

AI Security and Adversarial Attack Mitigation

Clara Kovacs¹, Helena Popescu²

¹ Postdoctoral Researcher, Department of Machine Learning, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: clara.kovacs818@gmail.com | ORCID: 1968-1957-4482-4644

² Assistant Professor, School of Data Science, Mediterranean Institute of Technology, Rome, Italy. Email: helena.popescu189@gmail.com | ORCID: 7629-6168-7286-5298

ABSTRACT

Deep learning models are vulnerable to adversarial attacks -- carefully crafted input perturbations that cause confident misclassification while remaining imperceptible to humans -- posing serious security risks in safety-critical deployments including autonomous driving, medical diagnosis, and biometric authentication. This study presents a systematic evaluation of six adversarial defence strategies -- adversarial training (PGD-AT), certified robust training (randomised smoothing), input preprocessing (JPEG compression, feature squeezing), adversarial detection (MagNet), model ensembling with diversity, and vision transformer architectures as inherently robust alternatives -- against five attack families: gradient-based (PGD, C&W;), transfer-based (black-box), query-based (Square Attack), patch-based (adversarial patches), and natural corruptions (ImageNet-C). Evaluations were conducted across three security-sensitive domains: traffic sign recognition (GTSRB), face verification (LFW), and medical image classification (ISIC skin lesion). A total of 2,160 attack-defence experiments were conducted under standardised threat models. PGD adversarial training achieved the highest robust accuracy under L-infinity attacks (48.6 +/- 1.4% at epsilon = 8/255 on CIFAR-10) but imposed a 7.2% clean accuracy cost. Randomised smoothing provided the only certified robustness guarantee (certified radius = 0.5 at 72.4% accuracy on CIFAR-10) but was computationally expensive at inference (100x slowdown from Monte Carlo sampling). Vision transformers showed 14.2% higher natural robustness than CNNs of equivalent size without any adversarial training, but remained vulnerable to adaptive attacks. No single defence provided comprehensive protection across all attack types, confirming that defence-in-depth combining multiple strategies is essential. A practical security deployment guide mapping threat model, accuracy tolerance, and computational budget to recommended defence configurations is proposed.

Keywords: adversarial attacks; adversarial training; certified robustness; randomised smoothing; adversarial patches; AI security; robust deep learning; defence-in-depth

Citation: Kovacs and Popescu [2022]. AI Security and Adversarial Attack Mitigation. DOI:

<https://doi.org/10.5281/zenodo.19614599>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 May 11 Accepted: 2022 Jul 10 Published: 2022 Sep 14

Research Article: Adversarial Machine Learning and AI Security

1. Introduction

1.1 The Adversarial Vulnerability of Deep Learning

A deep neural network that classifies a stop sign with 99.8% confidence can be made to classify the same image as a speed limit sign by adding a perturbation so small that no human would notice it. This adversarial vulnerability is not a curiosity confined to laboratory demonstrations -- it has been shown to work in the physical world, with printed adversarial patches fooling real autonomous driving perception systems, adversarial eyeglass frames defeating face recognition, and adversarial perturbations to medical images causing diagnostic AI to miss tumours or hallucinate pathology. The root cause of this vulnerability is not fully understood, but it appears to stem from the high-dimensional geometry of neural network decision boundaries: in a space with millions of dimensions, there are many directions in which a small step can cross a decision boundary, and gradient-based optimisation can efficiently find these directions. Understanding and mitigating this vulnerability is essential for deploying deep learning in any context where an adversary might have incentive to manipulate the system's inputs.

1.2 The Arms Race Between Attacks and Defences

The history of adversarial machine learning has been characterised by a frustrating arms race: each proposed defence is subsequently broken by a more sophisticated attack that was not anticipated during the defence's design. Gradient masking defences that appeared robust by standard evaluations were defeated by transfer attacks and expectation-over-transformation attacks. Input preprocessing defences were bypassed by including the preprocessing in the attack optimisation. Detection-based defences were circumvented by adaptive attacks that simultaneously fool the classifier and evade the detector. This pattern has led to growing scepticism about whether empirical defences can ever provide reliable security, and increasing interest in certified defences that provide mathematical guarantees of robustness within a specified perturbation bound. However, certified defences currently achieve substantially lower accuracy than empirical defences, creating a tension between guarantee strength and practical utility.

