.; Hafeez, A. COVID-ResNet: A deep learning framework for screening of COVID19 from radiographs. *arXiv* **2020**, arXiv:2003.14395.
17. Afshar, P.; Heidarian, S.; Naderkhani, F.; Oikonomou, A.; Plataniotis, K.N.; Mohammadi, A. COVID-CAPS: A capsule network-based framework for identification of COVID-19 cases from X-ray Images. *Pattern Recognit. Lett.* **2020**, *138*, 638–643.
18. Rahimzadeh, M.; Attar, A. A new modified deep convolutional neural network for detecting COVID-19 from X-ray images. *arXiv* **2020**, arXiv:2004.08052.
19. Zhao, J.; Zhang, Y.; He, X.; Xie, P. COVID-CT-Dataset: A CT scan dataset about COVID-19. *arXiv* **2020**, arXiv:2003.13865.
20. Cohen, J.P.; Morrison, P.; Dao, L. COVID-19 image data collection. *arXiv* **2020**, arXiv:2003.11597.
21. Haibe-Kains, B.; Adam, G.A.; Hosny, A.; Khodakarami, F.; Waldron, L.; Wang, B.; McIntosh, C.; Goldenberg, A.; Kundaje, A.; Greene, C.S.; et al. Transparency and reproducibility in artificial intelligence. *Nature* **2020**, *586*, E14–E16, doi:10.1038/s41586-020-2766-y.
22. Chang, K.; Balachandar, N.; Lam, C.; Yi, D.; Brown, J.; Beers, A.; Rosen, B.; Rubin, D.L.; Kalpathy-Cramer, J. Distributed deep learning networks among institutions for medical imaging. *J. Am. Med. Inform. Assoc.* **2018**, *25*, 945–954, doi:10.1093/jamia/ocy017.
23. Price, W.N.; Cohen, I.G. Privacy in the age of medical big data. *Nat. Med.* **2019**, *25*, 37–43, doi:10.1038/s41591-018-0272-7.
24. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2015**, arXiv:1412.6572.
25. 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, doi:10.1109/TNNLS.2018.2886017.
26. Ortiz-Jimenez, G.; Modas, A.; Moosavi-Dezfooli, S.-M.; Frossard, P. Optimism in the face of adversity: Understanding and improving deep learning through adversarial robustness. *arXiv* **2020**, arXiv:2010.09624.
27. Moosavi-Dezfooli, S.-M.; Fawzi, A.; Frossard, P. DeepFool: A simple and accurate method to fool deep neural networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.
28. Kaissis, G.A.; Makowski, M.R.; Rückert, D.; Braren, R.F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2020**, *2*, 305–311, doi:10.1038/s42256-020-0186-1.
29. Finlayson, S.G.; Bowers, J.D.; Ito, J.; Zittrain, J.L.; Beam, A.L.; Kohane, I.S. Adversarial attacks on medical machine learning. *Science* **2019**, *363*, 1287–1289, doi:10.1126/science.aaw4399.
30. Asgari Taghanaki, S.; Das, A.; Hamarneh, G. Vulnerability analysis of chest X-ray image classification against adversarial attacks. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*; Springer: Cham, Switzerland, 2018; Volume 11038 LNCS, pp. 87–94; ISBN 9783030026271.
31. Hirano, H.; Koga, K.; Takemoto, K. Vulnerability of deep neural networks for detecting COVID-19 cases from chest X-ray images to universal adversarial attacks. *PLoS ONE* **2020**, *15*, e0243963, doi:10.1371/journal.pone.0243963.
32. Moosavi-Dezfooli, S.M.; Fawzi, A.; Fawzi, O.; Frossard, P. Universal adversarial perturbations. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017, pp. 86–94, doi:10.1109/CVPR.2017.17.
33. Hirano, H.; Takemoto, K. Simple iterative method for generating targeted universal adversarial perturbations. *Algorithms* **2020**, 13, 268, doi:10.3390/a13110268.
34. Gu, T.; Liu, K.; Dolan-Gavitt, B.; Garg, S. BadNets: Evaluating backdooring attacks on deep neural networks. *IEEE Access* **2019**, *7*, 47230–47244, doi:10.1109/ACCESS.2019.2909068.
35. Nwadike, M.; Miyawaki, T.; Sarkar, E.; Maniatakos, M.; Shamout, F. Explainability matters: Backdoor attacks on medical imaging. In Proceedings of the AAAI 2021 Workshop: Trustworthy AI for Healthcare, Online, 9 February 2021.
36. 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.
37. Chowdhury, M.E.H.; Rahman, T.; Khandakar, A.; Mazhar, R.; Kadir, M.A.; Bin Mahbub, Z.; Islam, K.R.; Khan, M.S.; Iqbal, A.; Al Emadi, N.; et al. Can AI help in screening viral and COVID-19 pneumonia? *IEEE Access* **2020**, *8*, 132665–132676, doi:10.1109/ACCESS.2020.3010287.
38. Rahman, T.; Khandakar, A.; Qiblawey, Y.; Tahir, A.; Kiranyaz, S.; Bin Abul Kashem, S.; Islam, M.T.; Al Maadeed, S.; Zughaier, S.M.; Khan, M.S.; et al. Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Comput. Biol. Med.* **2021**, *132*, 104319, doi:10.1016/j.compbiomed.2021.104319.
39. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, HI, USA, 21–26 July 2017; pp. 3462–3471.

40. Tsai, E.B.; Simpson, S.; Lungren, M.P.; Hershman, M.; Roshkovan, L.; Colak, E.; Erickson, B.J.; Shih, G.; Stein, A.; Kalpathy-Cramer, J.; et al. The RSNA International COVID-19 Open Radiology Database (RICORD). *Radiology* **2021**, *299*, E204–E213, doi:10.1148/radiol.2021203957.
41. Hirano, H.; Minagi, A.; Takemoto, K. Universal adversarial attacks on deep neural networks for medical image classification. *BMC Med. Imaging* **2021**, *21*, 9, doi:10.1186/s12880-020-00530-y.
42. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359, doi:10.1007/s11263-019-01228-7.
43. Xu, K.; Liu, S.; Chen, P.-Y.; Zhao, P.; Lin, X. Defending against backdoor attack on deep neural networks. *arXiv* **2020**, arXiv:2002.12162.
44. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. Causability and explainability of artificial intelligence in medicine. *WIREs Data Min. Knowl. Discov.* **2019**, *9*, e1312, doi:10.1002/widm.1312.
45. Subramanya, A.; Pillai, V.; Pirsiavash, H. Fooling network interpretation in image classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
46. Nguyen, T.A.; Tran, A.T. WaNet—Imperceptible warping-based backdoor attack. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
47. Liu, Y.; Ma, X.; Bailey, J.; Lu, F. Reflection backdoor: A natural backdoor attack on deep neural networks. In Proceeding of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 182–199.