CalCErt: Bin-wise Certification of Confidence Calibration in Medical Image Classification

Leo Fillioux[1*], Stergios Christodoulidis[1], Paul-Henry Cournède[1], Maria Vakalopoulou[1] and Jose Dolz[2]

[1]Université Paris-Saclay, CentraleSupélec, Gustave Roussy, INSERM, CDSU, IHU PRISM, Gif-sur-Yvette
[2]LIVIA, ILLS, ETS Montréal
Corresponding author: leo.fillioux@centralesupelec.fr

Abstract. Deep neural networks remain vulnerable to adversarial perturbations, which can distort not only predictions but also confidence scores, undermining uncertainty calibration. While existing certification methods focus on preserving the predicted category, providing guarantees on how calibration behaves under adversarial attacks remains overlooked. In this work, we introduce CalCErt, a simple and efficient post-hoc strategy that certifies bin-wise confidence calibration for any pretrained differentiable classifier. Our approach combines empirical calibration estimates, statistical concentration bounds, and local Lipschitz estimates of the confidence function to derive data-dependent upper bounds on worst-case miscalibration within an ℓ_2-ball of radius R. We evaluate CalCErt across 11 medical image classification tasks and multiple adversarial perturbations, demonstrating substantially higher certified coverage than baseline strategies while maintaining competitive tightness. Our code is available at github.com/leofillioux/calcert.

Keywords: Confidence calibration · Certification · Medical Imaging.

1 Introduction

Adversarial attacks can drastically alter the predictions of modern neural networks, even when perturbations are visually imperceptible [31]. Consequently, models that appear robust under current evaluation protocols may remain vulnerable to these attacks. Moreover, label attacks [14,20,24,29], the most prominent form of adversarial perturbations, affect not only prediction correctness, but also the confidence scores, thereby degrading overall uncertainty calibration. This is particularly concerning in safety-critical domains, such as medical image analysis [4,23], where confidence scores often play a central role in downstream decision-making, and where failures can have severe real-world consequences.

Given these risks, there is a growing need for formal guarantees on model behavior under perturbations. While adversarial training and empirical defenses

[*] Work done during an internship at ILLS.

can improve robustness, they provide no provable guarantees and are frequently broken by stronger attacks. In contrast, certified robustness [5,8,30] has emerged as a promising direction, providing mathematical guarantees that a model's output remains stable for all inputs within a prescribed ℓ_p-norm ball of radius R. Nevertheless, computing exact adversarial robustness is coNP-hard [19], and attack-based evaluations provide only heuristic upper bounds on vulnerability [6,32]. This has motivated a rich and growing line of work on certification methods, such as Randomized Smoothing [5,8], which provide provable guarantees on robust accuracy. Yet, despite this progress, existing certification frameworks focus almost exclusively on providing guarantees that the model predicted class remains unchanged [18,21] given a prescribed perturbation radius, but they do not address how confidence scores behave. This is a critical omission, as robustness of accuracy does not directly imply robustness of calibration, and a model may remain correct while becoming arbitrarily over- or under-confident.

To our knowledge, certifying uncertainty calibration in the medical domain remains unexplored. Existing method provides no guarantees on how calibration may deteriorate under adversarial perturbations, or how bin-wise miscalibration behaves. This gap is critical in healthcare, where confidence (rather than accuracy alone) drives downstream actions, motivating the question: *Can we certify the uncertainty calibration of a pretrained model under ℓ_2-bounded perturbations without retraining or attack-specific fine-tuning?*

Thus, we can summarize our main contributions as:

- We take a first step toward certifying uncertainty calibration in medical image classification under adversarial perturbations. To this end, we derive a tractable, data-dependent upper bound on the worst-case change in confidence within an ℓ_2-ball by estimating local Lipschitz constants of the model confidence function. This provides a principled way to control how far confidence values can drift under input perturbations.
- Building on this, we obtain certified upper bounds on *bin-wise* miscalibration by combining empirical calibration estimates, statistical concentration inequalities, and our Lipschitz-based control of confidence variation. The resulting certificate is post-hoc, requires no adversarial training and is compatible with any differentiable model. This makes our approach simple to implement and applicable to a wide range of pretrained classifiers.
- Extensive experiments across 11 classification tasks spanning multiple medical imaging modalities, five diverse perturbations, including adversarial attacks, and four backbones (leading to a total of 220 different settings) demonstrate that `CalCErt` yields usable, non-vacuous bounds for calibration certification.