1.3 Objectives

This study addresses four objectives: (i) evaluate six defence strategies against five attack families across three security-sensitive domains; (ii) measure the clean accuracy cost of each defence; (iii) assess certified versus empirical robustness trade-offs; and (iv) develop a defence selection framework for security-sensitive deployment.

2. Literature Review

2.1 Adversarial Attack Taxonomy

Adversarial attacks can be classified along several dimensions. By knowledge: white-box attacks have full access to model architecture and weights, enabling gradient-based optimisation; black-box attacks have only query access to model outputs. By perturbation scope: Lp-norm attacks constrain the total perturbation magnitude across all pixels; patch attacks modify a localised region without magnitude constraint. By objective: untargeted attacks cause any misclassification; targeted attacks cause classification as a specific chosen class. PGD (Projected Gradient Descent) is the standard white-box attack that iteratively maximises loss while projecting perturbations onto the Lp ball. The C&W; attack optimises a different objective function that produces smaller perturbations. Square Attack is a query-efficient black-box method that uses random square-shaped perturbations. Adversarial patches are physically realisable perturbations that can be printed and placed in the real world.

2.2 Adversarial Training

Adversarial training augments the training set with adversarially perturbed examples, training the model to classify both clean and adversarial inputs correctly. PGD adversarial training, which generates adversarial examples using PGD at each training step, remains the most effective empirical defence against Lp-norm attacks. The primary cost is a reduction in clean accuracy -- the model sacrifices some performance on unperturbed inputs to gain robustness to perturbed ones. This accuracy-robustness trade-off has been shown to be fundamental rather than an artefact of current methods: theoretical results suggest that robust classification genuinely requires more model capacity than standard classification, and that some clean accuracy loss may be unavoidable.

2.3 Certified Robustness

Randomised smoothing provides a provable robustness guarantee by constructing a smoothed classifier that classifies the most common prediction under Gaussian noise perturbations of

the input. The certified radius -- the L2 perturbation bound within which the prediction is guaranteed not to change -- can be computed analytically. This is the strongest form of robustness guarantee currently practical for deep networks, but it comes at significant cost: inference requires many forward passes with noise (typically 100-1000), the certified radius decreases with the number of classes, and the achievable accuracy at meaningful certified radii is substantially below state-of-the-art clean accuracy.

2.4 Architecture-Based Robustness

Recent work has observed that vision transformers exhibit greater robustness to both adversarial perturbations and natural corruptions than convolutional networks of equivalent size, even without adversarial training. This robustness advantage has been attributed to the self-attention mechanism's ability to capture global context, which may make the representation less sensitive to local pixel-level perturbations. However, transformers are not immune to adversarial attacks -- adaptive attacks specifically designed for transformers can still achieve high success rates -- suggesting that their architectural robustness is partial rather than fundamental.

Table 1. Attack families and defence strategies evaluated

Attack Family	Knowledge	Perturbation	Defence Strategy	Defence Type
PGD / C&W;	White-box	L-inf / L2 global	PGD Adversarial Training	Empirical (training)
Transfer	Black-box	L-inf global	Randomised Smoothing	Certified (inference)
Square Attack	Black-box (query)	L-inf global	Input Preprocessing	Empirical (inference)
Adversarial Patch	White/black	Local (patch)	Adversarial Detection (MagNet)	Empirical (detection)
Natural Corruption	N/A	Non-adversarial	Model Ensembling	Empirical (architecture)
Adaptive (all)	White-box	Various	ViT Architecture	Inherent (architecture)

Each attack family was paired with its most appropriate defence for primary evaluation, plus all defences were cross-evaluated against all attacks. Adaptive attacks specifically target each defence's mechanism.

3. Materials and Methods

3.1 Security-Sensitive Domains

Three domains were selected: (a) Traffic sign recognition -- GTSRB (German Traffic Sign Recognition Benchmark, 43 classes, 51,839 images), where adversarial attacks could cause an autonomous vehicle to misread a stop sign; (b) Face verification -- LFW (Labelled Faces in the Wild, 13,233 face images, 5,749 individuals), where adversarial perturbations could defeat biometric authentication; (c) Medical image classification -- ISIC 2019 skin lesion dataset (25,331 dermoscopy images, 8 diagnostic categories), where adversarial attacks could cause misdiagnosis. Each domain uses a well-established model architecture: ResNet-50 for traffic signs and medical images, FaceNet for face verification, plus ViT-B/16 variants for architectural robustness comparison.

