Certification of Deep Learning Models for Medical Image Segmentation

Othmane Laousy[1,2,3](✉), Alexandre Araujo[4], Guillaume Chassagnon[2], Nikos Paragios[5], Marie-Pierre Revel[2], and Maria Vakalopoulou[1,3]

[1] MICS, CentraleSupélec, Paris-Saclay University, Gif-sur-Yvette, France
othmane.laousy@centralesupelec.fr
[2] Hôpital Cochin, AP-HP, Paris-Cité University, Paris, France
[3] Inria Saclay, Gif-sur-Yvette, France
[4] New York University, New York, USA
[5] Therapanacea, Paris, France

Abstract. In medical imaging, segmentation models have known a significant improvement in the past decade and are now used daily in clinical practice. However, similar to classification models, segmentation models are affected by adversarial attacks. In a safety-critical field like healthcare, certifying model predictions is of the utmost importance. Randomized smoothing has been introduced lately and provides a framework to certify models and obtain theoretical guarantees. In this paper, we present for the first time a certified segmentation baseline for medical imaging based on randomized smoothing and diffusion models. Our results show that leveraging the power of denoising diffusion probabilistic models helps us overcome the limits of randomized smoothing. We conduct extensive experiments on five public datasets of chest X-rays, skin lesions, and colonoscopies, and empirically show that we are able to maintain high certified Dice scores even for highly perturbed images. Our work represents the first attempt to certify medical image segmentation models, and we aspire for it to set a foundation for future benchmarks in this crucial and largely uncharted area.

Keywords: Certified Robustness · Randomized Smoothing · Denoising Diffusion Models · Segmentation

1 Introduction

For the past decade, deep neural networks have dominated the computer vision community and provided near human performance on many different tasks, including classification [18], segmentation [24], and image generation [16]. Given these impressive results, convolutional neural networks are now used on a daily basis in fields like healthcare, self-driving cars, and robotics, to cite a few. In medical imaging, convolutional neural networks are particularly used to segment organs or regions of interest on different modalities such as X-rays, CT

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43901-8_58.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14223, pp. 611–621, 2023.
https://doi.org/10.1007/978-3-031-43901-8_58

scans, MRIs, or ultrasound [36]. Indeed, segmentation techniques and variations of 2D and 3D U-Nets are currently the state-of-the-art to identify and isolate tumors, blood vessels, organs, or other structures within an image and provide crucial help to physicians for medical diagnosis, screening, and prognosis [32].

Nowadays, segmentation models are gaining widespread adoption in modern clinical practice and are being used with increasing frequency, making the results of these models critical for many patients. However, it is now commonly known that neural networks can be vulnerable to adversarial attacks [17,34], *i.e.*, small input perturbations invisible to humans crafted specifically such that the network performs errors. Over the past few years, a large body of work has devised empirical defenses against adversarial attacks for classification tasks [3,17,25], as well as segmentation tasks [37], including applications on medical imaging [27]. Although state-of-the-art empirical defenses provide significant robustness, these defenses do not guarantee *theoretical* robustness and stronger attacks can be crafted to break them [5]. Recently, *certified* defenses, for classification [2,11,26] and segmentation [15,23], have been proposed to guarantee the accuracy and reliability of neural networks. However, certified defenses for segmentation in the context of medical imaging are still lacking, even if models are getting market approvals (*e.g.*, FDA, CE) and are already adopted in clinical practice.

In this paper, we provide the first method for certified robustness in the context of segmentation for medical imaging. We leverage the *randomized smoothing* strategy [11,15], and the recent work on *diffusion models* [7] to achieve state-of-the-art certified robustness for segmentation models. Randomized smoothing consists in convolving the neural network with a Gaussian distribution (*i.e.*, by adding noise to the input) in order to obtain a smooth segmentation model. From the smoothness properties of the segmentation model, we can derive a robustness guarantee and compute a certified Dice score. We go even further by using diffusion models to first denoise the perturbed input and boost the certified robustness. By extension, we show that current diffusion models, trained on 'classical images' generalize well to medical datasets for denoising tasks. Extensive experiments on five public medical datasets of chest X-rays [21,31], skin lesions [10], and colonoscopies [6], and different popular segmentation models, prove the potential of our method. We hope that this study will provide the first step towards robustness guarantees for medical image segmentation.

