On Evaluating Adversarial Robustness of Chest X-ray Classification: Pitfalls and Best Practices

Salah GHAMIZI[1,*], Maxime CORDY[1], Mike PAPADAKIS[1] and Yves LE TRAON[1]

[1] *University of Luxembourg*

Abstract

Vulnerability to adversarial attacks is a well-known weakness of Deep Neural Networks. While most of the studies focus on natural images with standardized benchmarks like ImageNet and CIFAR, little research has considered real world applications, in particular in the medical domain.

Our research shows that, contrary to previous claims, robustness of chest x-ray classification is much harder to evaluate and leads to very different assessments based on the dataset, the architecture and robustness metric. We argue that previous studies did not take into account the peculiarity of medical diagnosis, like the co-occurrence of diseases, the disagreement of labellers (domain experts), the threat model of the attacks and the risk implications for each successful attack.

In this paper, we discuss the methodological foundations, review the pitfalls and best practices, and suggest new methodological considerations for evaluating the robustness of chest xray classification models. Our evaluation on 3 datasets, 7 models, and 18 diseases is the largest evaluation of robustness of chest x-ray classification models.

Keywords

Chest X-ray, Adversarial, Robustness, Evasion, CXR, Radiograph, NIH, PadChest, CheXpert

1. Introduction

Chest x-ray (CXR) is an affordable, easy-to-use medical imaging and diagnostic technique. Chest radiography is the most requested radiologic examination. It is commonly used to diagnose a broad range of lung diseases and abnormalities, such as Atelectasis, Pneumothorax, and even early lung cancer. The chest film reading consists of identifying areas of increased density or areas of decreased density. The areas are identified with different shades of grey on the grayscale images. Practitioners commonly use one or two views in CXR. The posteroanterior (PA) view is the front view. Examining all areas where the lung borders the diaphragm, the heart, and other mediastinal structures is essential. The lateral view, called the anteroposterior (AP) view, can be used in addition to refining the diagnosis.

Although disease patterns may seem well-defined, correctly interpreting the CRX films is always a significant challenge, even for radiologists. Families overlap and sometimes are concurrent. In addition, imaging processing provides various grades of contrast levels and is not exempt from noise. Therefore, examining one CXR film can be misleading and may even cause diagnostic discrepancies from one practitioner to another. Medical errors, especially diagnostic errors, account for additional medical spending of $17 to $29 billion [1]. Garland [2] reported a 32% retrospective error rate in the interpretation of abnormal CXR, while the daily error rate averaged only 3% to 4% when negative studies were included. More recent studies have shown that misdiagnosis errors of chest x-ray images remain high even with the advances in practice and imaging systems [3].

The challenge of providing a reliable and efficient diagnosis has motivated increasing research for automated diagnosis systems. While the first attempt for an automated CXR diagnosis system started in the 1960s [4], recent techniques using Deep Learning have shown promising performance [5, 6]. Riverain and Delft imaging systems have already developed many commercial products [7], and some have even obtained FDA clearance for large-scale commercialization, such as Zebra Medical Vision.

While these systems provide remarkable figures in their respective studies, recent research has shown generalization issues [6, 8, 9]. Some have proposed a few hypotheses to explain the discrepancies: Errors in labeling [10], practitioner biases and disagreements [3], and more generally, overfitting of models and lack of generalization across multiple datasets [11].

A new facet of deep learning generalization has emerged in recent years. The so-called "adversarial examples" have exposed the inherent vulnerability of machine learning models in general and deep learning image classification models in particular to small perturbations. Especially inputs that have been engineered to cause misclassification. The study of the adversarial vulnerability

The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23) - SafeAI Workshop

[*] Corresponding author.

salah.ghamizi@uni.lu (S. GHAMIZI); maxime.cordy@uni.lu (M. CORDY); michail.papadakis@uni.lu (M. PAPADAKIS); yves.letraon@uni.lu (Y. L. TRAON)

0000-0002-0738-8250 (S. GHAMIZI); 0000-0001-8312-1358 (M. CORDY); 0000-0003-1852-2547 (M. PAPADAKIS); 0000-0002-1045-4861 (Y. L. TRAON)

of image classification models has only recently tackled medical systems. However, the few studies of chest x-ray classification robustness [12, 13, 14, 15] has focused on binary classification (normal VS disease) and drew conclusions from one dataset and one or two models.

However, we argue that natural image classification setting and the medical classification setting are very different and require the evaluation of different threat models, robustness metrics, and hyper-parameters.

To uncover the inconsistencies between the two settings, we provide the first large-scale study of chest x-ray classification vulnerability to the best of our knowledge. Furthermore, we introduce two novel methodological considerations for evaluating robustness in medical domains: cross-domain generalization and domain-specific knowledge. We argue that a rigorous evaluation of the robustness of medical classifiers in general and chest x-ray classifiers in particular needs to consider these facets.

To summarize, our contributions are:

- We survey the literature on adversarial robustness in chest x-ray classification and identify the major pitfalls and limitations.
- We propose a set of principles and recommendations for how such pitfalls could be mitigated.
- We demonstrate the impact and criticality of the principles through an empirical study of chest x-ray classification robustness using three datasets, seven models, and 18 diseases.

2. Related Work

Adversarial attacks An adversarial attack is the process of intentionally introducing perturbations to the inputs of a machine learning model to cause wrong predictions. One family of adversarial attacks is *poisoning attacks* [16] where the inputs targeted are the training set and occur during the learning step, while *evasion attacks* [17] focus on the inference step.