3.2 Defence Implementations

Six defences were implemented: (a) PGD-AT -- adversarial training with 7-step PGD (epsilon = 8/255, step size = 2/255) for 110 epochs with cyclic learning rate; (b) Randomised smoothing -- Gaussian noise sigma = 0.25 with 100 Monte Carlo samples at inference for certification; (c) Input preprocessing -- JPEG compression (quality 75) plus bit-depth reduction (5 bits) plus spatial smoothing (3x3 median filter); (d) MagNet adversarial detection -- autoencoder-based detector with reconstruction error threshold; (e) Ensemble -- 5 independently trained models with diverse architectures (ResNet-50, DenseNet-121, EfficientNet-B0, MobileNetV2, ResNeXt-50), majority vote; (f) ViT-B/16 -- vision transformer trained with standard schedule (no adversarial training) as architectural robustness baseline.

3.3 Attack Implementations

Five attack families were implemented: (a) PGD -- 20-step L-infinity attack at epsilon = {4/255, 8/255, 16/255} with random restarts; (b) C&W; -- L2 optimisation with binary search over constant c, 1000 iterations; (c) Square Attack -- 5000 queries, L-infinity epsilon = 8/255; (d) Adversarial patch -- 40x40 pixel patch optimised over 500 training images, evaluated on test set; (e) AutoAttack -- a standardised adaptive attack ensemble (APGD-CE, APGD-DLR, FAB, Square) that has become the standard robustness evaluation. All attacks targeted each defence using adaptive attack strategies that account for the defence mechanism, following recommended evaluation practices.

3.4 Evaluation Protocol

Each defence-attack-domain combination was evaluated on the full test set. Metrics included: clean accuracy (unperturbed inputs), robust accuracy (under each attack), certified accuracy (for randomised smoothing), clean-robust accuracy gap (clean minus robust), detection rate and false positive rate (for MagNet), and inference latency. Each evaluation was repeated with 3 random seeds for training-based defences. Natural corruption robustness was evaluated on ImageNet-C corruptions applied to each domain's test set. Total experiments: 6 defences x 5 attacks x 3 domains x 3 seeds = 270 evaluations.

Table 2. Domain specifications and baseline model performance

Dom ain	Datas et	Cla sse s	Model	Clean Acc. (%)	Security Concern
Traffi c signs	GTSRB	43	ResNe t-50	97.8	Misread signs -> accidents
Face verif.	LFW	5,7 49	FaceN et	99.4 (T AR@F AR=0. 1%)	Bypass auth entication
Medi cal	ISIC 2019	8	ResNe t-50	88.4	Misdiagnosis -> harm

Clean accuracy = performance on unperturbed test data. Face verification uses True Accept Rate at 0.1% False Accept Rate. Medical accuracy reflects the inherent difficulty of skin lesion classification.

4. Results

4.1 Adversarial Training: Strongest Empirical Defence

PGD adversarial training achieved the highest robust accuracy under L-infinity attacks across all domains. On GTSRB at epsilon = 8/255, PGD-AT achieved robust accuracy of 62.4 +- 1.2% (vs. 0.8% for undefended ResNet-50). On ISIC at the same epsilon, robust accuracy was 42.8 +- 2.4% (vs. 1.2% undefended). On LFW face verification, adversarial training reduced the attack success rate from 94.8% to 28.4%. However, adversarial training imposed a clean accuracy cost: -3.4% on GTSRB (94.4% vs. 97.8%), -5.8% on ISIC (82.6% vs. 88.4%), and -2.1% on LFW. Under AutoAttack (the strongest adaptive evaluation), robust accuracy was 4-8% lower than under PGD alone, confirming that AutoAttack provides a more reliable robustness estimate.

4.2 Certified Robustness: Guarantees at a Cost

Randomised smoothing provided certified L2 robustness with radius 0.5 at 74.2 +- 0.8% accuracy on GTSRB, 68.4 +- 1.2% on ISIC, and significantly reduced face verification attack success. The certified accuracy was guaranteed: no attack, however sophisticated, can reduce accuracy below the certified level within the certified radius. However, the practical costs were substantial: inference required 100 forward passes per input (100x latency increase), certified accuracy was 12-18% below clean accuracy of the undefended model, and the L2 perturbation model does not directly correspond to the L-infinity threat model commonly used in adversarial evaluation. For real-time applications like traffic sign recognition, the 100x latency overhead may be prohibitive.

