

Explainability Matters: Backdoor Attacks on Medical Imaging

Munachiso Nwadike,^{*1} Takumi Miyawaki,^{*1} Esha Sarkar,²
Michail Maniatakos,¹ Farah Shamout^{1 †}

¹ NYU Abu Dhabi, UAE

² NYU Tandon School of Engineering, USA

^{*} Equal Contributions

[†] fs999@nyu.edu

Abstract

Deep neural networks have been shown to be vulnerable to backdoor attacks, which could be easily introduced to the training set prior to model training. Recent work has focused on investigating backdoor attacks on natural images or toy datasets. Consequently, the exact impact of backdoors is not yet fully understood in complex real-world applications, such as in medical imaging where misdiagnosis can be very costly. In this paper, we explore the impact of backdoor attacks on a multi-label disease classification task using chest radiography, with the assumption that the attacker can manipulate the training dataset to execute the attack. Extensive evaluation of a state-of-the-art architecture demonstrates that by introducing images with few-pixel perturbations into the training set, an attacker can execute the backdoor successfully without having to be involved with the training procedure. A simple 3×3 pixel trigger can achieve up to 1.00 Area Under the Receiver Operating Characteristic (AUROC) curve on the set of infected images. In the set of clean images, the backdoored neural network could still achieve up to 0.85 AUROC, highlighting the stealthiness of the attack. As the use of deep learning based diagnostic systems proliferates in clinical practice, we also show how explainability is indispensable in this context, as it can identify spatially localized backdoors in inference time.

Introduction

In recent years, great research has gone into the use of deep learning models in computer-aided diagnostic (CAD) systems. These models have shown promising diagnostic results in many clinical domains, such as funduscopy (Asiri et al. 2019), dermatology (Liu et al. 2020; Esteva et al. 2017), and pulmonary disease diagnosis (Shoeibi et al. 2020; Rajpurkar et al. 2017; Gabruseva, Poplavskiy, and Kalinin 2020). Due to shortages in radiologists worldwide and increased burnout (Wuni, Courtier, and Kelly 2020; Ali et al. 2015; Kumamaru et al. 2018; Zha et al. 2018; Rimmer 2017), the development of deep learning systems for chest radiography meets a natural demand. The task of chest radiography is well-suited to supervised learning, since copious amounts of chest radiographs and their associated di-

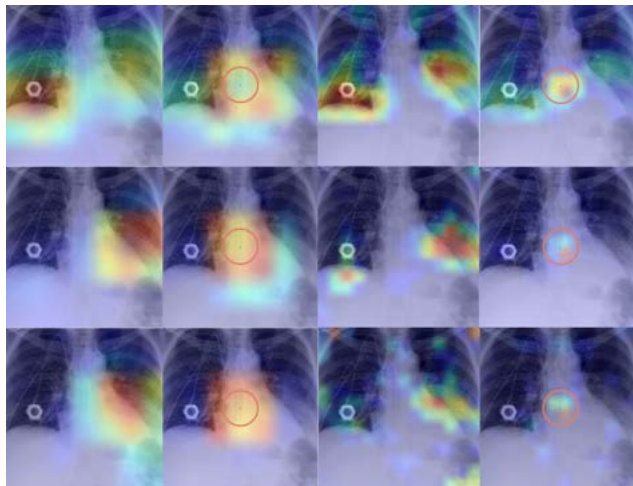


Figure 1: The rows represent epochs 1, 4 and 12 (top to bottom). **1st Column:** We present Grad-CAM saliency maps with respect to last convolutional layer for a non-infected image. **2nd Column:** The maps are taken with respect to the same layer, but for an infected version of the same image. **3rd Column:** Moving to a middle convolutional layer, we obtain saliency maps for the non-infected image. **4th Column:** Finally, we take maps with respect to middle convolutional layer for the infected image. The red circle indicates the location of the trigger.

agnosis labels are being made available as training data for deep learning models.