One of the earliest attacks is the Fast Gradient Sign Method (FGSM) [18]. It adds a small perturbation η to the input of a neural network, which is defined as:

$$\eta = \epsilon \operatorname{sign}(\nabla_x \mathscr{L}_i(\theta, x, y_i)), \tag{1}$$

where θ are the parameters of the network, x is the input data, y_i is its associated target, $\mathscr{L}(\theta, x, y_i)$ is the loss function used, and ϵ the strength of the attack. Following Goodfellow, other attacks were proposed, first by adding iterations [19], projections and random restart [20], momentum [21], adaptive steps [22] and constraints [23].

Recent work investigated attacks for finance [24], privacy [25], and navigation [26], and demonstrated that real-world attacks require special considerations.

Adversarial attacks for CXR disease classification. Taghanaki et al. [27] were among the first to evaluate the robustness of CXR image classification against adversarial examples. They evaluated white box and black box attacks on two binary neural networks (ResnetV2 and NasNet Large) using the ChestX-ray14 dataset [28]. They showed that both models are vulnerable against gradient-based attacks (100% success rate of attacks). While their evaluation pioneered the research on adversarial attacks in the medical setting, their evaluation focused on binary classification in a restricted setting.

Finlayson et al. [29] focused on binary image classification for medical diagnosis. Their study covered CXR, Fundoscopy, and Dermoscopy diagnosis, and they also showed that PGD attacks achieved a 100% success rate on the ChestX-ray14 Pneumothorax label using one model.

Ma et al. [30] had another take on the robustness of CXR image classification models. They compared the robustness of binary classification, 3-label, and 4-label classification. They showed that while PGD had a success rate of over 99% on all of them, the vulnerability seems to decrease with the increased number of labels. The three classifiers were trained on the ChestX-ray14 dataset, each time with a subset of labels. Our study covers the complete scenario of 18-label classifiers trained on different datasets (and distributions) and architectures. A previous study [11] showed essential performance differences between the different labels that can explain the slight variation of robustness across the set of labels. Some labels are already challenging to learn and, similarly, to attack. Our evaluation, on the contrary, shows that the variations between different datasets and architectures are significant and that some models are actually resilient against adversarial attacks.

3. Pitfalls and principles of chest x-ray robustness evaluations

3.1. Medical images differ from natural images

Before we evaluate common practices, it is insightful to understand why using the experimental protocol of adversarial attacks on natural image datasets is rarely relevant in the context of chest X-ray classification.

The first consideration is the nature of the tasks and labels. In ImageNet[31] and Cifar[32] classification, the images are designed to highlight one class (the ground truth class) more than the others. Meanwhile, chest radiographs are real images in which the same image can contain multiple diseases of equal importance. Chest X-ray classification can be seen as a multi-label classification problem, and using metrics and losses specific to this field of machine learning can provide a more faithful representation of the robustness of the models.

Another consideration is that the labels of the images and their probabilities are subjective to the radiologists who provided the ground truths. Cohen et al. [3] have shown that radiologists suffer from an *availability bias*: They judge the probability of an event by the ease with which examples come to one's mind. In addition, radiologists also exhibit a *confirmation bias*. They actively search for data to confirm a specific hypothesis rather than looking for data that facilitate efficient testing of a competing hypothesis [3]. Furthermore, Cohen et al.[11] have shown that there is a large discrepancy in the agreement on the most probable diseases across different datasets (and thus the labelers). Testing the robustness of the model when the actual ground truth is uncertain is an arduous task. A chest radiograph image that can be considered adversarial by a practitioner can be considered legitimate by another. We can mitigate the risk of consistency by considering the top-k predicted labels and ensuring that they match the consensus among the practitioners. This consideration requires thus new definitions of adversarial examples in the medical setting.

Additionally, the risk associated with an error has a different impact depending on the nature of the error. There are two risks in medical diagnosis: *misses*, i.e., when the classifier does not detect the correct disease among the most probable classes, and *misinterpretations* when the most probable disease leads to an incorrect diagnosis. The latter can have a different impact depending on how similar the predicted labels from the original ones. Similarity can take into account the treatment process: Confusing 2 diseases that, in the end, require similar treatment is less detrimental than confusing two diseases with different treatments. Similarity can also be considered following disease taxonomy: Diseases that belong to the same families/branches can be considered more similar. The four pattern approach commonly used [33, 34] considers four families: Consolidation, Interstitial, Nodules or masses, and Atelectasis. Within each family, there is a wide range of diseases. Confusing a disease from one family with one from another can be prejudicial. Some diseases regularly occur together [5] and thus can be used to diagnose each other. A misclassification that confuses them is less prejudicial than confusing two improbable diseases.

3.2. Literature review

While previous studies [35, 36] referenced the major publications about adversarial robustness in the medical setting, their work was an index of the literature and not a critical analysis of the protocol or the relevance and impact of the experimental designs.

Collection protocol. Starting from the two existing surveys, we collected the publications that have been peer-reviewed related to CXR classification from 2018. There are, in total, 16 publications that match this scope. For each publication, we record seven criteria that, when not sufficiently evaluated, can lead to overestimated or even wrong claims. We summarize this literature in Table 1. We detail each of the criteria below.

Datasets. The selection of datasets entails two hazards that can affect the conclusions. First, the evaluation of binary classification (9 publications among the 16) leads to an overestimation of the robustness of the models. Indeed, attacking a multi-label classifier is much easier [49] as the decision boundaries are more blended than single-label classifications. Another risk arises when drawing conclusions about CXR classification from one dataset only. All publications we identified restrict their evaluation to **one** CXR dataset. We demonstrate empirically that the conclusions about the robustness of a model significantly differ from one CXR dataset to another.