2 Related Work

Certified robustness for deep networks aims at guaranteeing that model predictions remain stable under bounded adversarial perturbations. Empirical

defenses such as adversarial training [2,10,13] can improve robustness in practice but offer no formal guarantees. Certified defenses instead provide provable robustness under specific threat models, most commonly via *randomized smoothing* (RS) [5,8,25,27,28] or Lipschitz-controlled architectures [7,11,17,36]. RS [27] constructs a Gaussian-smoothed classifier and yields probabilistic guarantees on robust accuracy. However, it certifies only label invariance, i.e., no guarantees on confidence or calibration, and requires retraining or noise-specific tuning, making it unsuitable for our goal of post-hoc calibration analysis. *Lipschitz-based certification methods* [7,11,17,36] enforce global or local Lipschitz constraints during training, limiting output variation under bounded perturbations. However, they require architectural or training modifications and are therefore not compatible with our post-hoc, model-agnostic setting.

Robustness certification in medical imaging. Providing accurate and reliable models is essential in medical imaging, especially as AI systems increasingly receive regulatory approval (e.g., FDA, CE) and are deployed in clinical practice. Thus, recent work has begun exploring certified robustness in this domain. [21] presented a method to verify robustness in the context of medical imaging segmentation based on RS and diffusion models. More recently, [18] used diffusion- and prompt-based smoothing to improve robustness to noise and distribution shift in medical vision systems. These advances highlight that certified robustness is becoming a critical component of trustworthy medical imaging systems and represents an important step toward safe and reliable clinical deployment.

Unlike prior work, we focus on *certifying confidence calibration* under adversarial perturbations, a problem still unexplored in medical imaging, and take a first step toward extending certification beyond accuracy to confidence reliability.

3 Methodology

3.1 Problem setup

Let $f_\theta : \mathbb{R}^d \to [0,1]^C$ be C-class classifier with predicted probabilities $p(x) = f_\theta(x)$, where the predicted label is defined as $\hat{y}(x) = \arg\max_c p_c(x)$, and whose associated confidence is $p_{\hat{y}(x)}(x)$. A classifier is said to be well calibrated if its predicted confidence matches the true likelihood of correctness. Formally, for a prediction with confidence $p_{\hat{y}(x)}(x)$, the model is calibrated when $\mathbb{P}\big(y = \hat{y}(x) \,\big|\, p_{\hat{y}(x)}(x) = t\big) = t$ for all $t \in [0,1]$. In practice, this conditional relationship is approximated by using a binning scheme: the confidence interval $[0,1]$ is partitioned into K bins $B_k = [a_k, b_k]$, and for each bin we compute empirical accuracy $\hat{A}_k$, mean confidence $\hat{P}_k$, and miscalibration $\widehat{\mathrm{Cal}}_k = |\hat{A}_k - \hat{P}_k|$, based on a given calibration dataset $\mathcal{D}_{cal} = \{(x^{(n)}, y^{(n)})\}_{n=1}^N$. Our goal is to bound how much bin-wise calibration may change under additive ℓ_2-bounded perturbations.

3.2 Proposed confidence certification

A central difficulty of this problem is that, under perturbations $\|\delta\|_2 \le R$, both the predicted confidence and the predicted label may change in ways that are

hard to characterize directly. To obtain certified guarantees, we therefore require an upper bound on how much the model confidence can vary within the perturbation ball. Since evaluating the classifier on all possible perturbed inputs is infeasible, we approximate this worst-case change using a local Lipschitz bound on the confidence function. This provides a tractable way to control the effect of perturbations and forms the basis of our certified calibration analysis.