2 Related Work

Since the discovery of adversarial attacks [17,34], numerous defenses [8,17,25] and attacks have been devised [8,25], demonstrating that neural networks are sensitive to small input perturbation and vulnerable to attacks. Adversarial training, which has been acknowledged as one of the most successful empirical defenses, consists in training a network directly on adversarial examples [25]. However, it is now known that even strong defenses can be bypassed by adaptive attacks [12]. Paschali et al. [27] were among the first to study adversarial attacks in the context of medical imaging. They conducted experiments using several neural network architectures [20,33] (*i.e.*, Inception V3, V4, MobileNet) and several attacks [17,25] to demonstrate that the vulnerability of neural networks is extended to medical images.

More specifically, in the context of classification, a previous work [4] has analyzed the robustness of neural networks for chest X-ray images and showed that gradient-based attacks were successful in fooling both machines and humans. In a similar line of work, Yao et al. [38] proposed an add-on to known attacks that bypasses state-of-the-art adversarial detectors making current defenses even less robust. On the other hand, several works have been focused on crafting defense strategies specifically in the context of medical imaging. For example, Almalik et al. [1] proposed a self-ensembling method to enhance the robustness of Vision Transformers in the presence of adversarial attacks. In the context of segmentation in medical imaging, [30] introduced a vector quantization approach by learning a discrete representation in a low dimensional embedding space and improving the robustness of a segmentation model. Finally, Daza et al. [13] proposed a lattice architecture that segments organs and lesions on MRI and CT scans and leveraged an efficient approach of adversarial training to defend against adversarial examples.

Although a large body of work has focused on constructing defenses for classification and segmentation tasks in the context of medical imaging, *certified* defenses are under-studied by the medical community. In this paper, we propose to leverage randomized smoothing and diffusion models for certified segmentation on medical datasets, setting the first baseline for this challenging problem and certifying popular segmentation architectures.

3 Randomized Smoothing

Randomized smoothing is a model agnostic technique, proposed by Cohen et al. [11], used to improve and certify the robustness of neural networks against adversarial attacks. This method consists in adding random noise (*e.g.*, noise generated from a Gaussian distribution) to the input data and then classifying the perturbed data using the neural network. Let $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$ denote the data distribution where $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = \{1, \ldots, k\}$ represent the input space and target space respectively and k is the number of classes. Let $f : \mathcal{X} \to \mathcal{Y}$ be a neural network such that for $(x, y) \in \mathcal{D}$, the classifier correctly classifies if $f(x) = y$. An adversarial attack is a small norm-bounded perturbation $\delta \in \mathbb{R}^d$ with $\|\delta\|_2 \leq \epsilon$ such that: $f(x + \delta) \neq y$. Randomized smoothing is a procedure to construct a new *smooth* classifier g given any base classifier f. Let $\mathcal{N}(0, \sigma^2\mathbf{I})$ be a Gaussian distribution of mean 0 and variance σ, then, the smooth classifier g is defined as follows:

$$g(x) = \mathbb{P}_{\eta \sim \mathcal{N}(0,\sigma^2\mathbf{I})} \left[f(x + \eta) = y \right]$$

Cohen et al. [11] have shown that if $R = \sigma \Phi^{-1}(g(x))$ where Φ is the cumulative distribution function of the standard Gaussian distribution and R can be considered the certified radius, then, $g(x + \delta) = y$ for all δ satisfying $\|\delta\|_2 \leq R$.

However, since it is not possible to compute g at x exactly, they proposed using Monte Carlo algorithms as an alternative approach for estimating $g(x)$ using random sampling. In order to obtain a reliable estimate of the probability $g(x)$, they also suggested a method that involves generating n samples of η from a normal distribution $\mathcal{N}(0, \sigma^2\mathbf{I})$ and evaluating $f(x + \eta)$ for each sample. The resulting counts for each class in $\mathcal{Y}$ are then used to estimate probability p_y and the radius R with confidence $1 - \alpha$ (where α is a value between 0 and 1). If the confidence level cannot be achieved (for example, due to insufficient samples), the method will abstain from providing an estimate. More recently, Fischer et al. [15] built upon the work of [11] by introducing SegCertify, the first certified approach for image segmentation. The segmentation process involves assigning a segmentation class to every pixel in the image, which can be viewed as a form of classification at the pixel level. In the segmentation settings, the output space consists of regions or categories to be segmented, such as cars, roads, pedestrians, etc. The classifier function $f : \mathcal{X} \rightarrow \mathcal{Y}^d$ determines the class for each pixel and categorizes each component independently. In this context, the certification algorithm proposed by Cohen et al. [11] can be extended to accommodate the segmentation task.