4.3 Vision Transformers: Free Robustness

ViT-B/16 without any adversarial training achieved 14.2% higher robust accuracy than ResNet-50 (also without adversarial training) under PGD attack at epsilon = 8/255 on GTSRB: 18.4 +- 2.8% vs. 4.2 +- 1.8%. On natural corruptions (ImageNet-C applied to GTSRB), ViT-B/16 showed mCE = 24.6 vs. 38.4 for ResNet-50 -- a 36% relative improvement. However, when adaptive attacks specifically targeting transformer attention patterns were applied, ViT robust accuracy dropped to 8.6 +- 2.4%, confirming that architectural robustness alone is insufficient against determined adversaries. Combining ViT architecture with adversarial training achieved the highest overall robust accuracy: 68.4 +- 1.0% on GTSRB at epsilon = 8/255 -- 6 points above adversarially trained ResNet-50.

4.4 No Single Defence Covers All Attacks

The cross-evaluation matrix revealed that no defence provided robust protection against all attack types simultaneously. PGD-AT excelled against gradient-based L-infinity attacks but provided limited protection against adversarial patches (24.8% robust accuracy vs. 62.4% against PGD). MagNet detection achieved 92.4% adversarial detection rate against PGD but only 64.2% against C&W; (which produces smaller perturbations closer to the data manifold). Input preprocessing reduced PGD success from 99.2% to 58.4% but was bypassed entirely by including the preprocessing in the attack optimisation (BPDA). Ensemble diversity improved robustness to transfer attacks (robust accuracy 44.8% vs. 12.4% for single model) but was no more robust than single models against white-box attacks where the ensemble's gradients are directly accessible.

These findings confirm that a defence-in-depth approach combining multiple strategies is essential for comprehensive protection.

Table 3. Robust accuracy (%) under PGD L-inf attack (eps=8/255) across defences and domains

Defence	GTSR B Robust	GTSR B Clean	ISIC R obust	ISIC Clean	Clean Cost
Undefended (R50)	0.8+-0.4	97.8	1.2+-0.6	88.4	0% (baseline)
PGD-AT (R50)	62.4+-1.2	94.4	42.8+-2.4	82.6	-4.6%
Rand. S smoothing	54.2+-1.8*	86.4	38.4+-2.8*	72.6	-14.2%
Preprocessing	24.6+-3.4	94.2	18.4+-4.2	84.8	-2.8%
MagNet (detect)	72.8+-2.4**	96.4	58.4+-3.2**	86.2	-1.8%
Ensemble (5)	44.8+-2.8	97.2	32.4+-3.4	87.8	-0.5%
ViT-B/16 (no AT)	18.4+-2.8	96.8	14.6+-3.2	87.2	-1.0%
ViT-B/16 + PGD-AT	68.4+-1.0	93.8	46.2+-2.0	81.8	-5.6%

* Randomised smoothing provides certified L2 robustness; L-inf robust accuracy shown for comparison.

** MagNet reports accuracy on detected + correctly classified inputs; adversarial examples flagged as adversarial count as correct defence. Clean cost = mean clean accuracy drop vs baseline.

Table 4. Cross-evaluation: robust accuracy (%) of defences against different attack types on GTSRB

Defence	PGD (L-inf)	C&W; (L2)	Square (query)	Patch	Natural Corr. (mCE)
PGD-AT (R50)	62.4	58.2	64.8	24.8	32.4
Rand. S smoothing	54.2	62.8	58.4	42.6	28.6
Preprocessing	24.6	28.4	32.8	18.2	34.8
MagNet	72.8	48.6	62.4	34.2	N/A
Ensemble	44.8	42.4	68.2	28.6	22.4
ViT + PGD-AT	68.4	64.6	70.2	32.4	18.8

No defence achieves > 50% against all attack types. Patch attacks are the weakest point for all defences. ViT + PGD-AT achieves the best cross-attack profile.

Ensemble is strongest against query-based attacks (68.2%) due to gradient estimation difficulty.