Local Lipschitz constants for confidence. For a fixed input x, let us define the scalar confidence function $g(x) = p_{\hat{y}(x)}(x)$. Then, we approximate its local Lipschitz constant by the gradient norm $L^{\text{local}}(x) \approx \|\nabla_x g(x)\|_2$, computed via automatic differentiation by treating $g(x)$ as a scalar objective[1]. Under this approximation, any perturbation δ with $\|\delta\|_2 \leq R$ satisfies

$$|g(x+\delta) - g(x)| \leq L^{\text{local}}(x)\, R, \tag{1}$$

which provides a pointwise, first-order certificate on how much confidence can change within the perturbation ball.

Bin-wise Lipschitz constants. We adopt the standard binning strategy for calibration, partitioning the confidence interval $[0,1]$ into K bins $B_k = [a_k, b_k]$. To obtain certified guarantees, we must ensure that samples assigned to a bin remain in that bin under all admissible perturbations. For each bin B_k, we define a *robust sub-bin* $\widetilde{B}_k = [a_k+\gamma_k,\ b_k-\gamma_k]$, where $\gamma_k > 0$ is chosen so that any sample whose confidence lies in $\widetilde{B}_k$ cannot cross a bin boundary under perturbations of size R. Using the Lipschitz bound in (1), a sufficient condition is:

$$\gamma_k \geq L_k R \geq \max_{n \in \widetilde{\mathcal{I}}_k} L^{\text{local}}(x^{(n)})\, R, \tag{2}$$

where $\widetilde{\mathcal{I}}_k = \{n : p_{\hat{y}(x^{(n)})}(x^{(n)}) \in \widetilde{B}_k\}$, and the bin-wise Lipschitz constant is $L_k = \max_{n \in \widetilde{\mathcal{I}}_k} L^{\text{local}}(x^{(n)})$, which controls the worst-case change in confidence for all samples that remain robustly inside bin k.

Certified bin-wise calibration under perturbations. For samples in $\widetilde{\mathcal{I}}_k$, let $A_k(\delta)$ and $P_k(\delta)$ denote the accuracy and mean confidence under perturbation:

$$A_k(\delta) = \frac{1}{|\widetilde{\mathcal{I}}_k|} \sum_{n \in \widetilde{\mathcal{I}}_k} \mathbb{1}\{y^{(n)} = \hat{y}(x^{(n)} + \delta)\},$$

$$P_k(\delta) = \frac{1}{|\widetilde{\mathcal{I}}_k|} \sum_{n \in \widetilde{\mathcal{I}}_k} p_{\hat{y}(x^{(n)}+\delta)}(x^{(n)} + \delta).$$

We can now decompose the bin-wise miscalibration by adding and subtracting the empirical $\hat{A}_k$ and $\hat{P}_k$ and applying the triangle inequality twice:

$$|A_k(\delta) - P_k(\delta)| \leq |A_k(\delta) - \hat{A}_k| + |\hat{A}_k - \hat{P}_k| + |\hat{P}_k - P_k(\delta)|. \tag{3}$$

This decomposition isolates the three sources of uncertainty: statistical, empirical, and noise-based, and allows us to bound each one separately.

[1] This provides a tractable, first-order approximation of the local Lipschitz constant needed for the certified calibration bound.