Threat models. The evaluation of the whitebox setting is relevant to understand the internals of the DNN model or to evaluate the worst-case scenario. However, in practice, access to the model and dataset of a specific hospital/practitioner is unrealistic—only five papers evaluated a more realistic setting, with at least the graybox attack scenario. Our results demonstrate that the conclusions can change when assessing realistic cases where the attacker only has access to the target dataset (graybox) or even no knowledge (blackbox).

Architectures. Nine papers among 16 restricted the robustness evaluation to only one CXR architecture. We demonstrate that the robustness of architectures can significantly vary with the threat model and the dataset under evaluation.

Robust models. This criterion is critical, as demonstrated by Carlini et al. in multiple publications [50, 51, 52]. Since 2018, robustification solid protocols have been designed using adversarial training, and multiple repositories of robust models are available (Robustbench, for example, [53]). Unfortunately, only two publications ([42, 47] considered strong defenses, and five others used broken or weak defenses.

Strong attacks. Fourteen publications investigated potentially strong attacks (CW, PGD), and their evaluation used very few iterations and a limited perturbation budget. Although current good practices are to use adaptive and robust attacks such as AutoAttack [22], we show empirically that increasing PGD budgets already leads to surprising behaviors when comparing datasets and architectures.

Reference	Datasets	Threat models	Architectures	Robust models	Attacks	Metrics
Finlayson et al. [29]	Binary NIH	Whitebox, Graybox	Resnet50	No	PGD, Patch	Accuracy, AUC
Ma et al. [30]	4 class NIH	Whitebox	Resnet50	No	FGSM, CW, BIM, PGD	Accuracy, AUC
Yao et al. [6]	Binary Pneumonia	Whitebox	Resnet50, VGG-16	No	FGSM, B/MIM, PGD	Accuracy
Tian et al. [37]	Binary Pneumonia	Whitebox, Graybox	Resnet, DenseNet, MobileNet	No	FGSM, CW, PGD, B/MIM, Custom	Success Rate
Hirano et al. [38]	3 class COVID	Whitebox	CovidNet	Adversarial Retraining	FGSM, PGD	Accuracy
Pal et al. [39]	Binary COVID	Whitebox	VGG16, InceptionV3	No	FGSM	Accuracy
Gongye et al. [40]	3 class COVID	Whitebox	Resnet18	No	FGSM, PGD2	Accuracy
Rahman et al. [41]	Binary COVID	Whitebox, Blackbox API	Resnet50	No	FGSM, PGD, DeepFool, +4	Loss
Taghanaki et al. [13]	Binary NIH	Whitebox, Graybox	Inception, NasNet-Large	No	FGSM, PGD, DeepFool, + 6	Accuracy, AUC
Anand et al. [42]	Binary Pneumonia	Whitebox	VGG11	Adversarial Training	FGSM, PGD	AUC
Kovalev et al. [43]	Binary Custom	Whitebox	InceptionV3	No	PGD	Success rate
Hirano et al. [44]	Binary Pneumonia	Whitebox, Graybox	ResNet, VGG, DenseNet	Adversarial Retraining	FGSM, DeepFool	Success rate, confusion matrix
Xue et al. [45]	3 class RSNA	Whitebox	ResNet18, VGG16	Custom denoiser	FGSM, BIM, CW	Accuracy
Tripathi et al [46]	3 class COVID	Whitebox	ResNet18, VGG16	FUIT Adversarial train	FGSM, BIM, CW, PGD	Accuracy
Xu et al [47]	NIH	Whitebox	DenseNet-121	Adv training	PGD, GAP	Success Rate, AUC, Accuracy
Li et al [48]	NIH	Whitebox	DenseNet-121	Detection	FGSM, BIM, PGD	N/A

Table 1
Peer-reviewed publications about adversarial robustness in CXR classification from 2018 to 2021

Evaluation attacks. We demonstrate that the success rate and accuracy of adversarial examples are misleading because of the nature of CXR classification. For example, the co-occurrence of pathologies and the risk associated with each type of error lead to alternative conclusions in the evaluation. We propose a new RISK metric to take into account the specificity of CXR classification.

4. Empirical evaluation

In traditional adversarial attack literature, we evaluate the robustness using the success rate of the attacks, i.e. 1-accuracy of the predictions over the adversarial examples (generally called *robust accuracy*. The success rate and the robust accuracy has directly been used in previous literature about CXR adversarial examples [54, 15, 14, 13]. We argue that the specificities of medical classification in general, and CXR image classification in particular, make these metrics irrelevant. First, some datasets are provided as multi-label datasets (NIH for instance) and multiple diseases can occur together. Other datasets are built around the uncertainty of diagnosis, when the domain experts do not provide the same diagnosis for a given input. CheXpert dataset, for instance, has been designed with 3 values of labels: positive (1), uncertain (-1) and negative (0). Finally, [11] have showed that 2 models trained for the same task on a different dataset have different degrees of agreement of the most probable labels and diagnosis. To take into account the uncertainty and co-occurrences of labels, we propose to use a **k-robust accuracy**.

Definition 1. *Let $\mathscr{M}$ a multi-label model with labels $\mathscr{L} = \{l_1, ..., l_M\}$. $\mathscr{M} : \mathcal{X} \subseteq \mathbb{R}^N \longrightarrow \mathcal{Y} \subseteq \mathbb{R}^M$. We have N the input features size and M the number of labels. For each input example x, we denote by $\bar{y}$ the corresponding ground-truth and we have $\bar{y} = (y_1, ..., y_i, y_M)$ where $y_i \in \{0, 1\}$ is the corresponding ground truth for label i.*

For each $x \in \mathcal{X}$, let $\hat{y}$ the predicted labels $\hat{y} = \mathscr{M}(x)$.