To obtain a smooth classifier, it is necessary to add random noise to the input of the classifier. However, this creates a trade-off between accuracy and robustness. If low variance noise is added, accuracy won't be impacted significantly, but the certified radius will remain low. Conversely, adding high variance noise can improve certificates but at the expense of accuracy. To address this issue, Cohen et al. proposed a simple trick of training the network with noise injection during the training phase. While this method may reduce accuracy when evaluating the classifier with noise during the certification process, it can also help mitigate the trade-off between accuracy and robustness. One can note that during training, the network's objective is to learn to ignore the noise and classify at the same time. To improve the natural as well as the certified accuracy, Salman et al. [29] proposed to separate the two tasks with two networks trained separately. First, a network, $h : \mathcal{X} \rightarrow \mathcal{X}$, is trained to denoise the data such that for $\eta \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, we have $h(x + \eta) \approx x$, then, the output of the denoiser is given to the classifier.

In this paper, we leverage randomized smoothing and diffusion probabilistic models to obtain state-of-the-art results on certified segmentation for medical imaging. To the best of our knowledge, we are the first to propose a comprehensive study on certified segmentation for medical imaging.

4 Diffusion Probabilistic Models for Certification

The training of a Denoising Diffusion Probabilistic Model (DDPM) is an iterative process that involves adding a small amount of noise at every step of the diffusion process until random noise is reached. The reverse process then starts from random noise and generates a new image that conforms to the data distribution. Since DDPMs are inherently iterative denoising models, we can leverage this property for randomized smoothing. The idea would be to start the reverse process with a noisy image, rather than Gaussian noise, enabling the DDPM to output an image that resembles the original image.

Similar to Carlini et al. [7], our proposed pipeline is composed of two main steps: we denoise, then we certify. In order to complete the denoising, we need to first map between the noise model utilized in diffusion models and the one used in randomized smoothing. Randomized smoothing needs a data point that is enhanced with Gaussian noise added to it, given by $x_{\text{rs}} = x + \delta$ with $\delta \sim \mathcal{N}(x, \sigma^2 \mathbf{I})$. On the other hand, diffusion models suppose the noise model for $x_{\text{DDPM}} \sim \mathcal{N}(\sqrt{\alpha_t} x, (1 - \alpha_t)\mathbf{I})$. Programmatically, we start by adding Gaussian noise to an image x, obtaining x_{rs}. Then the timestep t^* on which we can use the diffusion model for randomized smoothing is defined. Depending on the scheduler of the denoiser, we compute t^* such that $\sigma^2 = \frac{1-\alpha_{t^*}}{\alpha_{t^*}}$ (obtained by scaling x_{rs} with $\sqrt{\alpha_t}$ and pairing the variances). We then calculate $x_{\text{DDPM}} = \sqrt{\alpha_{t^*}}(x + \delta), \delta \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$. After that, we apply a single-step denoiser and predict the completely denoised image. A single-step denoising involves directly predicting the image from t^* to $t = 0$. A multi-step denoising strategy implies iteratively predicting all images at t^*, $t^* - 1, \dots$ until $t = 0$. Both techniques are explored in the next section and supplementary material.

Since randomized smoothing is applied to each pixel separately with a probability of $1 - \alpha$, considering the entire segmentation region would imply considering a union bound with significantly reduced confidence. Similar to Fischer et al. [15], we leverage the Holm-Bonferroni method [19] and perform multiple-testing corrections. For each image, we repeat this process $n = 100$ times, identifying pixels on which the model abstains, and computing the certified scores. We extend the work of Fischer et al. [15] to also compute a certified Dice score that is calculated ignoring the abstain class ($\oslash$). Our approach has a significant advantage compared to SegCertify since it leverages off-the-shelf and state-of-the-art pre-trained denoisers and segmentation models. SegCertify, on the other hand, relies on models trained with Gaussian noise.

