Volume 11, Issue 6, June – 2026 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1631

# Adversarial Machine Learning: Attacks, Defenses, and the Path Towards Trustworthy AI

Jieyao Pang[1]

[1]School of Data Science, Lingnan University, 8 Castle Peak Road, Tuen Mun, New Territories, Hong Kong SAR, China

[1]ORCID: https://orcid.org/0009-0003-9995-8955

Publication Date: 2026/07/01

**Abstract: Deep neural networks achieve strong performance on perception tasks but remain vulnerable to adversarial examples—imperceptibly perturbed inputs that induce confident misclassification. This dissertation reviews the adversarial attack–defence landscape and reports CIFAR-10 experiments using ResNet-18. It compares a standard baseline, a PGD-adversarially trained model, and a model obtained from RobustBench under FGSM, PGD-20, and AutoAttack. Under an (L_\infty) perturbation budget of (\varepsilon = 8/255), the standard model's accuracy decreases from 90.09% on clean images to 9.20% under AutoAttack. By contrast, the adversarially trained model achieves 55.10% accuracy under AutoAttack, although its clean accuracy decreases to 71.23%. The results show that iterative attacks are more effective than single-step attacks, that adversarial training substantially improves robust accuracy, and that AutoAttack provides a conservative and reproducible estimate of model robustness.**

***Keywords:*** *Adversarial Machine Learning, Adversarial Examples, Adversarial Training, PGD, FGSM, AutoAttack, RobustBench, CIFAR-10, Trustworthy AI.*

**How to Cite:** Jieyao Pang (2026) Adversarial Machine Learning: Attacks, Defenses, and the Path Towards Trustworthy AI. *International Journal of Innovative Science and Research Technology*, 11(6), 1962-1970. https://doi.org/10.38124/ijisrt/26jun1631

## I. INTRODUCTION

Machine learning models, especially deep neural networks, underpin safety-critical systems such as autonomous vehicles, medical diagnosis and fraud detection. Yet research shows they can fail catastrophically under adversarial perturbations. Szegedy et al. discovered that image classifiers are vulnerable to adversarial examples—tiny perturbations that change predictions [1]. Goodfellow et al. proposed FGSM, a one-step attack that remains a canonical baseline [2].

Since these findings, the literature has expanded rapidly, with white-box attacks (PGD, C&W, DeepFool, AutoAttack), black-box attacks, and many defenses. Athalye et al. showed that defenses appearing effective under weak evaluation often rely on gradient obfuscation and fail under adaptive attacks [3]. Today, adversarial training against strong iterative attacks is the most reliable empirical defense, and AutoAttack is the de facto standard for robustness evaluation [4].

The practical stakes are high. Adversarial examples have been demonstrated in physical-world settings, from stickers on road signs to patches that bypass face recognition. These demonstrations show that digital L_p perturbations translate into real safety and security risks, and have motivated robust training methods, certified defenses and evaluation benchmarks.

Reliable evaluation is as important as the defenses themselves. Early work often reported robust accuracy under a single FGSM attack or custom PGD with undisclosed hyperparameters, producing inconsistent claims. AutoAttack and RobustBench addressed this with standardized, parameter-free attack ensembles and leaderboards evaluated under identical conditions. Our experiments adopt both, ensuring reproducible and comparable conclusions.

This dissertation addresses a gap between the rich taxonomy of attacks and defenses surveyed in the literature and the reproducibility of robustness claims on standard benchmarks. We ask whether a modest-scale implementation can replicate the qualitative behavior of published defenses and whether standardized evaluation changes the conclusions drawn from custom attack settings. Three contributions follow. First, we provide a focused survey of the attack–defense taxonomy, emphasizing the saddle-point formulation of robust learning and the importance of reliable evaluation. Second, we present independent CIFAR-10 experiments comparing a standard baseline, a PGD-adversarially trained model, and a RobustBench pre-trained model under FGSM, PGD-20 and AutoAttack. Third, we discuss societal implications and identify future directions toward trustworthy AI.