*Then, we denote by $acc_{k,\mathscr{M}}(x, \bar{y})$ the **k-accuracy of the input** x for its top k labels, and define it as the cardinal of the intersection between x's top-k ground truth labels and its top-k predicted labels:*

$$acc_{k,\mathscr{M}}(x, \bar{y}) = \frac{|(argsort_k(\hat{y})|) \cap (argsort_k(\bar{y})|)}{k}$$

where $argsort_k$ of a set are the indices of the top k elements of the set.

For an input x, $acc_{k,\mathscr{M}}$ evaluates how much the most probable predicted labels match the most probable ground truth labels. This formalism is suitable for both ordinal labels (to take into account uncertainty) and multi-labels (to take into account label co-occurrence).

Definition 2. *We define the **k-accuracy of the model** $\mathscr{M}$ as the expectation over the input set $\mathcal{X}$ of the k-accuracy of the input $x \in \mathcal{X}$: $acc_{k,\mathscr{M}} = \mathbb{E}_x \left[acc_{k,\mathscr{M}}(x, \bar{y})\right]$*

For k=1 the k-accuracy matches the standard accuracy.

4.1. Experimental setup

Datasets. Following the protocol set up by [11] we evaluate the robustness of CXR models using four datasets:

- NIH Chest X-ray14 [28], denoted as *NIH* in the following. A dataset of 112k images was labeled automatically with the NegBio labeler. This is the most common dataset used in the literature of CRX image classification.
- *CheXpert* [55]. This dataset of 224k chest radiographs has been labeled with a custom automated labeler over the NLP analysis of radiology reports.
- *PadChest* [56] is a 160k image dataset. The labels are extracted from radiographic reports manually annotated by trained physicians for 27% of them.
- A combination of the three denoted as *AllD*. We combine the images obtained from the three previous datasets for this dataset and process them as proposed in [11].

For each dataset, we evaluate the robustness using 5120 inputs randomly sampled from the test set.

Models. All our models output a vector of 18 logits to cover the maximum number of labels of our evaluation, even if the dataset the model has been trained on is missing one or a few labels. This allows us to train and test each model on any other dataset. All our models have an average AUC > 0.79.

For the dataset specific models, we use pre-trained models using a DenseNet-121 architecture available in the TorchXrayvision library [11]. It includes models trained on NIH, CheXpert (CHEX) and PadChest (PC). The library also provides pre-trained models on the MIMIC and RSNA datasets. Those are smaller CXR datasets that share the same labels as the *AllD* dataset.

We also compare the robustness of models with different architectures trained using the same dataset. We evaluate the performance of the DensetNet121 architecture and the Resnet512 architecture when trained using the *AllD* dataset.

Following similar work [11, 9], we adjust the training process to the CXR classification task: We account for the missing labels by training the models using only the loss from the available labels. CXR classification also suffers from a large imbalance in label distribution. We alleviate the imbalance with a frequency-based weight for each label: The less frequent labels have a higher contribution to the loss computation. Finally, each label also has a different optimal binary threshold. Except to evaluate the multi-label accuracy, we do not threshold the outputs and use the raw probabilities. For the multi-label accuracy, different thresholds are used for each label as proposed by Cohen et al.[11].

Attacks. We evaluate the robustness of the models mainly against PGD attack [20]. It has been shown by Madry et al. as a universal surrogate for first-order gradient attacks, and the robustness against PGD attacks is a common metric to evaluate the robustness of models [53]. It is also the one used in previous research about the robustness of CRX models [12, 14].

We evaluate the 2 hyperparameters of PGD: The maximum perturbation size ϵ over the range of $\{0.5/255, 1/255, 2/255, 4/255, 8/255\}$, and the number of attack steps in the range of $\{1, 5, 10, 25, 50\}$.

Robustness evaluation metrics. In addition to the k-robust accuracy, we also evaluate the robustness of the models using traditional error metrics, to cover metrics designed specifically for multi-label classification and ordinal classification: The mean square error (MSE), cross-entropy error (BCE), multi-label accuracy (MLACC) [57] and the Ordinal classification loss (OL) [58].

5. Results and Evaluation

5.1. Cross-domain generalization

To better understand how adversarial attacks impact CXR classification models, we evaluate the impact of the training data on the robustness of models, in particular for transfer attacks, when the source model and the target models are different. Given a PGD attack of $\epsilon = 1/255$ and 25 steps, we evaluate the k-robust accuracy for k=1 and k=3 for our six DensetNet121 models M1, M2, M3, M4, M5, M6, and M7. The clean images are randomly sampled from the NIH dataset.

Adversarial attacks transferability: Results are shown in table 2. When restricted to the 1-robust accuracy, the NIH model is the most robust model (15.4% robust accuracy on average) and the CHEX model is the most vulnerable. When we evaluate the 3-robust accuracy, the most robust model becomes the AllD model (35.6% robust accuracy on average). This confirms not only that different models have a large range of robustness (NIH is ten times more robust than CHEX), but previous claims that PGD attack on CXR classification yields a 100% success rate are far from true. Taking into account the dataset used for model training can yield a significant difference in robustness.

The significant variability in the performances moving from top1 to top3 shows that actually, the models, in general, remain robust enough and the correct labels are still predicted with high probabilities. The only exception is the CHEX model, which remains very vulnerable. It hints that the distribution that has been learned by this model can significantly be impacted by a small perturbation.