5 Experiments and Results

Datasets: We perform experiments on 5 different publicly available datasets. All datasets were divided to 70% for training, 10% for validation, and 20% for testing. The testing set is the one used to compute certified results.

Chest X-rays Datasets: JSRT dataset [31] with annotations of lung, heart, and clavicles provided by [35] is used. This dataset contains 247 images. For lung segmentation only, we use both the Montgomery and Shenzen datasets [21]. Montgomery consists of 138 and Shenzen of 662 annotated images.

Skin Lesion: Skin images with their annotations provided by the ISIC 2018 boundary segmentation challenge [10] were used. This dataset consists of 2694 RGB dermatoscopy images.

Table 1. Comparison of our approach with three different model architectures on chest X-ray datasets. We report certified Dice, IoU and percentage of abstentions ($\%\oslash$) for different noise levels $\boldsymbol{\sigma}$ and radii $\boldsymbol{R}$.

$\boldsymbol{\sigma}$	$\boldsymbol{R}$	**JSRT**							**Montgomery**			**Schenzen**		
		Lung		Heart		Clavicles			Lung			Lung		
		Dice	IoU	Dice	IoU	Dice	IoU	$\%\oslash$	Dice	IoU	$\%\oslash$	Dice	IoU	$\%\oslash$
UNet [28]														
0.25	0.17	0.94	0.91	0.88	0.79	0.75	0.63	0.07	0.93	0.89	0.07	0.95	0.90	0.05
0.50	0.34	0.90	0.83	0.88	0.79	0.61	0.45	0.09	0.89	0.80	0.07	0.93	0.90	0.02
1.00	0.67	0.87	0.79	0.84	0.75	0.23	0.15	0.15	0.88	0.80	0.14	0.89	0.83	0.10
ResUNet++ [22]														
0.25	0.17	0.95	0.91	0.93	0.87	0.78	0.65	0.05	0.96	0.93	0.02	0.95	0.91	0.01
0.50	0.34	0.94	0.88	0.91	0.83	0.63	0.48	0.08	0.94	0.89	0.03	0.93	0.90	0.02
1.00	0.67	0.90	0.82	0.87	0.77	0.28	0.19	0.12	0.89	0.83	0.07	0.90	0.85	0.06
DeeplabV2 [9]														
0.25	0.17	0.94	0.91	0.91	0.86	0.85	0.75	0.04	0.93	0.91	0.07	0.80	0.71	0.07
0.50	0.34	0.88	0.81	0.87	0.79	0.63	0.49	0.10	0.91	0.87	0.02	0.34	0.25	0.15
1.00	0.67	0.88	0.80	0.83	0.74	0.20	0.11	0.14	0.85	0.79	0.17	0.04	0.02	0.11

Colonoscopy Images: CVC-ClinicDB dataset [6] containing 612 colonoscopy images in RGB together with their annotations were utilized.

Implementation Details: We train three different segmentation models namely, a UNet [28], a ResUNet++ [22], and a DeeplabV2 [9] with and without noise. The models trained without noise are used exclusively with our method. The models trained with a Gaussian noise of 0.25 are used to compute SegCertify scores. All 6 models use an image input size of 512 × 512 for X-ray images, 384 × 512 for skin lesions, and 288 × 384 for colonoscopy. As a denoiser, we use an off-the-shelf denoising diffusion probabilistic model provided by [14]. We perform our experiments with the 256 × 256 class unconditional denoiser with a linear scheduler and without timestep respacing. For each noise level, our method follows the steps described in the previous section and uses $n_0 = 10$, n=100 for each image, and $\alpha = 0.001$, and $\tau = 0.75$. Our code is made publicly available at: https://github.com/othmanela/medical_cert_seg.

Results and Discussion: For all five datasets, we compute a certified Dice score and certified mean Intersection over Union (IoU). We also report the percentage of abstentions ($\%\oslash$) representing the mean number of pixels on which the model's prediction confidence was insufficient with respect to the radius R. The lower the percentage of abstentions the better the segmentation model is.