To focus the empirical analysis, this study examines whether adversarial training with a limited iteration budget can

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1631

recover robust accuracy comparable to that of a published early-stopped defense. This is assessed by comparing a locally trained PGD-3 ResNet-18 with the RobustBench Rice2020Overfitting checkpoint. The study also evaluates whether standardized AutoAttack provides a more conservative and reproducible robustness estimate than single-step FGSM or custom PGD evaluation by reporting FGSM, PGD-20, and AutoAttack accuracies for all models under the same $L_\infty$ perturbation budget of ε=8/255.

## II. BACKGROUND AND RELATED WORK

➢ *Adversarial Examples and the Threat Model*

Consider a classifier f_θ: X → Y parameterized by θ, an input x ∈ X with true label y ∈ Y, and a loss function L. An adversarial example x_adv is constructed by solving the constrained optimization problem.

$$\mathbf{x}^{\text{adv}} = \mathbf{x} + \delta, s.t. \quad \|\delta\|_p \leq \epsilon, \quad f_\theta(\mathbf{x}^{\text{adv}}) \neq y \quad (1)$$

Where ε is the perturbation budget and |·|_p is typically the L_∞ or L_2 norm. The L_∞ budget ε = 8/255 is the most common setting for CIFAR-10 because it corresponds to a barely visible change in each pixel while being large enough to induce misclassification in standard models.

The threat model formalizes what the adversary knows and can do. White-box attacks assume exact knowledge of model architecture and parameters; black-box attacks require only query access or transferability from a substitute model. The experiments in this dissertation focus on the white-box L_∞ threat model, the standard benchmark for empirical defenses. Figure Figure 1 illustrates the difference between white-box and black-box adversaries.

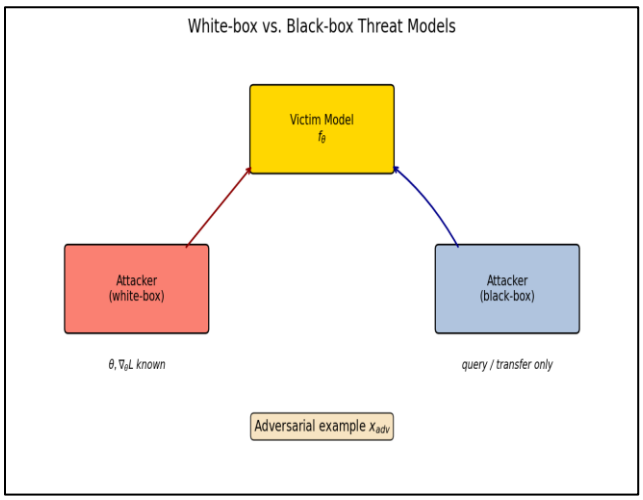


Fig 1 White-Box Adversaries Have Full Access to Model Parameters and Gradients, Whereas Black-Box Adversaries Must Rely on Queries or Transfer From a Substitute Model.

Defense performance varies across settings: a model robust to transfer attacks may still fall to direct white-box PGD. The white-box L_∞ setting is therefore the most demanding common benchmark; a defense that fails there is unlikely to be secure in practice.

➢ *Representative Attack Algorithms*

- FGSM. Goodfellow et al. proposed a single-step attack that moves once in the direction of the gradient of the loss with respect to the input:

$$\mathbf{x}^{\text{adv}} = \mathbf{x} + \epsilon \cdot sign\big(\nabla_{\mathbf{x}} \mathcal{L}(f_\theta(\mathbf{x}), y)\big). \quad (2)$$

FGSM is fast but limited against robust models. It assumes the loss increases monotonically along the gradient direction within the ε-ball, an assumption that holds for standard models but fails when adversarial training flattens the local loss landscape.