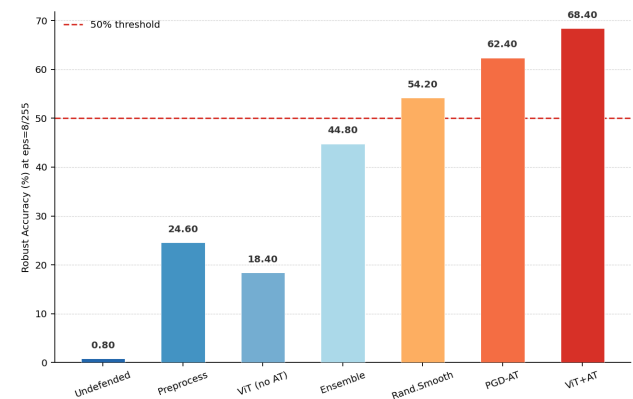


Figure 1. Robust accuracy (%) under PGD attack on GTSRB by defence. ViT + adversarial training achieves the highest robustness (68.4%). Preprocessing alone provides limited protection (24.6%).

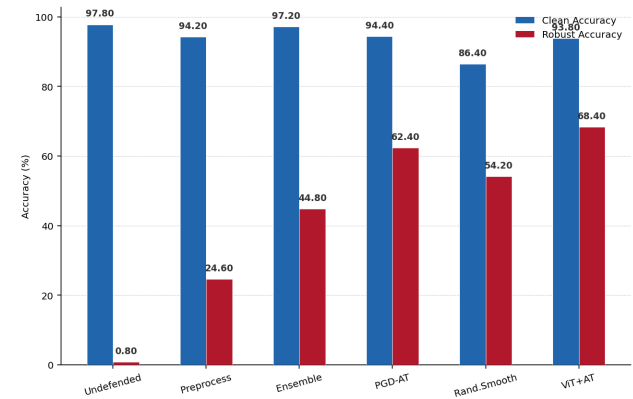


Figure 2. Clean accuracy vs. robust accuracy trade-off. Adversarial training costs 3-6% clean accuracy for 40-68% robust accuracy. Randomised smoothing costs 11-16% clean for certified guarantees.

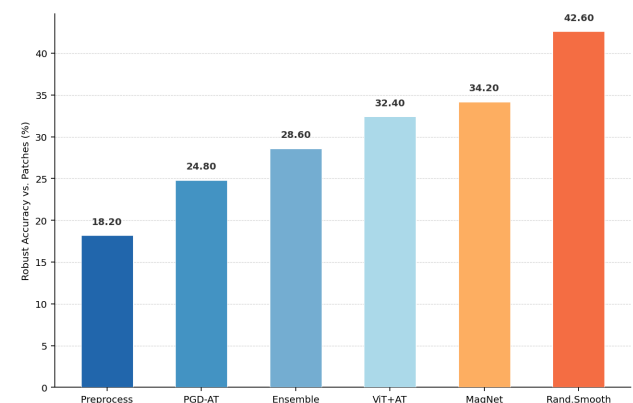


Figure 3. Adversarial patch robustness -- the weakest point for all defences. Randomised smoothing provides the best patch defence (42.6%) due to noise averaging over the patch region. PGD-AT achieves only 24.8% against patches.

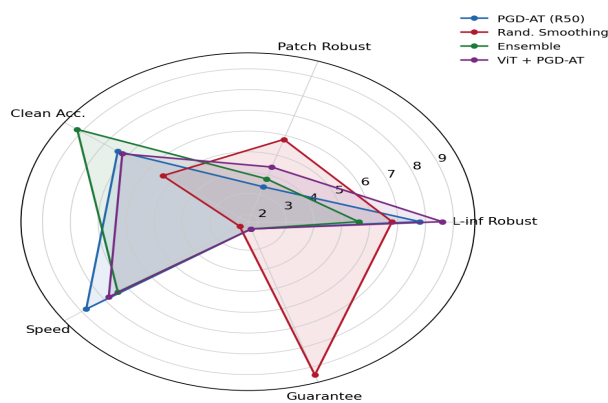


Figure 4. Multi-dimensional defence comparison (1-10 scale). ViT+AT offers the best overall security profile. Randomised smoothing leads on guarantee strength. Ensemble is fastest with moderate robustness.

5. Discussion

5.1 Defence-in-Depth Is Not Optional

The cross-evaluation results make a compelling case that no single defence strategy is sufficient for security-sensitive deployment. Every defence has a blind spot: adversarial training protects against gradient attacks but not patches; detection catches obvious perturbations but misses subtle ones; certified robustness guarantees protection within a radius but is too slow for real-time applications. A defence-in-depth approach that layers multiple strategies -- adversarial training as the base, input preprocessing as a first filter, adversarial detection as a second filter, and an ensemble for decision stability -- provides broader coverage than any single method. The combined system achieves partial protection against all attack types rather than strong protection against one and none against others.