In Table 1, we compare our method using 3 different and popular architectures (UNet, ResUNet++, and DeeplabV2) on the chest X-rays datasets. We

Table 2. Certified segmentation results of our technique and SegCertify [15] on the chest X-ray JSRT dataset. We report Dice, IoU, and percentage of abstentions (%⊘) for each class.

Model	**Trained with noise**	$\boldsymbol{\sigma}$	$\boldsymbol{R}$	**Lung**		**Heart**		**Clavicles**		
				Dice	IoU	Dice	IoU	Dice	IoU	%⊘
ResUNet++ [22]	✗	0.00	0.00	0.97	0.94	0.94	0.91	0.93	0.91	0.00
	✓	0.25	0.00	0.91	0.90	0.89	0.87	0.84	0.79	0.00
SegCertify [15]	✓	0.25	0.17	**0.96**	**0.92**	**0.93**	**0.88**	**0.83**	**0.72**	**0.04**
	✓	0.50	0.34	0.89	0.84	0.85	0.79	0.58	0.43	0.13
	✓	1.00	0.67	0.07	0.04	0.02	0.01	0.00	0.00	0.24
Ours	✗	0.25	0.17	0.95	0.91	**0.93**	0.87	0.78	0.65	0.05
	✗	0.50	0.34	**0.94**	**0.88**	**0.91**	**0.83**	**0.63**	**0.48**	**0.08**
	✗	1.00	0.67	**0.90**	**0.82**	**0.87**	**0.77**	**0.28**	**0.19**	**0.12**

notice that our method maintains overall good results on all three model backbones. A similar table with SegCertify results is provided in Table S2 of the supplementary material. Overall, for both methods, ResUNet++ is the most robust architecture followed by UNet and then DeeplabV2 for all $\boldsymbol{\sigma}$ and $\boldsymbol{R}$ combinations. Moreover, certified metrics for lungs and heart remain high for our method, even with high levels of noise. However, the increasing level of noise affects the clavicles since these are smaller structures.

A comparison of our method and SegCertify using the ResUNet++ architecture is presented in Table 2 for the three chest X-ray datasets. We observe that we outperform SegCertify, especially for high sigma values. For $\sigma = 0.25$, SegCertify performs slightly better. This is due to the fact that the model used with SegCertify is trained with a noise level of 0.25. The main drawback however is that its Dice on unperturbed images drops considerably (*e.g.*, from 0.96 to 0.91 on lung segmentation). On the other hand, our pipeline does not require training a segmentation model with noise or even a denoising model. Our methodology relies only on off-the-shelf models. For the highest noise level of $\sigma = 1.0$, we notice that the certified Dice and IoU with SegCertify both drop to 0 whereas our proposed method is able to maintain high certified scores.

Qualitative results are provided in Fig. 1 for our proposed method and SegCertify for the different datasets and different levels of noise. Regarding the structures to segment, we notice that the abstentions around the clavicles (the smallest benchmarked region of interest on chest X-rays) get bigger. We also notice that the fine segmentation boundaries (*e.g.*, area around the skin lesion) may not be as sharp after denoising. As we increase the noise, the decision boundary is harder to find for all models. This may be due to the fact that fine details on the image are lost after the denoising step. However, our method is still able to segment the large majority of pixels properly on the image, contrary to its competitor, especially for high noise levels (third row on chest X-rays).

Table 3 reports certified segmentation results for skin lesions and colonoscopy on both techniques. We notice that our method is still performing better than

Table 3. Results on skin lesions [10] and CVC-ClinicDB [6] segmentation.