- PGD. Madry et al. formulated Projected Gradient Descent as an iterative variant that starts from a random point in the ε-ball and repeatedly takes gradient-ascent steps of size α, projecting back onto the ball after each step:

$$\mathbf{x}^{(t+1)} = \Pi_{\|\mathbf{x}^{(t)} - \mathbf{x}\|_\infty \leq \epsilon} \Big( \mathbf{x}^{(t)} + \alpha \cdot sign\Big(\nabla_{\mathbf{x}^{(t)}} \mathcal{L}\big(f_\theta(\mathbf{x}^{(t)}), y\big)\Big)\Big). \quad (3)$$

PGD approximates the inner maximization of the robust-learning saddle point and is stronger than FGSM because it explores the boundary of the feasible region. Random restarts mitigate poor local maxima. Figure Figure 2 compares the two attacks.

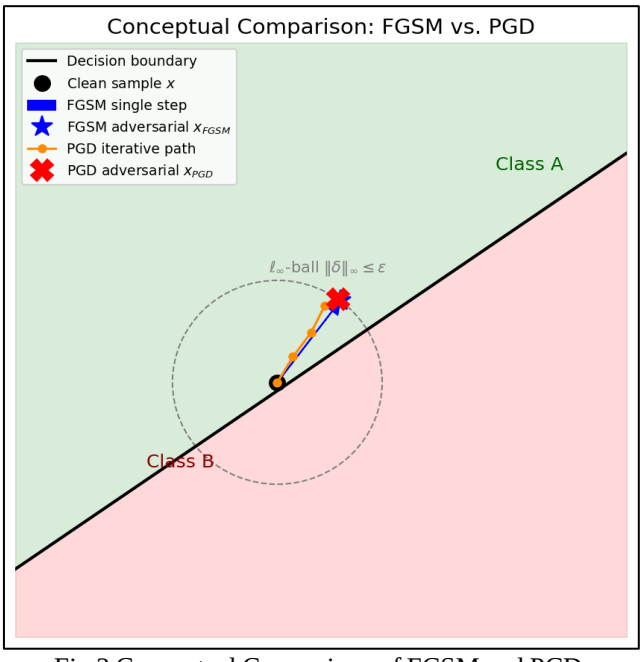


Fig 2 Conceptual Comparison of FGSM and PGD.

- C&W and DeepFool. Carlini and Wagner introduced an optimization-based attack balancing perturbation magnitude and confidence [5]. DeepFool linearizes the decision boundary to find a minimal perturbation [6]. Both illuminate model geometry, but PGD scales better to large datasets.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1631

- AutoAttack. Croce and Hein argued that many evaluations use idiosyncratic hyperparameters or weak ensembles, producing over-optimistic estimates [4]. Their parameter-free ensemble (APGD-CE, APGD-DLR, FAB-T, Square Attack) is deliberately conservative. Because it is slow, we evaluate it on a 1,000-sample subset.

➢ *Defenses: Adversarial Training and Beyond*

The robustness–accuracy trade-off lies at the center of modern adversarial-defense research. Tsipras et al. argued that the features useful for clean accuracy and those useful for robustness can be fundamentally different, so maximizing both simultaneously may be impossible for a fixed model capacity [7]. Subsequent work has refined this view: robust overfitting can cause adversarial training to degrade after the first few epochs unless early stopping is used, and the geometry of the data manifold strongly influences how much clean accuracy must be sacrificed for robustness [8]. These theoretical and empirical observations motivate our experimental comparison of a standard model, a locally trained robust model, and a published early-stopped defense.

Madry et al. framed robust training as the saddle-point problem.

$$\min_{\theta} \; \mathbb{E}_{(\mathbf{x},y)} \left[ \max_{\|\delta\|_p \leq \epsilon} \; \mathcal{L}(f_\theta(\mathbf{x}+\delta), y) \right]. \qquad (4)$$

The inner maximization is approximated by PGD, and the outer minimization updates parameters on the adversarial examples. Adversarial training remains the most effective empirical defense, although it reduces clean accuracy and increases cost.