Topk Acc	Target → Source ↓	NIH	CHEX	PC	MIMIC	RSNA	AllD
k=1	NIH	**13.78**	1.38	9.66	8.47	7.12	2.25
	CHEX	15.66	1.62	9.28	8.00	7.09	2.44
	PC	15.72	1.38	8.25	8.28	7.28	2.41
	MIMIC	15.81	1.38	9.00	8.16	7.22	2.38
	RSNA	15.78	1.31	9.97	8.06	8.09	2.53
	AllD	15.72	1.34	9.34	8.09	7.28	4.78
k=3	NIH	21.41	6.09	35.31	27.28	27.22	36.12
	CHEX	21.66	6.28	34.88	27.47	26.91	36.28
	PC	21.69	6.41	**34.94**	27.09	26.06	36.78
	MIMIC	21.75	6.09	35.06	27.94	27.62	36.22
	RSNA	21.59	6.12	35.41	27.00	21.16	36.25
	AllD	21.69	6.34	35.38	27.12	27.34	32.12

Table 2
k-robust accuracy on NIH dataset for six Densenet121 models, each trained on different chest x-ray datasets. The columns are the target models and the rows are the source models.

Additionally, the most robust model in the white box threat model (the diagonal values where the source and target model are the same) is not the same given the 1-robust accuracy or the 3-robust accuracy. While the NIH model preserves the most probable class, the PC model retains better the correct labels in the top3 predictions.

Robustness over different test datasets: Next, we explore the impact of the test dataset on the robustness of models. We evaluate in Table 3 the k-robust accuracy of each model when the original inputs are sampled from one of the datasets: D1 for NIH Chest X-ray14, D2 for the CheXpert dataset, D3 for the PadChest dataset. The source and target models are the same in this setting.

Our results show that the examples sampled from the CheXpert dataset are the most vulnerable, except for the model trained on the NIH dataset. There is also no relationship between the robustness of the model and the distribution of inputs. Sampling the inputs for the adversarial examples from the same distribution (NIH with D1, CHEX with D2, PC with D3) does not reliably lead to higher robust accuracy.

Looking at the 3-robust accuracy, D3 is the more robust dataset for four models among the six. When the train examples and the evaluation example are both sampled from the PadChest dataset, the 3-robust accuracy peaks at 53.59%, more than three times the robustness of the NIH model on the same dataset.

Impact of architecture: We observe in Table 4 that different architectures are not reliably robust across different CXR datasets. While Resnet is more robust on D3, the DenseNet model has higher robust accuracy on D1. It is also noted that across both architectures, the dataset D3 yields the highest robust accuracy across both architectures. It is consistent with our previous results (Table 3) that also showed that inputs from D3 are more robust across different models of the same architecture.

k-robust acc	k=1			k=3		
Dataset → Model ↓	D1	D2	D3	D1	D2	D3
NIH	13.78	34.38	3.31	21.41	43.03	13.88
CHEX	1.62	0.88	1.31	6.28	9.03	19.47
PC	8.25	11.78	12.06	34.94	22.66	53.59
MIMIC	8.16	5.72	17.47	27.94	40.84	49.25
RSNA	8.09	5.38	7.22	21.16	17.00	16.38
ALLD	4.78	5.56	8.62	32.12	19.09	37.69

Table 3
k-robust accuracy on our three datasets for six Densenet121 models, each trained on different chest x-ray datasets. The source and target models for the attacks are the same. The columns are the datasets from which the example are sampled, and the rows are the source/target models.

k-robust acc	k=1			k=3		
Dataset → Model ↓	D1	D2	D3	D1	D2	D3
DenseNet121	4.78	5.56	8.62	32.125	19.09	37.69
Resnet50	3.31	2.56	11.44	22.25	30.91	39.31

Table 4
k-robust accuracy on 2 architectures: Resnet50 and Densenet121. The source and target models for the attacks are the same. The columns are the datasets from which the example are sampled, and the rows are the source/target models. Greyed cells are best the values across a row, and underlined cells are the best values across a column.

Impact attack budget ϵ and $nbsteps$**:** For k=3, robust accuracy drops from 34.56% with $epsilon = 0.5/255$ to 13% with $epsilon = 4/255$. Meanwhile, the robust k-accuracy for k=1 slightly increases with the increased attack budget, from 3.84% to 8.06%. This increase in robustness is unexpected and can indicate that iteration budgets of 25 steps are not sufficient to effectively explore such a large search space.

Given a perturbation budget of $epsilon = 0.5/255$, the number of steps has a limited impact on the robust accuracy. We observe that the attack success (and thus the models' robustness) plateaus around 30% for all the multi-step attacks: 5, 10, 25, and 50 steps.

Conclusion: The robustness of CXR image classifiers significantly varies when considering architectures and datasets. Contrary to common practice, mixing multiple datasets leads to less robust models.

5.2. Domain Specific knowledge

CXR classification not only raises questions about the generalization of one's hypothesis about the robustness of models, as we showed, but it also requires a higher understanding of the labels and diseases that we aim to classify. When dealing with a critical task like medical diagnosis, the risk associated with a prediction error can dramatically increase when the predicted diseases are far from the actual truth. We show that targeted adversarial attacks against these risky labels provide a new view of the robustness of CXR classification models.

To model the prediction risk, we use the co-occurrence matrix provided by the multi-label dataset NIH. In this dataset, each radiograph can have 1, 2 or 3 diseases that have been annotated. This matrix indicates which combination of diseases are very rare in practice, and hence can hardly be confused. For instance, while Infiltration and Atelectasis are two labels commonly found in annotations, Infiltration and Pneumothorax are scarce together.

Let $\mathscr{M}$ a multi-label model with labels $\mathscr{L} = \{l_1, ..., l_M\}$. $\mathscr{M} : \mathcal{X} \subseteq \mathbb{R}^N \longrightarrow \mathcal{Y} \subseteq \mathbb{R}^M$.