Model	**Method**	σ	R	**Skin Lesions**			**CVC-ClinicDB**		
				Dice	IoU	%⊘	Dice	IoU	%⊘
ResUNet++ [22]	SegCertify [15]	0.25	0.17	0.79	0.68	0.07	0.63	0.56	0.05
		0.50	0.34	0.41	0.27	0.06	0.15	0.10	0.01
		1.00	0.67	0.00	0.00	0.01	0.00	0.00	0.00
	Ours	0.25	0.17	0.85	0.77	0.03	0.65	0.57	0.04
		0.50	0.34	0.83	0.76	0.04	0.45	0.39	0.07
		1.00	0.67	0.77	0.69	0.06	0.26	0.23	0.14

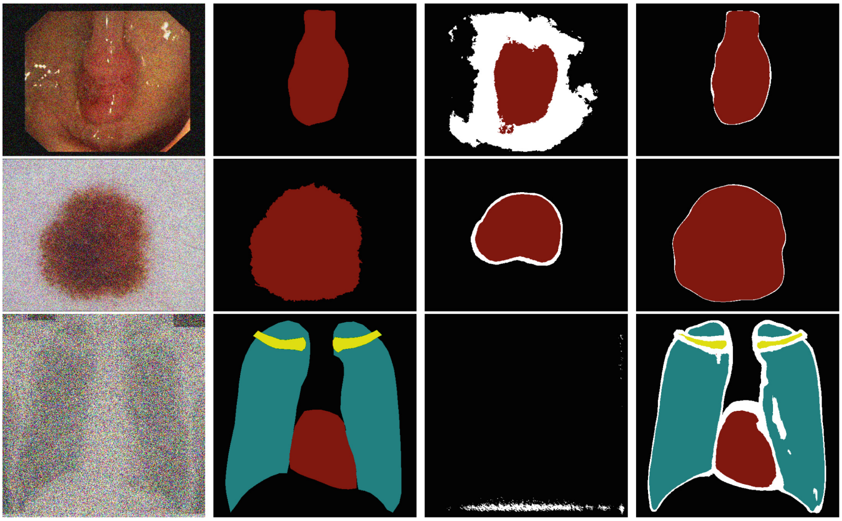

Fig. 1. Qualitative results of SegCertify and our method on colonoscopy, skin lesion, and chest X-ray images. From left to right: image with added noise, ground truth, SegCertify segmentation, our segmentation. White pixels denote abstention areas of the segmentation models. We increase the noise level from top to bottom: $\sigma = 0.25, 0.5$, and 1.0.

SegCertify. This supports our claim that DDPMs generalize quite well to medical images and that harnessing their potential boosts the state-of-the-art. Regarding the denoiser, we used a single-step denoising strategy, *i.e.*, we perform a single call to the DDPM to compute the denoised image from t^* to $t = 0$. Another strategy could be to iteratively denoise from t^*, $t^* - 1$, ... until $t = 0$. However, this implies predicting a denoised image multiple times and in the end, may result in images with unwanted artifacts. We perform multi-step denoising experiments and report results in Table S1 of the supplementary material. We note that the single-step denoising performs best since it relies more on the denoising power of DDPMs rather than their generative capabilities, and is also faster than the multi-step approach. Finally, we perform a comparison with another denoiser architecture. We train three UNet models (one for each noise level) on the JSRT dataset. We report results in Table S3 and notice that even with custom-trained denoisers, the DDPM outperforms the UNet denoising architecture. A comparison of denoised images is provided in Figure S1. We notice that the DDPM is able to keep high-fidelity images compared to the UNet and is therefore more relevant for certified medical image segmentation.

6 Conclusion

In this paper, we present the first work on certified segmentation for medical imaging, and extensively evaluate it on five different datasets and three deep learning segmentation models. Our technique leverages off-the-shelf denoising and segmentation models and provides the highest certified Dice and mIoU on multi-class and binary segmentation of five different datasets. With that, we are able to remove the overhead of having to train and fine-tune models specifically for robustness. This paradigm shift alleviates the dilemma of having to choose between highly accurate segmentation models or models robust to attacks. We hope that this work serves as a baseline for the unexplored yet critical topic of certified segmentation in medical imaging. Future work will involve extending our approach to 3D medical imaging modalities as well as exploring the realm of certified classification.

Acknowledgements. This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011013308R1 made by GENCI and it was partially supported by the ANR Hagnodice ANR-21-CE45-0007.