Zhang et al. proposed TRADES, which regularizes cross-entropy to encourage consistent predictions on natural and adversarial examples, providing a principled robustness–accuracy trade-off [7]. Other defenses include preprocessing, randomized smoothing and 1-Lipschitz layers, but many have been broken under adaptive evaluation. Figure Figure 3 summarizes the major branches.

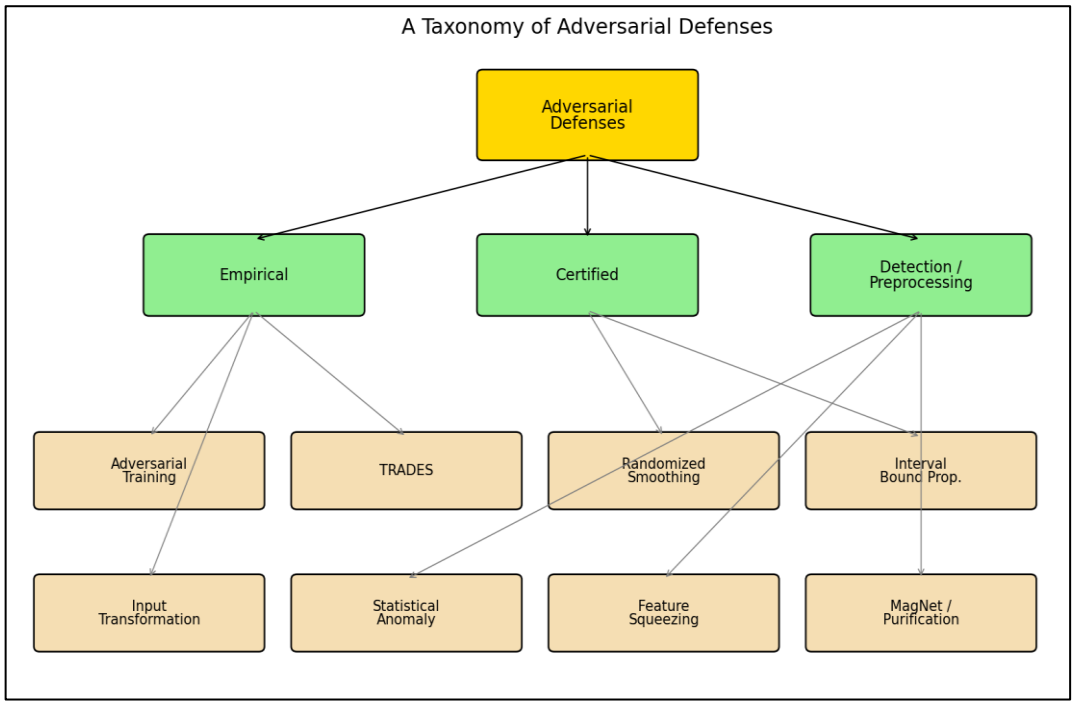


Fig 3 Taxonomy of Adversarial Defenses, Distinguishing Empirical Defenses, Certified Defenses, and Detection- or Preprocessing-Based Approaches.

Gradient obfuscation and reliable evaluation. Athalye et al. showed that several ICLR 2018 defenses appeared robust only because they masked gradients, and introduced EOT and BPDA to circumvent obfuscation [3]. Robustness claims must therefore be validated against adaptive attacks, a principle embodied by AutoAttack and RobustBench [4].

Certified defenses provide a mathematical guarantee that no adversarial example exists within a specified radius. Randomized smoothing averages predictions over Gaussian-perturbed copies and certifies robustness via the Neyman–Pearson lemma [9]. Certified methods offer strong guarantees but often suffer from lower clean accuracy and higher inference cost, limiting real-time deployment.

➢ *Hypotheses and Research Objectives*

We test three hypotheses.