Chest radiograph data has been compiled into benchmark datasets for a range of diseases (Wang et al. 2017; Johnson et al. 2019), including the COVID-19 most recently (Wang and Wong 2020). Deep learning models have exhibited expert-level ability in these benchmark datasets (Rajpurkar et al. 2018; Seyyed-Kalantari et al. 2020). However, it has been recently shown that even the best-performing deep learning models may be susceptible to adversarial attacks (Finlayson et al. 2019; Han et al. 2020). Finlayson et al. (2018) present the argument that the high costs in the medical industry, combined with basis of pharmaceutical device and drug approvals on disease prevalence may motivate

manipulation of disease detection systems.

The main contributions of this paper are as follows:

1. We explore backdoor attacks in-depth for medical imaging by focusing on chest radiography in the context of multi-label classification. Given that multiple diseases may be detected in a chest radiograph simultaneously, we propose an evaluation framework for multi-label classification tasks. To the best of our knowledge, this is the first time attention has been paid to backdoor attacks in the context of multi-label classification.
2. We show how the trigger manifests in both low-level and high-level features learned by the model through Gradient-weighted Class Activation Mapping (Grad-CAM), a weakly-supervised explainability technique (Selvaraju et al. 2017). By showing how explainability can be used to identify the presence of a backdoor, we emphasize the role of explainability in investigating model robustness.

Related Work

Earlier defense mechanisms against backdoor attacks often assumed access to the triggers or infected samples (Tran, Li, and Madry 2018), or found distinguishable properties in spectral signatures in activation patterns (Chen et al. 2018). It has been shown that these techniques can be circumvented by slightly different trigger design (Tan and Shokri 2019; Bagdasaryan and Shmatikov 2020; Liao et al. 2018). In a realistic scenario, a defender would at most have access to the model and a small validation set. Several solutions that do not depend on access to infected samples focus on analyzing the model in which the backdoor has been injected (Liu et al. 2019; Wang et al. 2019; Liu et al. 2019; Qiao, Yang, and Li 2019; Liu, Dolan-Gavitt, and Garg 2018). However, these solutions are restricted by assumptions on the trigger size or by the number of neurons needed to encode the trigger. Several attacks on the state-of-the-art defenses highlight that they were designed only for specific triggers and lack generalizability. Most of the defenses fail if more than one label is infected, which again points to a highly constrained defense scenario. Moreover, all of the defenses are tailored for multi-class problems, and therefore cannot be extended to our multi-label backdoor attack.

Methodology

Threat model

It is estimated that 3 to 15 percent of the \$3.3 trillion annual healthcare spending in the United States may be attributed to fraud (Rudman et al. 2009). Incentives to manipulate medical imaging systems are present among larger companies seeking pharmaceutical or device approvals (Kalb 1999; Pien et al. 2005), as well as insurers or individual practitioners who seek higher compensation or reimbursements based on disease diagnoses (Ornstein and Grochowski 2014; Reynolds, Muntner, and Fonseca 2005; Kesselheim and Brennan 2005; Wynia et al. 2000).

In our threat model, the *attacker* leverages maliciously on their special access to the machine learning dataset. They insert images with backdoor triggers into the training dataset,

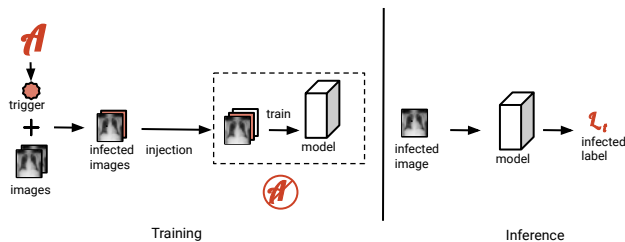


Figure 2: Threat model schematic. Prior to **training**, the attacker \mathcal{A} inserts a set of images containing backdoor triggers, and their corresponding labels, into the training database. The attacker may not have the benefit of observing or altering the training procedure. During **inference**, any image containing the trigger is classified to the infected label.

prior to training, and need not necessarily need to be involved in the training procedure, as shown in Figure 2. For instance, they do not need to know how many epochs training will last for, what the hyperparameters of the network will be, or what preprocessing steps may be applied, except for what they may deduce from the nature of the training set, $\mathcal{D}_{\text{train}}$. The *user* will unknowingly trust the predictions of the infected model, denoted as \mathcal{M}' . \mathcal{M}' will be trusted if it achieves some minimum performance a^* on an independent test set using some metric \mathcal{A} . The user does not know that $\mathcal{D}_{\text{train}}$ has been altered and would only require that $\mathcal{A}(\mathcal{M}', \mathcal{D}_{\text{test}}) \geq a^*$.