- **(i) Statistical deviation:** $|A_k(\delta) - \hat{A}_k|$ measures how much the *true* accuracy under perturbations may differ from the empirical accuracy computed on the clean dataset. Since $A_k(\delta)$ depends on the model predictions for all δ in the perturbation ball, it cannot be computed directly and must be bounded using concentration inequalities, yielding a statistical term ϵ_k^{stat}. In particular, we resort to Hoeffding's inequality [16] to compute ϵ_k^{stat}, which results in $\epsilon_k^{\text{stat}} = \sqrt{\frac{1}{2m_k} \ln(\frac{2}{\alpha})}$, where m_k is the number of samples per bin, i.e., $|\tilde{\mathcal{I}}_k|$, and α the probability of making an error, set to 0.05.
- **(ii) Empirical miscalibration:** $|\hat{A}_k - \hat{P}_k|$ is exactly the empirical bin-wise calibration error computed on clean calibration data. This is the standard calibration quantity used in ECE and requires no additional approximation.
- **(iii) Lipschitz deviation:** $|\hat{P}_k - P_k(\delta)|$ captures how much the mean confidence in the bin may change under perturbations. This term is controlled by the Lipschitz constant of the confidence function, leading to the bound $|\hat{P}_k - P_k(\delta)| \leq L_k R$ (Eq. 1 and Eq. 2).

Proposed upper bound to certify confidence calibration. Combining these terms yields the proposed certified bin-wise confidence calibration guarantee. Specifically, for every admissible perturbation δ with $\|\delta\|_2 \leq R$, we obtain

$$|A_k(\delta) - P_k(\delta)| \leq \widehat{\text{Cal}}_k + \epsilon_k^{\text{stat}} + L_k R, \qquad \forall \|\delta\|_2 \leq R. \tag{4}$$

This bound certifies that, for all samples whose confidence remains within the robust sub-bin $\widetilde{B}_k$, the worst-case bin-wise miscalibration cannot exceed the sum of the empirical calibration error, a statistical uncertainty term, and a Lipschitz-controlled perturbation term.

4 Experiments

4.1 Experiment setup

Models. We use Vision Transformers (ViTs) [9] as our main backbone, given their strong performance and enhanced calibration. Their structured size variants (Ti, S, B) also let us study how model capacity affects certified calibration.
Datasets. We use MedMNIST [35], a standardized medical image classification benchmark to evaluate CalCErt. In particular, we use the 2-dimensional datasets and exclude ChestMNIST because it is a multi-label task, leaving us with 11 total datasets. We use the provided training, validation, and test splits, where we use the validation split for certification and report the results on the test set. We note that [3] reports that the train and test, which are used to build OrganA (OA), OrganC (OC) and OrganS (OS) contain some differences. Thus, we split the datasets in *ID*: Blood (Bl), Breast (Br), Derma (D), OCT (O), Pathology (P), Pneumonia (Pn), Retina (R) and Tissue (T), and *OOD*: OA, OC and OS.
Baselines. We compare CalCErt against two baselines: L_2–*Cal*, where the model is calibrated on ℓ_2-perturbed data, and the bin-wise calibration error computed is used as the upper bound, regardless of the attack used. Similarly, we

Table 1. Coverage and tightness. Results across three ViT sizes and five attack types. $\mathcal{FT}$ indicates if the methods requires specific noise fine-tuning.