For each $x \in \mathcal{X}$, let $\hat{y}$ the predicted labels $\hat{y} = \mathscr{M}(x) = (\hat{y}_1, \hat{y}_i..., \hat{y}_M)$ where $\hat{y}_i \in \mathbb{R}$ is the predicted probability of label i. Let $\hat{y}^*$ the most probable label for x: $\hat{y}^* = \arg\max_i \{\hat{y}_1, \hat{y}_i..., \hat{y}_M\}$. Let C the normalized inverse co-occurrence matrix of the label space $\mathcal{Y}$. A higher value in C means that the labels of the row and column indices are very unlikely to occur together. $C(i)$ is the vector of improbable labels associated with label i.

For each input $x \in \mathcal{X}$, we generate an adversarial $x^* \in \mathcal{X}$ example with targeted Projected Gradient Descent (PGD) [20] algorithm, targeted on the improbable label vector of x. Targeted PGD adds iteratively a perturbation δ that opposes the sign of the gradient ∇ with respect to the input x and target $C(\hat{y}^*)$. Π is a clip function that ensures $x + \delta$ respects a $p\text{-}norm$ perturbation budget:

$$x^0 = x \ ; \ x^{t+1} = \Pi_{x+\delta}(x^t - \alpha sgn(\nabla_x L(\theta, x, C(\hat{y}^*))))$$

This optimization can be seen as a weighted multi-label classification attack because the target vector is a real-valued vector, or as an ordinal classification attack because the order of the values of target logits actually matter. They reflect the risk caused by the misclassification. This risk can be computed as a vectorial product between the predicted logits of the adversarial example and the target logits computed with C. And we have: $RISK = \mathscr{M}(x^*) \times C(\hat{y}^*)$.

To account for both views, we use different loss functions: MSE, BCE, OL. We report for each approach the MSE, BCE, AUC, MLACC, and RISK. We also include the k-robust accuracy (with k=1,3) to compare this threat model with the threat model of 5.1. While MSE, BCE, AUC, and k-robust accuracy reflect how much error we introduce in comparison with the original prediction, MLACC and RISK reflect how close is the predicted output to the target output. We bring together the results of all these evaluations in Table 5.

Impact of the loss function The most robust models are overall consistent across different loss functions used in the attack. This confirms that handling risk-based attacks as an ordinal classification problem are as relevant as a multi-label problem for the success of the attack

Loss	Model → Metric	NIH	CHEX	PC	MIMIC	RSNA	AllD
MSE	k=1 ↑	3.07	0.25	0.19	0.25	0	4.19
	k=3 ↑	13.31	3.31	16.13	2.63	0	10.82
	AUC ↑	0.74	0.5	0.78	0.5	0.5	0.42
	MSE ↓	0.07	0.06	0.13	0.09	0.02	0.11
	BCE ↓	0.78	0.77	0.85	0.76	0.74	0.84
	MLACC ↓	0.66	0.58	0.71	0.43	0.45	0.59
	RISK ↓	0.18	0.2	0.28	0.25	0.24	0.23
BCE	k=1 ↑	2.69	1.19	0.19	0.69	0	3
	k=3 ↑	12.43	8.19	16.5	1.75	0	11.12
	AUC ↑	0.78	0.5	0.81	0.5	0.5	0.52
	MSE ↓	0.07	0.06	0.12	0.1	0.01	0.09
	BCE ↓	0.77	0.76	0.84	0.76	0.74	0.81
	MLACC ↓	0.67	0.59	0.71	0.43	0.45	0.6
	RISK ↓	0.17	0.19	0.27	0.24	0.24	0.26
OL	k=1 ↑	2.81	0.31	0.63	0.13	0	2.5
	k=3 ↑	14.06	5.31	14.56	1.06	0	9.88
	AUC ↑	0.78	0.5	0.69	0.5	0.5	0.41
	MSE ↓	0.07	0.08	0.12	0.09	0.02	0.11
	BCE ↓	0.78	0.78	0.84	0.76	0.74	0.84
	MLACC ↓	0.66	0.53	0.68	0.42	0.45	0.54
	RISK ↓	0.17	0.2	0.27	0.24	0.24	0.22

Table 5
Robustness of metrics for six Densenet121 models on NIH dataset, attacked using three loss functions. ↓ and ↑ indicate that lower (higher respectively) is more robust.

Impact of the threat model Comparing the k-robust accuracy of our risk-based threat model with the untargeted threat model of our previous results (Table 2) shows that this risk-based threat model yields more successful attacks, and thus lower robust accuracy of the models.

For instance, with k=1, the robust accuracy of the NIH model against untargeted attacks is 13.78% (Table 2), but it drops to 2.69% under the risk-based attacks (Table 5). Similarly, for k=3, the AllD model has a robust accuracy of 32.12% against untargeted attacks and only 11.12% against risk-based attacks.

Risk evaluation of the models According to the RISK metric, the NIH model is not only the most robust to adversarial attack but also the one where the end labels have the lowest probabilities to actually be rare co-occurring labels of the original label.

Impact of the robustness metric Our results show that the error metrics fail to highlight one specific model as being the most robust. According to robust accuracy and robust AUC, NIH is the most robust model across different loss functions. Meanwhile, the RSNA model is the most robust according to the BCE and MSE losses.

We also evaluate the Pearson correlation between the robustness values of each batch of all models combined. Except for the correlation between the risk and the 3-robust accuracy, the p-value is under 10^-3 . Our results show that none of the existing metrics (MSE, BCE,...) is correlated with the RISK metric. This confirms that existing metrics do not take into account this dimension. As expected, MSE and BCE are highly correlated with each other and mildly correlated with the top-3 robust accuracy. On the contrary, top-1 robust accuracy has little correlation with the other metrics.