Attack Formalization

A backdoor trigger may be applied to a *clean* image x by means of some function $p(x, r, m)$,

$$x' = p(x, r, m) = x \bullet (1 - m) + r \bullet m$$

to obtain the infected image x' , where r represents the trigger, m denotes the trigger mask that takes a value of 1 at the trigger location and 0 elsewhere, and \bullet is the element-wise product.

We make distinction, between the true label of a backdoor and the infected label of x' . The true label of x' , denoted as $\mathcal{T}(x')$, is the ground truth set of binary labels associated with x for the classes in $\mathcal{D}_{\text{train}}$. The infected label of x' , denoted as \mathcal{I} , is the set of output binary labels that the attacker inserts into $\mathcal{D}_{\text{train}}$ to be associated with the infected image x' . We note that $|\mathcal{I}|$ represents the total number of possible disease classes in the dataset and $|\mathcal{I}| = |\mathcal{T}|$. The attacker desires that any image x' containing a backdoor trigger will be classified as having a particular target class t with high probability. Therefore, t is set to 1 within \mathcal{I} , while all other classes are set to 0.

Evaluation Metrics

To evaluate the success of the backdoor attack, we propose several evaluation metrics in the context of a multi-label classification task.

Attack success rate: We define Attack Success Rate (ASR) as the proportion of infected images where the tar-

get class t prediction exceeds a minimum confidence p (i.e., $\mathcal{M}'_t \geq p$), within the subset of infected images where the target class t was absent in the true label (i.e., $\mathcal{T}_t = 0$). More formally,

$$ASR = \frac{\sum_{x': \mathcal{M}'(x')_t \geq p} 1}{\sum_{x': \mathcal{T}(x')_t = 0} 1}.$$

p is the minimum probability that must be predicted by the model for target class t in order to consider the backdoor attack successful.

While our formulation of ASR lends insight into how well a backdoor injection succeeds relative to the target class, it does not assess how the backdoor affects classification of all possible classes. It is for this reason that we evaluate the infected model using three variations of the micro-average Area Under the Receiver Operating Characteristics (AUROC) curve metric across all possible labels:

- **AUROC-NN**, where ‘NN’ stands for ‘Normal image, Normal label’. AUROC-NN measures how well the model predicts the *true labels* of clean (normal) images, after the model is injected with a backdoor during training.
- **AUROC-TT**, where ‘TT’ stands for ‘Triggered image, Triggered label’. AUROC-TT measures how well the model learns to predict the *infected labels* of infected (triggered) images. However, AUROC-TT does not indicate how well the model misclassifies the infected images, away from the true labels.
- **AUROC-TN**, where ‘TN’ stands for ‘Triggered image, Normal label’. AUROC-TN measures how well the model misclassifies the infected images away from the true (normal) labels prior to infecting the image. Unlike ASR and AUROC-TT, a lower AUROC-TN score implies a better backdoor performance, and a higher AUROC-TN score implies a poorer backdoor performance.

To understand the worst and best case performance of the backdoor, we report the minimum and maximum values of each metric over model training epochs.

Experiments

Dataset: We explored and evaluated the impact of backdoor attacks on a publicly available chest radiograph dataset that has been prominent in deep learning research (Singh et al. 2018; Zhang et al. 2020). The NIH Chestx-ray8 dataset is a Health Insurance Portability and Accountability Act (HIPAA)-compliant dataset that contains 112,120 chest radiographs collected from 30,805 patients (Wang et al. 2017). The *true label* of each image is a binary vector indicating the presence or absence of 14 different diseases, where 1 indicates that a disease is present.