				Coverage					Tightness				
		Method	$\mathcal{FT}$	L2	FGSM	BIM	RFGSM	PGD	L2	FGSM	BIM	RFGSM	PGD
ID	ViT-Ti	L_2–CAL	✔	0.36	0.34	0.36	0.34	0.34	0.11	0.30	0.18	0.32	0.32
		ASC	✔	0.36	0.41	0.41	0.44	0.44	0.11	0.21	0.17	0.22	0.22
		`CalCErt`	✘	0.99	0.91	0.98	0.91	0.91	0.33	0.34	0.45	0.32	0.32
	ViT-S	L_2–CAL	✔	0.38	0.36	0.38	0.36	0.36	0.12	0.26	0.19	0.27	0.27
		ASC	✔	0.38	0.44	0.38	0.46	0.48	0.12	0.16	0.17	0.18	0.18
		`CalCErt`	✘	0.98	0.90	0.98	0.90	0.90	0.39	0.44	0.50	0.43	0.43
	ViT-B	L_2–CAL	✔	0.36	0.36	0.38	0.36	0.36	0.12	0.34	0.19	0.36	0.36
		ASC	✔	0.36	0.45	0.39	0.45	0.45	0.12	0.23	0.16	0.23	0.23
		`CalCErt`	✘	0.98	0.91	0.98	0.91	0.91	0.38	0.42	0.53	0.40	0.40
OOD	ViT-Ti	L_2–CAL	✔	0.20	0.20	0.20	0.20	0.20	0.18	0.43	0.24	0.45	0.45
		ASC	✔	0.20	0.37	0.20	0.37	0.37	0.18	0.17	0.23	0.20	0.20
		`CalCErt`	✘	0.60	0.33	0.60	0.27	0.27	0.21	0.26	0.24	0.29	0.29
	ViT-S	L_2–CAL	✔	0.20	0.20	0.20	0.20	0.20	0.15	0.44	0.23	0.47	0.47
		ASC	✔	0.20	0.40	0.20	0.37	0.37	0.15	0.21	0.23	0.23	0.23
		`CalCErt`	✘	0.93	0.40	0.80	0.40	0.40	0.30	0.23	0.28	0.24	0.24
	ViT-B	L_2–CAL	✔	0.20	0.20	0.20	0.20	0.20	0.15	0.41	0.19	0.44	0.44
		ASC	✔	0.20	0.43	0.20	0.43	0.43	0.15	0.20	0.17	0.24	0.21
		`CalCErt`	✘	0.77	0.73	0.77	0.73	0.73	0.98	0.94	1.06	0.94	0.94

define "attack-specific calibration" (ASC), where the empirical calibration upper bound is computed with data perturbed with the specific adversarial attack used.

Adversarial attacks. To evaluate the robustness of `CalCErt` we resort to well-known adversarial attacks: FGSM [14], BIM [20], PGD [24] and RFGSM [33]. In addition, we report results with bounded Gaussian noise, denoted as L2. For all perturbations, we use an ℓ_2-certified radius $R = 0.1$.

Implementation details. Since several bins sometimes contain too few samples to yield stable estimates, we approximate bin-specific ϵ_k by a global ϵ.

Evaluation metrics. Coverage $= \frac{1}{K}\sum_k \mathbb{1}\{\text{Cal}_{k,\text{test}} \leq \text{UB}_k\}$ represents the fraction of bins for which the bound is respected, where UB_k is the upper bound for bin k computed for each method. We define by Tightness $= \frac{1}{K}\sum_k |\text{Cal}_{k,\text{test}} - \text{UB}_k|$, the mean distance between the bound and the test calibration across bins.

4.2 Results

Main results **(Table 1).** *ID scenario:* Across all architectures and adversarial attacks, our method yields the highest coverage, between **0.90-0.99**, representing a **2-3× increase** over the baselines, which hover around 0-30-0.44. Importantly, this improvement does not typically come at the cost of overly conservative or loose bounds. Furthermore, while ASC sometimes yields tighter estimates, it fails to provide useful coverage, and requires retraining for each noise level and attack, making it impractical. In contrast, `CalCErt` **certifies substantially more samples, while maintaining competitive tightness, all without requiring costly per-noise or per-attack retraining.** *OOD scenario:* As expected, performance drops under distribution shift. Nonetheless, `CalCErt` still provides the strongest guarantees, maintaining valid certifications in most cases,

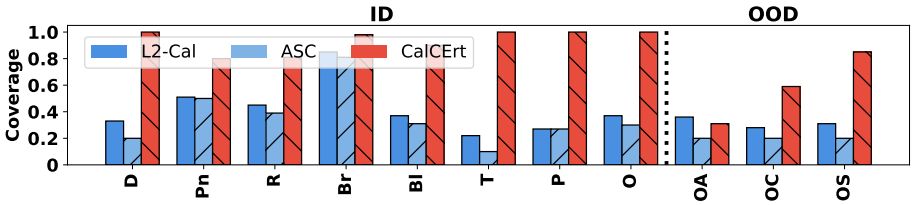

Fig. 1. Dataset-wise coverage. Coverage for L_2–Cal, ASC, and CalCErt across 11 MedMNIST datasets. Results are averaged across three ViT sizes.