References

1. Almalik, F., Yaqub, M., Nandakumar, K.: Self-ensembling vision transformer (sevit) for robust medical image classification. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13433, pp. 376–386. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-16437-8_36
2. Araujo, A., Havens, A., Delattre, B., Allauzen, A., Hu, B.: A unified algebraic perspective on lipschitz neural networks. In: ICLR (2023)
3. Araujo, A., Meunier, L., Pinot, R., Negrevergne, B.: Advocating for multiple defense strategies against adversarial examples. In: ECML (2020)
4. Asgari Taghanaki, S., Das, A., Hamarneh, G.: Vulnerability analysis of chest x-ray image classification against adversarial attacks. In: iMIMIC (2018)
5. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. In: ICML (2018)
6. Bernal, J., et al.: Wm-dova maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput. Med. Imaging Graph. **43**, 99–111 (2015)
7. Carlini, N., Tramer, F., Kolter, J.Z., et al.: (certified!!) adversarial robustness for free!. In: ICLR (2023)
8. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: IEEE on Security and Privacy (2017)

9. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Pattern Anal. Mach. Intell. **40**, 834–848 (2016)
10. Codella, N.C.F., et al.: Skin lesion analysis toward melanoma detection 2018: a challenge hosted by the international skin imaging collaboration (ISIC). CoRR (2019)
11. Cohen, J., Rosenfeld, E., Kolter, Z.: Certified adversarial robustness via randomized smoothing. In: ICML (2019)
12. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: ICML (2020)
13. Daza, L., Pérez, J.C., Arbeláez, P.: Towards robust general medical image segmentation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 3–13. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_1
14. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: NeurIPS (2021)
15. Fischer, M., Baader, M., Vechev, M.: Scalable certified segmentation via randomized smoothing. In: ICML (2021)
16. Goodfellow, I., et al.: Generative adversarial networks. ACM (2020)
17. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv:1412.6572 (2014)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
19. Holm, S.: A simple sequentially rejective multiple test procedure. J. Stat. **6**, 65–70 (1979)
20. Howard, A.G., et al.: Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861 (2017)
21. Jaeger, S., Candemir, S., Antani, S., Wáng, Y.X.J., Lu, P.X., Thoma, G.: Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. Quant. Imaging Med. Surg. **4**, 475 (2014)
22. Jha, D., et al.: Resunet++: an advanced architecture for medical image segmentation. In: IEEE on Multimedia (2019)
23. Laousy, O., et al.: Towards better certified segmentation via diffusion models. In: UAI (2023)
24. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
25. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083 (2017)
26. Meunier, L., Delattre, B., Araujo, A., Allauzen, A.: A dynamical system perspective for lipschitz neural networks. In: ICML (2022)
27. Paschali, M., Conjeti, S., Navarro, F., Navab, N.: Generalizability *vs.* robustness: investigating medical imaging networks using adversarial examples. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 493–501. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_56
28. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
29. Salman, H., Sun, M., Yang, G., Kapoor, A., Kolter, J.Z.: Denoised smoothing: a provable defense for pretrained classifiers. In: NeurIPS (2020)

30. Santhirasekaram, A., Kori, A., Winkler, M., Rockall, A., Glocker, B.: Vector quantisation for robust segmentation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13434, pp. 663–672. Springer, Heidelberg (2022). https://doi.org/10.1007/978-3-031-16440-8_63
31. Shiraishi, J., et al.: Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. Am. J. Roentgenol. **174**, 71–74 (2000)
32. Siddique, N., Paheding, S., Elkin, C.P., Devabhaktuni, V.: U-net and its variants for medical image segmentation: a review of theory and applications. IEEE Access **9**, 82031–82057 (2021)
33. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI (2017)
34. Szegedy, C., et al.: Intriguing properties of neural networks. arXiv:1312.6199 (2013)
35. Van Ginneken, B., Stegmann, M.B., Loog, M.: Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. Med. Image Anal. **10**, 19–40 (2006)
36. Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., Nandi, A.K.: Medical image segmentation using deep learning: a survey. arXiv:2009.13120v3 (2020)
37. Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., Yuille, A.: Adversarial examples for semantic segmentation and object detection. In: CVPR (2017)
38. Yao, Q., He, Z., Lin, Y., Ma, K., Zheng, Y., Zhou, S.K.: A hierarchical feature constraint to camouflage medical adversarial attacks. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12903, pp. 36–47. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87199-4_4