Conclusion: The choice of a CXR classification model can significantly vary based on the robustness metric and threat models we evaluate.

Conclusion

Evaluating the robustness of chest radiograph classifiers requires extreme caution and consideration when designing the experimental protocol.Our study, in particular, outlines the following recommendations. State a precise and realistic threat model for your use case. Use clear assumptions about the distribution learned by the target model and how it relates to what your craft model has learned. Choose the right robustness metric depending on the task and threat model: Single-label classification, multi-label classification, and ordinal classification metrics reflect different vulnerabilities. Identify the risks and their impacts, and evaluate the robustness of robust models use strong baselines and risky attack strategies, e.g., with larger attack budgets.

We do not intend for this paper to be the definitive answer, nor that the items contained above are exhaustive. We encourage future research to confront their protocol with actual use cases, and we hope that our work paves the way to critical thinking about the protocols of adversarial attack evaluation in the real world.

References

[1] L. T. Kohn, J. M. Corrigan, M. S. Donaldson, et al., Errors in health care: a leading cause of death and injury (2000).

[2] L. H. Garland, Studies on accuracy of diagnostic procedures, AJR 82 (1959) 25–38.

[3] L. P. Busby, J. L. Courtier, C. M. Glastonbury, Bias in radiology: The how and why of misses and misinterpretations, RadioGraphics 38 (2018) 236–247. URL: https://doi.org/10.1148/rg.2018170107. doi:10.1148/rg.2018170107. arXiv:https://doi.org/10.1148/rg.2018170107, pMID: 29194009.

[4] G. S. Lodwick, T. E. Keats, J. P. Dorst, The coding of roentgen images for computer analysis as applied to lung cancer, Radiology 81 (1963) 185–200. URL: https://doi.org/10.1148/81.2.185. doi:10.1148/81.2.185. arXiv:https://doi.org/10.1148/81.2.185, pMID: 14053755.

[5] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, A. Y. Ng, Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017. arXiv:1711.05225.

[6] L. Yao, J. Prosky, B. Covington, K. Lyman, A strong baseline for domain adaptation and generalization in medical imaging, 2019. arXiv:1904.01638.

[7] C. Qin, D. Yao, Y. Shi, Z. Song, Computer-aided detection in chest radiography based on artificial intelligence: a survey, BioMedical Engineering OnLine 17 (2018).

[8] E. H. P. Pooch, P. L. Ballester, R. C. Barros, Can we trust deep learning models diagnosis? the impact of domain shift in chest radiograph classification, 2020. arXiv:1909.01940.

[9] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, A. Saalbach, Comparison of deep learning approaches for multi-label chest x-ray classification, 2019. arXiv:1803.02315.

[10] L. Oakden-Rayner, Exploring large scale public medical image datasets, 2019. arXiv:1907.12720.

[11] J. P. Cohen, M. Hashir, R. Brooks, H. Bertrand, On the limits of cross-domain generalization in automated x-ray prediction, 2020. arXiv:2002.02497.

[12] S. Finlayson, I. Kohane, A. Beam, Adversarial attacks against medical deep learning systems (2018).

[13] S. A. Taghanaki, A. Das, G. Hamarneh, Vulnerability analysis of chest x-ray image classification against adversarial attacks (2018). arXiv:1807.02905.

[14] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, F. Lu, Understanding adversarial attacks on deep learning based medical image analysis systems, Pattern Recognition 110 (2021) 107332. URL: https://www.sciencedirect.com/science/article/pii/S0031320320301357. doi:https://doi.org/10.1016/j.patcog.2020.107332.

[15] X. Li, D. Zhu, Robust detection of adversarial attacks on medical images, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 1154–1158. doi:10.1109/ISBI45749.2020.9098628.

[16] B. Biggio, B. Nelson, P. Laskov, Poisoning attacks against support vector machines, arXiv preprint arXiv:1206.6389 (2012).

[17] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, F. Roli, Evasion attacks against machine learning at test time, in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), volume 8190 LNAI, 2013, pp. 387–402. URL: http://arxiv.org/abs/1708.06131http://dx.doi.org/10.1007/978-3-642-40994-3_25. doi:10.1007/978-3-642-40994-3{_}25.

[18] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples (2015). arXiv:1412.6572.

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

[20] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks (2019). arXiv:1706.06083.

[21] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, J. Li, Boosting adversarial attacks with momentum, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 9185–9193.

[22] F. Croce, M. Hein, Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks (2020). arXiv:2003.01690.

[23] T. Simonetto, S. Dyrmishi, S. Ghamizi, M. Cordy, Y. L. Traon, A unified framework for adversarial attack and defense in constrained feature space, arXiv preprint arXiv:2112.01156 (2021).

[24] S. Ghamizi, M. Cordy, M. Gubri, M. Papadakis, A. Boystov, Y. Le Traon, A. Goujon, Search-based adversarial testing and improvement of constrained credit scoring systems, in: Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2020, Association for Computing Machinery, New York, NY, USA, 2020, p. 1089–1100. URL: https://doi.org/10.1145/3368089.3409739. doi:10.1145/3368089.3409739.

[25] S. Ghamizi, M. Cordy, M. Papadakis, Y. Le Traon, Evasion attack steganography: Turning vulnerability of machine learn-

ing to adversarial attacks into a real-world application, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2021, pp. 31–40.

[26] S. Ghamizi, M. Cordy, M. Papadakis, Y. L. Traon, Adversarial robustness in multi-task learning: Promises and illusions, Proceedings of the AAAI Conference on Artificial Intelligence 36 (2022) 697–705. URL: https://ojs.aaai.org/index.php/AAAI/article/view/19950. doi:10.1609/aaai.v36i1.19950.