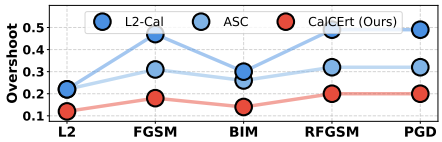

Fig. 2. Overshoot (↓) across types of attacks: mean across ViT backbones.

e.g., **mean coverage often $\geq$ 0.6, with a highest value of 0.93,** whereas the highest coverage of baselines is 0.43. Interestingly, CalCErt bounds tend to appear tighter (or at least comparable to the baselines) in this setting, largely due to the mismatch between the validation data (used to estimate empirical miscalibration) and the OOD test distribution. While this effect should not be interpreted as a genuine advantage, since it stems from a violation of the i.i.d. assumption underlying our certificate, CalCErt still achieves the overall highest certified coverage among the evaluated approaches. Finally, *coverage* and *tightness* must be interpreted jointly: a certificate that applies to only a small fraction of samples can appear tight, but is of limited practical value, whereas CalCErt provides meaningful, data-dependent bounds for most of the inputs under a wide spectrum of adversarial perturbations.

***Per-dataset results* (Figure 1).** We observe that CalCErt consistently yields strong coverage gains across all ID datasets, **achieving perfect coverage in five out of 8 datasets**, and clearly outperforming both baselines. Under domain shift, improvements become more modest and occasionally only competitive, as in the OA dataset, reflecting the expected degradation when validation and test distributions diverge. Still, even in the OD regime, CalCErt remains the only method that preserves meaningful coverage rather than collapsing entirely.

***What happens when the bounds are violated?* (Figure 2).** We define *overshoot* as the mean positive gap between the true bin-wise miscalibration and the certified bound, evaluated over all bins where the bound is violated. In addition to its superior coverage, **CalCErt exhibits much smaller violation magnitudes whenever the bound is not satisfied.**

***Robustness across models* (Table 2).** Consistent with our ViT results, CalCErt achieves **near-full coverage** in the ID setting even when using a

Table 2. Results with ResNet18 as backbone.

			Coverage					Tightness				
	Method	$\mathcal{FT}$	L2	FGSM	BIM	RFGSM	PGD	L2	FGSM	BIM	RFGSM	PGD
ID	L_2–Cal	✔	0.35	0.34	0.33	0.34	0.34	0.12	0.32	0.25	0.36	0.36
	ASC	✔	0.35	0.41	0.39	0.43	0.43	0.12	0.22	0.24	0.27	0.27
	`CalCErt`	✘	0.99	0.95	0.95	0.95	0.95	0.44	0.53	0.59	0.49	0.49
OOD	L_2–Cal	✔	0.20	0.20	0.20	0.20	0.20	0.15	0.40	0.19	0.43	0.43
	ASC	✔	0.20	0.27	0.23	0.27	0.27	0.15	0.23	0.19	0.25	0.26
	`CalCErt`	✘	0.57	0.27	0.50	0.23	0.23	0.17	0.30	0.20	0.32	0.32

ResNet-18 backbone. Under distributional shift, however, ResNet-18 exhibits a more pronounced degradation. This is expected, as compared to ViTs, ResNets generally have lower representational capacity and weaker calibration, making their confidence estimates more sensitive to OOD drift. Moreover, because our certificate depends on bin-wise confidence stability, architectures whose confidence distributions shift more under OOD conditions naturally yield fewer samples that remain within robust sub-bins. Despite this drop, `CalCErt` still substantially outperforms the baselines in ID and remains competitive in OOD.