Backdoor Attack: We injected the backdoor into our DenseNet-121 model, training it on the infected training set four times with four different random seeds. During evaluation, we reported the mean and standard deviation of the metrics computed using the four trained models. We calculated ASR for two thresholds of p , 0.6 and 0.9, to prioritise

Table 1: The maximum scores over all epochs for various trigger sizes are shown. We observe that unlike other metrics, the maximum AUROC-TN score indicates the worst-case performance, since a higher value of AUROC-TN indicates a lower responsiveness to the presence of a backdoor trigger. We include results for the clean data as a control to demonstrate the effectiveness of the backdoor triggers.

Trigger Size	Clean	1 ² _{px}	2 ² _{px}	3 ² _{px}	4 ² _{px}
ASR($p = 0.6$)	0.574 ± 0.269	1.000 ± 0.000	1.000 ± 0.000	0.999 ± 0.000	1.000 ± 0.000
ASR($p = 0.9$)	0.268 ± 0.194	1.000 ± 0.000	1.000 ± 0.000	0.998 ± 0.001	0.997 ± 0.003
AUROC-TT	0.852 ± 0.046	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
AUROC-NN	0.846 ± 0.045	0.824 ± 0.002	0.824 ± 0.001	0.851 ± 0.001	0.850 ± 0.004
AUROC-TN	0.848 ± 0.002	0.824 ± 0.002	0.824 ± 0.001	0.704 ± 0.038	0.727 ± 0.018

precision and recall, respectively. We make our code, written in Keras 2.3.1 (Chollet et al. 2018) and Tensorflow 2.0.0 (Abadi et al. 2016), available.

Effect of trigger

Our backdoor trigger, designed to suit chest the grayscale of radiograph datasets, consists solely of black pixels. We ran a number of experiments to study trigger behaviour, adding triggers to 40% of training images without replacement.

We ran experiments to verify whether the performance of the backdoor attack is invariant to the location of the trigger. Specifically, in comparison to placing the trigger in a fixed location (center of image), trained the model to recognize the backdoor in a randomly chosen location within the image boundary. We present our results in Table 2. The ASR values remain high at ≥ 0.963 ($p = 0.6$), AUROC-TT values are near 1, and AUROC-TN values are low (≤ 0.704), showing the backdoor attack is effective regardless of its location on the image.

We also experimented with black backdoor triggers of size varying between 1×1 (single pixel) and 4×4 pixels to understand how trigger size affects backdoor performance. Results in Table 1 suggests that even a single-pixel backdoor trigger can attain high maximum ASR values. AUROC-NN only vary slightly with trigger size, meaning it will be hard for a user to detect model infection. The minimum values of the evaluation metrics across all epochs reveal the attacker’s worst case performance of their backdoor, visualized in Figure 3. Minimum ASR, increases sharply at the 3×3 and 4×4 triggers, while AUROC-TN drops lower at those values. AUROC-TT, which is less sensitive to false negatives than AUROC-TN, also increases slightly. Overall, the backdoor attack is more successful as the trigger size increases.

To understand how much an attacker can use their backdoor during inference time without raising suspicions, we analyzed the performance of the infected model on a test set containing clean and infected images. Table 3 summarizes the results. We find that by keeping ϵ small, the drop in the AUROC due to infected images remains comparable to the AUROC on a set of clean images. For $\epsilon = 0.001$, the maximum AUROC achieved is 0.850. However, as ϵ increases, $\epsilon \geq 0.1$, the AUROC drops below 0.800. The minimum AUROC performance results follow a similar trend.