[27] S. Asgari Taghanaki, A. Das, G. Hamarneh, Vulnerability analysis of chest x-ray image classification against adversarial attacks, in: Understanding and interpreting machine learning in medical image computing applications, Springer, 2018, pp. 87–94.

[28] 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, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017). URL: http://dx.doi.org/10.1109/CVPR.2017.369. doi:10.1109/cvpr.2017.369.

[29] S. G. Finlayson, H. W. Chung, I. S. Kohane, A. L. Beam, Adversarial attacks against medical deep learning systems, arXiv preprint arXiv:1804.05296 (2018).

[30] X. Ma, Y. Niu, L. Gu, Y. Wang, Y. Zhao, J. Bailey, F. Lu, Understanding adversarial attacks on deep learning based medical image analysis systems, Pattern Recognition 110 (2021) 107332.

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[32] A. Krizhevsky, Learning multiple layers of features from tiny images, University of Toronto (2012).

[33] G. de Lacey, S. Morley, L. Berman, 1 - chest radiology: The basic basics, in: G. de Lacey, S. Morley, L. Berman (Eds.), The Chest X-Ray: A Survival Guide, W.B. Saunders, Edinburgh, 2008, pp. 2–13. URL: https://www.sciencedirect.com/science/article/pii/B9780702030468500065. doi:https://doi.org/10.1016/B978-0-7020-3046-8.50006-5.

[34] J. R. Ledford, Chest radiology: Plain film patterns and differential diagnoses, 6th ed., American Journal of Roentgenology 197 (2011) W1159–W1159. URL: https://doi.org/10.2214/AJR.11.7214. doi:10.2214/AJR.11.7214. arXiv:https://doi.org/10.2214/AJR.11.7214.

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

[36] S. Kaviani, K. J. Han, I. Sohn, Adversarial attacks and defenses on ai in medical imaging informatics: A survey, Expert Systems with Applications (2022) 116815.

[37] B. Tian, Q. Guo, F. Juefei-Xu, W. Le Chan, Y. Cheng, X. Li, X. Xie, S. Qin, Bias field poses a threat to dnn-based x-ray recognition, in: 2021 IEEE international conference on multimedia and expo (ICME), IEEE, 2021, pp. 1–6.

[38] 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 (2020) e0243963.

[39] B. Pal, D. Gupta, M. Rashed-Al-Mahfuz, S. A. Alyami, M. A. Moni, Vulnerability in deep transfer learning models to adversarial fast gradient sign attack for covid-19 prediction from chest radiography images, Applied Sciences 11 (2021) 4233.

[40] 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.

[41] A. Rahman, M. S. Hossain, N. A. Alrajeh, F. Alsolami, Adversarial examples—security threats to covid-19 deep learning systems in medical iot devices, IEEE Internet of Things Journal 8 (2020) 9603–9610.

[42] D. Anand, D. Tank, H. Tibrewal, A. Sethi, Self-supervision vs. transfer learning: robust biomedical image analysis against adversarial attacks, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), IEEE, 2020, pp. 1159–1163.

[43] V. Kovalev, D. Voynov, Influence of control parameters and the size of biomedical image datasets on the success of adversarial attacks, in: International Conference on Pattern Recognition and Information Processing, Springer, 2019, pp. 301–311.

[44] H. Hirano, A. Minagi, K. Takemoto, Universal adversarial attacks on deep neural networks for medical image classification, BMC medical imaging 21 (2021) 1–13.

[45] F.-F. Xue, J. Peng, R. Wang, Q. Zhang, W.-S. Zheng, Improving robustness of medical image diagnosis with denoising convolutional neural networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 846–854.

[46] A. M. Tripathi, A. Mishra, Fuzzy unique image transformation: Defense against adversarial attacks on deep covid-19 models, arXiv preprint arXiv:2009.04004 (2020).

[47] M. Xu, T. Zhang, Z. Li, M. Liu, D. Zhang, Towards evaluating the robustness of deep diagnostic models by adversarial attack, Medical Image Analysis 69 (2021) 101977.

[48] X. Li, D. Zhu, Robust detection of adversarial attacks on medical images, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), IEEE, 2020, pp. 1154–1158.

[49] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, arXiv preprint arXiv:1706.06083 (2017).

[50] A. Athalye, N. Carlini, D. Wagner, Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples, 2018. arXiv:1802.00420.

[51] W. He, J. Wei, X. Chen, N. Carlini, D. Song, Adversarial example defense: Ensembles of weak defenses are not strong, in: 11th {USENIX} workshop on offensive technologies ({WOOT} 17), 2017.

[52] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, A. Kurakin, On evaluating adversarial robustness, 2019. arXiv:1902.06705.

[53] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, M. Hein, Robustbench: a standardized adversarial robustness benchmark, arXiv preprint arXiv:2010.09670 (2020).

[54] M. Paschali, S. Conjeti, F. Navarro, N. Navab, Generalizability vs. robustness: Adversarial examples for medical imaging, 2018. arXiv:1804.00504.

[55] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, A. Y. Ng, Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019. arXiv:1901.07031.

[56] A. Bustos, A. Pertusa, J.-M. Salinas, M. de la Iglesia-Vayá, Padchest: A large chest x-ray image dataset with multi-label annotated reports, Medical Image Analysis 66 (2020) 101797. URL: http://dx.doi.org/10.1016/j.media.2020.101797. doi:10.1016/j.media.2020.101797.

[57] Q. Song, H. Jin, X. Huang, X. Hu, Multi-label adversarial perturbations, 2019. arXiv:1901.00546.

[58] E. Frank, M. Hall, A simple approach to ordinal classification, in: European conference on machine learning, Springer, 2001, pp. 145–156.