5 Discussion and limitations

This work should be viewed as an exploratory first step toward certifying confidence calibration under adversarial perturbations. Along the way, we encountered several conceptual and empirical limitations that highlight important challenges and opportunities for future work. **First**, our bin-wise certificates depend on the empirical distribution of samples within each confidence bin. In several datasets, some bins contain very few samples, leading to loose statistical bounds and, consequently, large certified miscalibration values. Adaptive binning, where bin counts are equalized, could mitigate this imbalance, but we observed that it often produces *near-overlapping* bins, which may complicate the construction of robust sub-bins, and violate the assumptions required for our certificates. **Second,** `CalCErt` relies on local Lipschitz estimates of the confidence function. While this provides a principled way to control confidence variation under perturbations, it inherits well-known limitations of Lipschitz-based robustness methods. In particular, local Lipschitz constants can be large for some architectures, which increases the perturbation term in the certified bound and may render the certificate overly conservative or even vacuous. This reflects a broader challenge in Lipschitz-based certification: global bounds (e.g., products of spectral norms) are computationally efficient but notoriously loose [22,34], whereas tighter local [15] or semidefinite programming (SDP)-based [12] bounds are more accurate but scale poorly and are impractical for modern architectures. Our approach sits within this landscape and therefore shares these structural trade-offs. Furthermore, techniques such as spectral normalization, Jacobian penalties, or explicitly training low-Lipschitz networks [1,26] could reduce these constants and yield tighter certificates, but integrating such training-time strategies is

incompatible with our post-hoc, model-agnostic setting. **And third**, our certificates address only the robustness of confidence calibration and do not guarantee classification accuracy under adversarial perturbations. A model may remain well-calibrated while still misclassifying perturbed inputs, and conversely, accuracy robustness does not imply calibration robustness. Combining both within a unified framework is non-trivial, as the two objectives interact in complex ways and may require fundamentally different bounding strategies. Developing joint guarantees is therefore an important direction, but lies beyond the scope of this study. These limitations do not alter the main message of this work, where our goal is not to provide a definitive solution, but to explore how certification can be meaningfully extended to confidence calibration in safety-critical settings.

Acknowledgements. This work has benefited from state financial aid, managed by the Agence Nationale de Recherche under the investment program integrated into France 2030, project reference ANR-21-RHUS-0003. This work was granted access to the HPC resources of IDRIS under the allocation 2025-AD011014802R1 made by GENCI. The author gratefully acknowledges support from ILLS during the internship at ETS Montreal.

Disclosure of Interests. The authors declare no competing interests.

References

1. Anil, C., Lucas, J., Grosse, R.: Sorting out lipschitz function approximation. In: International conference on machine learning. pp. 291–301 (2019)
2. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: International conference on machine learning. pp. 274–283. PMLR (2018)
3. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). Medical image analysis **84**, 102680 (2023)
4. Bortsova, G., González-Gonzalo, C., Wetstein, S.C., Dubost, F., Katramados, I., Hogeweg, L., Liefers, B., Van Ginneken, B., Pluim, J.P., Veta, M., et al.: Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. Medical image analysis **73**, 102141 (2021)
5. Carlini, N., Tramer, F., Dvijotham, K.D., Rice, L., Sun, M., Kolter, J.Z.: (certified!!) adversarial robustness for free! In: The Eleventh International Conference on Learning Representations (2023)
6. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. Ieee (2017)
7. Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., Usunier, N.: Parseval networks: Improving robustness to adversarial examples. In: International conference on machine learning. pp. 854–863. PMLR (2017)
8. Cohen, J., Rosenfeld, E., Kolter, Z.: Certified adversarial robustness via randomized smoothing. In: international conference on machine learning. pp. 1310–1320. PMLR (2019)
9. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)