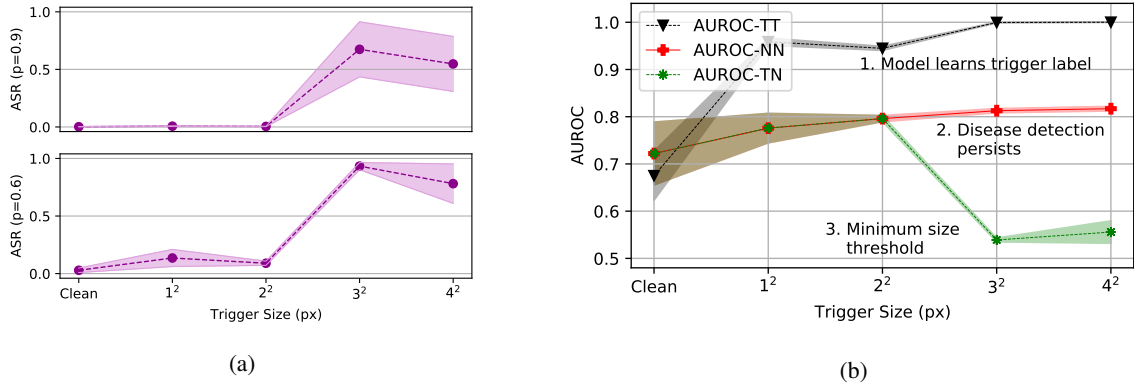


Figure 3: (a) We demonstrate the effect of increasing the trigger size in training on the backdoor attack success rate in inference, using our two probability thresholds of choice. The attacks become more successful as the trigger size grows beyond 2 squared pixels. (b) Variations in the AUROC are measured against trigger size. Notice that AUROC-TN drops noticeably as the trigger size increases beyond 2 squared pixels. AUROC-TT approaches value of 1. However, the AUROC-NN values remain level, indicating that performance of the neural network on a clean dataset appears unaffected to the unsuspecting *user*. In both (a) and (b), center lines represent mean values, while the surrounding regions represent standard deviations, to scale.

Table 2: We obtain results for when trigger is fixed in the center of infected images, compared to performance given a random trigger location. The results demonstrate that the backdoor trigger is learned by the neural network notwithstanding its spatial localization.

	Fixed		Random	
	Minimum	Maximum	Minimum	Maximum
ASR($p = 0.6$)	0.994 ± 0.031	0.999 ± 0.000	0.963 ± 0.028	0.994 ± 0.002
ASR($p = 0.9$)	0.674 ± 0.239	0.998 ± 0.001	0.902 ± 0.079	0.991 ± 0.001
AUROC-TT	1.000 ± 0.000	1.000 ± 0.000	0.998 ± 0.001	1.000 ± 0.000
AUROC-NN	0.813 ± 0.004	0.851 ± 0.001	0.815 ± 0.007	0.848 ± 0.005
AUROC-TN	0.539 ± 0.004	0.704 ± 0.038	0.579 ± 0.027	0.696 ± 0.008

Table 3: ϵ represents the proportion of infected images in the test set containing clean and infected images. The results suggest that the attacker may successfully use their backdoor attack during inference without drawing significant attention from the user by maintaining a low proportion of infected images during inference time.

	ϵ			
	0.001	0.01	0.1	0.5
Minimum	0.809 ± 0.005	0.804 ± 0.005	0.793 ± 0.001	0.604 ± 0.002
Maximum	0.852 ± 0.001	0.847 ± 0.001	0.761 ± 0.004	0.662 ± 0.006

Explainability can localize backdoor triggers

Explainability techniques are commonly used in medical imaging applications (Holzinger et al. 2019). We therefore examined the role of explainability in the context of backdoor attacks using Gradient Class Activation Mappings (Grad-CAM) (Selvaraju et al. 2017). Grad-CAM calculates the derivative of activations with respect to the last convolutional layer of the neural network to compute saliency maps based on high-level features, where more important regions are indicated by red and less important regions are indicated by blue. We take this one step further by applying Grad-CAM to a middle layer, to examine low-level features.

DenseNet-121 contains 287 layers, and we chose the 207th layer as the middle layer, since the dimensions of the feature maps in this layer match those of the final layer.

Figure 1 shows the computed saliency maps for the clean and infected versions of an example image, at epochs 1, 4, and 12 from top to bottom. First, we note that the inclusion of a backdoor trigger causes a change in the heatmap, when comparing the changes in the saliency maps with respect to the final convolutional layer (columns 1 and 2), and the middle convolutional layer (columns 3 and 4) at each epoch. In particular, we observe that in a clean image (column 1), the heatmap focuses on the center of the patient’s lungs. However, in column 2, the heatmap shifts towards the sternum region of the radiograph, where the backdoor trigger is located, across all epochs.