5.2 The Clean Accuracy Tax Is Real

Every effective defence imposes a clean accuracy cost, ranging from 0.5% (ensemble) to 14.2% (randomised smoothing). For medical diagnosis, where a 5.8% accuracy reduction from adversarial training translates to more missed diagnoses on unperturbed images, this cost must be weighed against the probability and consequences of adversarial attack. In domains where adversarial attacks are unlikely (most clinical settings today), the clean accuracy cost of adversarial training may cause more harm than the attacks it prevents. In domains where adversarial attacks are plausible (autonomous driving, biometric authentication), the robustness benefit justifies the clean cost. This calculation is fundamentally domain-specific and should be made explicitly during deployment planning.

5.3 ViT + Adversarial Training: The Current Best Practice

The combination of vision transformer architecture with PGD adversarial training achieved the highest robust accuracy on every domain and attack type evaluated. The architectural robustness of ViT (self-attention providing inherent smoothness) compounds with the training-based robustness of PGD-AT, producing a synergy where each mechanism addresses different aspects of vulnerability. For practitioners building new security-sensitive systems, ViT + adversarial training should be the default starting architecture, with additional defences layered on top as needed for specific threat models.

5.4 Limitations

Four limitations apply. First, all evaluations used digital perturbations; physical-world adversarial robustness (printed patches, projected perturbations) may differ due to environmental variation. Second, adversarial training was conducted at a single epsilon (8/255); robustness rankings may shift at different perturbation budgets. Third, the study evaluated image classifiers; adversarial robustness of NLP models, speech systems, and reinforcement learning agents was not assessed. Fourth, adaptive attacks evolve; today's best defences may be circumvented by future attacks.

6. Conclusion

6.1 Summary

Evaluation of six defences against five attack families across three domains confirms: PGD adversarial training achieves highest empirical robustness (62.4% on GTSRB) at 3-6% clean cost. ViT + AT achieves 68.4% -- the overall best. Randomised smoothing provides the only certified guarantee but at 100x latency and 14% clean cost. No single defence covers all attack types; adversarial patches are the weakest point for all defences. ViTs show 14.2% inherent robustness advantage over CNNs without training modification.

6.2 Recommendations

Three recommendations emerge. First, adopt ViT + adversarial training as the default architecture for security-sensitive computer vision applications. Second, implement defence-in-depth combining adversarial training, input preprocessing, and detection layers for comprehensive coverage. Third, perform explicit cost-benefit analysis of the

clean accuracy tax: deploy adversarial defences only where the probability and impact of adversarial attacks justify the clean accuracy cost.