10. Engstrom, L., Tran, B., Tsipras, D., Schmidt, L., Madry, A.: Exploring the landscape of spatial robustness. In: International conference on machine learning. pp. 1802–1811. PMLR (2019)
11. Fazlyab, M., Entesari, T., Roy, A., Chellappa, R.: Certified robustness via dynamic margin maximization and improved lipschitz regularization. Advances in Neural Information Processing Systems **36**, 34451–34464 (2023)
12. Fazlyab, M., Robey, A., Hassani, H., Morari, M., Pappas, G.: Efficient and accurate estimation of lipschitz constants for deep neural networks. Advances in neural information processing systems **32** (2019)
13. Ghiasi, A., Shafahi, A., Goldstein, T.: Breaking certified defenses: Semantic adversarial examples with spoofed robustness certificates. In: International Conference on Learning Representations (2020)
14. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
15. Hashemi, N., Ruths, J., Fazlyab, M.: Certifying incremental quadratic constraints for neural networks via convex optimization. In: Learning for Dynamics and Control. pp. 842–853. PMLR (2021)
16. Hoeffding, W.: Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association **58**(301), 13–30 (1963)
17. Huang, Y., Zhang, H., Shi, Y., Kolter, J.Z., Anandkumar, A.: Training certifiably robust neural networks with efficient local lipschitz bounds. Advances in Neural Information Processing Systems **34**, 22745–22757 (2021)
18. Hussein, N., Shamshad, F., Naseer, M., Nandakumar, K.: Promptsmooth: Certifying robustness of medical vision-language models via prompt learning. In: MICCAI. pp. 698–708 (2024)
19. Katz, G., Barrett, C., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: An efficient SMT solver for verifying deep neural networks. In: International conference on computer aided verification. pp. 97–117. Springer (2017)
20. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: Artificial intelligence safety and security, pp. 99–112. Chapman and Hall/CRC (2018)
21. Laousy, O., Araujo, A., Chassagnon, G., Paragios, N., Revel, M.P., Vakalopoulou, M.: Certification of deep learning models for medical image segmentation. In: MICCAI. pp. 611–621 (2023)
22. Leino, K., Wang, Z., Fredrikson, M.: Globally-robust neural networks. In: International Conference on Machine Learning. pp. 6212–6222. PMLR (2021)
23. Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., Lu, F.: Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern recognition **110**, 107332 (2021)
24. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (ICLR) (2018)
25. Mohapatra, J., et al.: Higher-order certification for randomized smoothing. Advances in Neural Information Processing Systems **33**, 4501–4511 (2020)
26. Pauli, P., Koch, A., Berberich, J., Kohler, P., Allgöwer, F.: Training robust neural networks using lipschitz bounds. IEEE Control Systems Letters **6**, 121–126 (2021)
27. Rosenfeld, E., Winston, E., Ravikumar, P., Kolter, Z.: Certified robustness to label-flipping attacks via randomized smoothing. In: International Conference on Machine Learning. pp. 8230–8241. PMLR (2020)

28. Salman, H., Sun, M., Yang, G., Kapoor, A., Kolter, J.Z.: Denoised smoothing: A provable defense for pretrained classifiers. Advances in Neural Information Processing Systems **33**, 21945–21957 (2020)
29. Shafahi, A., Huang, W.R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., Goldstein, T.: Poison frogs! targeted clean-label poisoning attacks on neural networks. Advances in neural information processing systems **31** (2018)
30. Singh, G., Gehr, T., Püschel, M., Vechev, M.: Boosting robustness certification of neural networks. In: International conference on learning representations (2019)
31. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: International Conference on Learning Representations (ICLR) (2014)
32. Tramer, F., et al.: On adaptive attacks to adversarial example defenses. Advances in neural information processing systems **33**, 1633–1645 (2020)
33. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: Ensemble adversarial training: Attacks and defenses (2020)
34. Tsuzuku, Y., Sato, I., Sugiyama, M.: Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. Advances in neural information processing systems **31** (2018)
35. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. Scientific Data **10**(1), 41 (2023)
36. Zhang, B., Jiang, D., He, D., Wang, L.: Rethinking lipschitz neural networks and certified robustness: A boolean function perspective. Advances in neural information processing systems **35**, 19398–19413 (2022)