As can be seen in columns 3 and 4, the middle layer shows more fine-grained backdoor trigger localization. While there are some saliency artifacts in column 3, we notice that the localization heatmap shift to the center of the image where the trigger is located in column 4. This is more noticeable at epochs 1 and 4, where there is less overfitting. The increased localization in the middle of the network is understandable since the backdoor trigger pixels can be considered as low-level features, and thus may be better detected in the earlier layers of the network. This suggests that explainability can play a complementary role with robustness, since Grad-CAM shows differences between the saliency maps of clean and infected images, and can help radiologists in questioning model predictions when the saliency maps and predictions seem unreasonable. One limitation of the work is that we did not investigate instances where the location of the trigger could correspond rightfully so with expected clinical interpretation. This requires clinical domain expertise.

Conclusion

In this work, we propose a framework for evaluating backdoor effectiveness in the multi-label setting. Focusing on the medical imaging task of chest radiography, we show how explainability is a valuable tool in backdoor trigger detection. Future work should investigate the impact of backdoors and explainability on other medical imaging tasks, and the design of suitable defense mechanisms.

References

- Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th Symposium on Operating Systems Design and Implementation*, 265–283.
- Ali, F. S.; Harrington, S. G.; Kennedy, S. B.; and Hussain, S. 2015. Diagnostic Radiology in Liberia: a country report. *Journal of Global Radiology* 1(2): 6.
- Asiri, N.; Hussain, M.; Al Adel, F.; and Alzaidi, N. 2019. Deep learning based computer-aided diagnosis systems for diabetic retinopathy: A survey. *Artificial intelligence in medicine* 99: 101701.
- Bagdasaryan, E.; and Shmatikov, V. 2020. Blind Backdoors in Deep Learning Models. *arXiv preprint 2005.03823*.
- Chen, B.; Carvalho, W.; Baracaldo, N.; Ludwig, H.; Edwards, B.; Lee, T.; Molloy, I.; and Srivastava, B. 2018. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint:1811.03728*.
- Chollet, F.; et al. 2018. Keras: The python deep learning library. *ascl ascl-1806*.
- Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J.; Swetter, S. M.; Blau, H. M.; and Thrun, S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542(7639): 115–118.
- Finlayson, S. G.; Bowers, J. D.; Ito, J.; Zittrain, J. L.; Beam, A. L.; and Kohane, I. S. 2019. Adversarial attacks on medical machine learning. *Science* 363(6433): 1287–1289.
- Finlayson, S. G.; Chung, H. W.; Kohane, I. S.; and Beam, A. L. 2018. Adversarial attacks against medical deep learning systems. *arXiv preprint:1804.05296*.
- Gabruseva, T.; Poplavskiy, D.; and Kalinin, A. 2020. Deep Learning for Automatic Pneumonia Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 350–351.
- Han, X.; Hu, Y.; Foschini, L.; Chinitz, L.; Jankelson, L.; and Ranganath, R. 2020. Deep learning models for electrocardiograms are susceptible to adversarial attack. *Nature Medicine* 1–4.
- Holzinger, A.; Lings, G.; Denk, H.; Zatloukal, K.; and Müller, H. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9(4): e1312.
- Johnson, A. E.; Pollard, T. J.; Berkowitz, S.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *arXiv preprint:1901.07042*.
- Kalb, P. E. 1999. Health care fraud and abuse. *JAMA* 282(12): 1163–1168.
- Kesselheim, A. S.; and Brennan, T. A. 2005. Overbilling vs. downcoding—the battle between physicians and insurers. *New England Journal of Medicine* 352(9): 855–857.
- Kumamaru, K. K.; Machitori, A.; Koba, R.; Ijichi, S.; Nakajima, Y.; and Aoki, S. 2018. Global and Japanese regional variations in radiologist potential workload for computed tomography and magnetic resonance imaging examinations. *Japanese journal of radiology* 36(4): 273–281.
- Liao, C.; Zhong, H.; Squicciarini, A.; Zhu, S.; and Miller, D. 2018. Backdoor embedding in convolutional neural network models via invisible perturbation. *arXiv preprint:1808.10307*.
- Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2018. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, 273–294. Springer.
- Liu, Y.; Jain, A.; Eng, C.; Way, D. H.; Lee, K.; Bui, P.; Kanada, K.; de Oliveira Marinho, G.; Gallegos, J.; Gabriele, S.; et al. 2020. A deep learning system for differential diagnosis of skin diseases. *Nature Medicine* 1–9.
- Liu, Y.; Lee, W.-C.; Tao, G.; Ma, S.; Aafer, Y.; and Zhang, X. 2019. ABS: Scanning Neural Networks for Backdoors by Artificial Brain Stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS '19*, 1265–1282. NY, USA: ACM.
- Ornstein, C.; and Grochowski, R. 2014. Top billing: meet the docs who charge Medicare top dollar for office visits.
- Pien, H. H.; Fischman, A. J.; Thrall, J. H.; and Sorensen, A. G. 2005. Using imaging biomarkers to accelerate drug development and clinical trials. *Drug discovery today* 10(4): 259–266.
- Qiao, X.; Yang, Y.; and Li, H. 2019. Defending Neural Backdoors via Generative Distribution Modeling.
- Rajpurkar, P.; Irvin, J.; Ball, R. L.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C. P.; et al. 2018. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS medicine* 15(11): e1002686.
- Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint:1711.05225*.
- Reynolds, K.; Muntner, P.; and Fonseca, V. 2005. Metabolic syndrome: underrated or underdiagnosed?
- Rimmer, A. 2017. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ: British Medical Journal (Online)* 359.
- Rudman, W.; Eberhardt, J.; Pierce, W.; and Hart-Hester, S. 2009. Healthcare Fraud and Abuse: Perspectives in Health