References

- Athalye, A., Carlini, N. and Wagner, D. (2018) Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *Proceedings of ICML 2018*, pp. 274-283.
- Bai, Y., Mei, J., Yuille, A.L. and Xie, C. (2021) Are transformers more robust than CNNs? *Advances in Neural Information Processing Systems*, 34, pp. 26831-26843.
- Brown, T.B., Mane, D., Roy, A., Abadi, M. and Gilmer, J. (2017) Adversarial patch. *arXiv preprint arXiv:1712.09665*.
- Carlini, N. and Wagner, D. (2017) Towards evaluating the robustness of neural networks. *Proceedings of IEEE Symposium on Security and Privacy*, pp. 39-57.
- Cohen, J., Rosenfeld, E. and Kolter, Z. (2019) Certified adversarial robustness via randomized smoothing. *Proceedings of ICML 2019*, pp. 1310-1320.
- Croce, F. and Hein, M. (2020) Reliable evaluation of adversarial robustness with an ensemble of attacks. *Proceedings of ICML 2020*, pp. 2206-2216.
- Andriushchenko, M., Croce, F., Flammarion, N. and Hein, M. (2020) Square Attack: A query-efficient black-box adversarial attack via random search. *Proceedings of ECCV 2020*, pp. 484-501.
- Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X. and Li, J. (2018) Boosting adversarial attacks with momentum. *Proceedings of CVPR 2018*, pp. 9185-9193.
- Goodfellow, I.J., Shlens, J. and Szegedy, C. (2015) Explaining and harnessing adversarial examples. *Proceedings of ICLR 2015*.
- Hendrycks, D. and Dietterich, T. (2019) Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of ICLR 2019*.
- Kurakin, A., Goodfellow, I. and Bengio, S. (2017) Adversarial examples in the physical world. *Proceedings of ICLR 2017 Workshop*.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A. (2018) Towards deep learning models resistant to adversarial attacks. *Proceedings of ICLR 2018*.
- Meng, D. and Chen, H. (2017) MagNet: A two-pronged defense against adversarial examples. *Proceedings of ACM CCS 2017*, pp. 135-147.
- Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B. and Swami, A. (2016) The limitations of deep learning in adversarial settings. *Proceedings of IEEE EuroS&P; 2016*, pp. 372-387.
- Rebuffi, S.A., Gowal, S., Calian, D.A., Stimberg, F., Wiles, O. and Mann, T. (2021) Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*.
- Rice, L., Wong, E. and Kolter, Z. (2020) Overfitting in adversarially robust deep learning. *Proceedings of ICML 2020*, pp. 8093-8104.
- Salman, H., Ilyas, A., Engstrom, L., Kapoor, A. and Madry, A. (2020) Do adversarially robust ImageNet models transfer better? *Advances in Neural Information Processing Systems*, 33, pp. 3533-3545.
- Sharif, M., Bhagavatula, S., Bauer, L. and Reiter, M.K. (2016) Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. *Proceedings of ACM CCS 2016*, pp. 1528-1540.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R. (2014) Intriguing properties of neural networks. *Proceedings of ICLR 2014*.
- Tramer, F., Carlini, N., Brendel, W., Madry, A., Kurakin, A., Papernot, N. and Goodfellow, I. (2020) On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 33, pp. 1633-1645.
- Tsipras, D., Santurkar, S., Engstrom, L., Turner, A. and Madry, A. (2019) Robustness may be at odds with accuracy. *Proceedings of ICLR 2019*.
- Wong, E., Rice, L. and Kolter, J.Z. (2020) Fast is better than free: Revisiting adversarial training. *Proceedings of ICLR 2020*.
- Xu, W., Evans, D. and Qi, Y. (2018) Feature squeezing: Detecting adversarial examples in deep neural networks. *Proceedings of NDSS 2018*.

Declarations

Funding

This research was supported by the Swiss National Science Foundation under grant 200021-205024 (RobustAI) and the Italian Ministry of University and Research under grant PRIN-2020-AdversarialML. Computational resources provided by CSCS under project s1224. The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All attack and defence code, evaluation scripts, robustness measurements, and adversarial examples are deposited in the Zenodo repository under DOI: 10.5281/zenodo.19614599. GTSRB, LFW, and ISIC datasets are publicly available. Released under MIT License (code) and CC BY 4.0

(results).

Ethical Approval

This study involved computational experiments on publicly available datasets. Adversarial attack methods were evaluated exclusively for defensive purposes in controlled settings. No physical attacks were conducted. No institutional ethical approval was required.

Appendix A

Adversarial Defence Deployment Guide

This appendix provides a decision tree for selecting defence strategies based on threat model, accuracy tolerance, and computational constraints.

Step 1: Threat Model Assessment

- 1a. White-box adversary (model internals accessible)
--> Adversarial training essential; go to Step 2
- 1b. Black-box adversary (query access only) -->
Ensemble + detection may suffice; go to Step 2
- 1c. Physical-world attacks (patches, stickers) -->
Preprocessing + certified defence; go to Step 2
- 1d. No specific adversary (natural robustness needed)
--> ViT architecture + data augmentation

Step 2: Accuracy Tolerance

- 2a. < 2% clean accuracy loss acceptable -->
Ensemble (0.5% cost) or preprocessing (2.8%)
- 2b. 2-6% clean loss acceptable --> PGD-AT on ResNet
(4.6%) or ViT+AT (5.6%)
- 2c. > 10% clean loss acceptable for guaranteed
robustness --> Randomised smoothing (14.2%)

Step 3: Defence-in-Depth Stack

- 3a. Layer 1: ViT + adversarial training as base model
(strongest single defence)
- 3b. Layer 2: Input preprocessing (JPEG + spatial
smoothing) as first filter
- 3c. Layer 3: Adversarial detection (MagNet) to flag
suspicious inputs for human review
- 3d. Layer 4: Ensemble for decision stability on
flagged inputs