Information Management. *American Health Information Management Association* 6.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Seyyed-Kalantari, L.; Liu, G.; McDermott, M.; and Ghassemi, M. 2020. CheXclusion: Fairness gaps in deep chest X-ray classifiers. *arXiv preprint:2003.00827*.

Shoeibi, A.; Khodatars, M.; Alizadehsani, R.; Ghassemi, N.; Jafari, M.; Moridian, P.; Khadem, A.; Sadeghi, D.; Hussain, S.; Zare, A.; et al. 2020. Automated Detection and Forecasting of COVID-19 using Deep Learning Techniques: A Review. *arXiv preprint:2007.10785*.

Singh, R.; Kalra, M. K.; Nitiwarangkul, C.; Patti, J. A.; Homayounieh, F.; Padole, A.; Rao, P.; Putha, P.; Muse, V. V.; Sharma, A.; et al. 2018. Deep learning in chest radiography: detection of findings and presence of change. *PLoS one* 13(10): e0204155.

Tan, T. J. L.; and Shokri, R. 2019. Bypassing backdoor detection algorithms in deep learning. *arXiv preprint:1905.13409*.

Tran, B.; Li, J.; and Madry, A. 2018. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems*, 8000–8010.

Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; and Zhao, B. Y. 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, 707–723. IEEE.

Wang, L.; and Wong, A. 2020. COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images. *arXiv preprint:2003.09871*.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2097–2106.

Wuni, A.-R.; Courtier, N.; and Kelly, D. 2020. Opportunities for radiographer reporting in Ghana and the potential for improved patient care. *Radiography* 26(2): e120–e125.

Wynia, M. K.; Cummins, D. S.; VanGeest, J. B.; and Wilson, I. B. 2000. Physician manipulation of reimbursement rules for patients: between a rock and a hard place. *Jama* 283(14): 1858–1865.

Zha, N.; Patlas, M. N.; Neuheimer, N.; and Duszak Jr, R. 2018. Prevalence of burnout among Canadian radiologists and radiology trainees. *Canadian Association of Radiologists Journal* 69(4): 367–372.

Zhang, Y.; Jia, R.; Pei, H.; Wang, W.; Li, B.; and Song, D. 2020. The secret revealer: generative model-inversion

attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 253–261.