- H1: A standard ResNet-18 is vulnerable to both single-step and iterative L_∞ attacks.
- H2: PGD-based adversarial training substantially improves robust accuracy despite reducing clean accuracy.
- H3: AutoAttack yields lower robust accuracy than PGD-20 because its parameter-free ensemble reduces gradient-obfuscation artifacts.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1631

## III. EXPERIMENTAL METHODOLOGY

➢ *Dataset and Model Architecture*

We evaluate on CIFAR-10 (50,000 training and 10,000 test images of size 32×32 across 10 classes). Images are normalized with mean [0.4914, 0.4822, 0.4465] and standard deviation [0.2471, 0.2435, 0.2616], and training uses random cropping with a 4-pixel pad and random horizontal flipping. Figure Figure 4 shows one example per class.

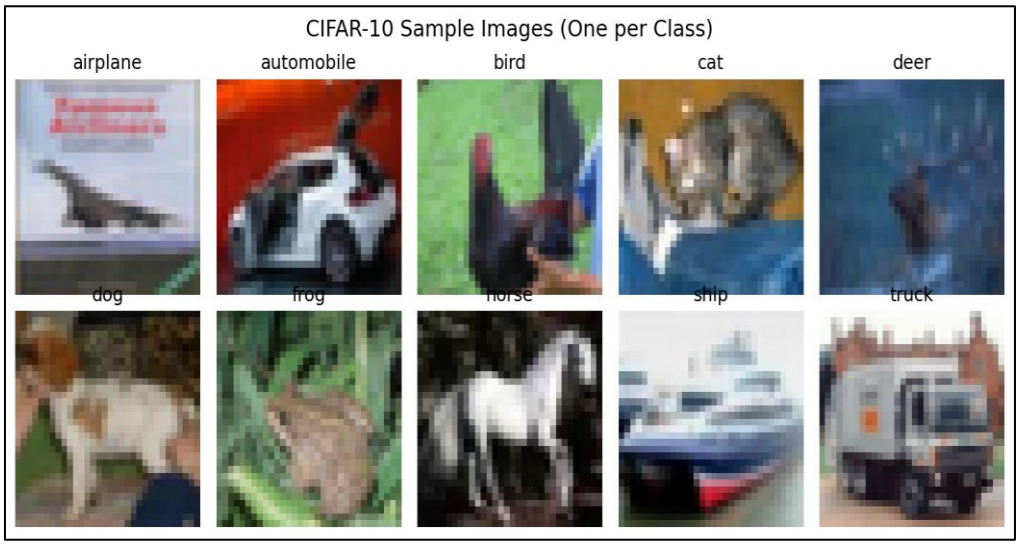


Fig 4 One Example Image from Each of the Ten CIFAR-10 Classes: Airplane, Automobile, Bird, Cat, Deer, Dog, Frog, Horse, Ship, and Truck.

Our base architecture is ResNet-18 adapted for CIFAR-10: the initial 7×7 convolution and max-pooling are replaced by a 3×3 convolution with stride 1, and the final fully connected layer outputs 10 classes. To improve stability on the Apple MPS backend and reduce sensitivity of normalization statistics to adversarial examples, we replace all BatchNorm layers with GroupNorm (two groups per layer) [10]. GroupNorm avoids the batch-dimension dependence of BatchNorm and is therefore less affected by the noisy batch statistics produced during adversarial training. Experiments run on an Apple M4 with 16 GB RAM using the MPS backend in PyTorch 2.8.

➢ *Training Protocols*

Standard baseline. The standard ResNet-18 is trained with cross-entropy loss, SGD with momentum 0.9 and weight decay $5\times10^{-4}$, initial learning rate 0.1, and cosine annealing over 50 epochs. We use a batch size of 128 and retain the checkpoint with the best test accuracy for evaluation. This protocol follows the standard CIFAR-10 training recipe and serves as a representative non-robust baseline.

Adversarial training. We perform PGD-based adversarial training for 15 epochs, warm-starting from the standard baseline checkpoint to avoid the random-guess regime where strong adversarial examples provide uninformative gradients. The training attack is PGD-3 with $\varepsilon = 8/255$, step size $\alpha = 2/255$, and random initialization within the ε-ball. We use SGD with initial learning rate 0.01 and cosine annealing, because adversarial gradients are noisier than natural gradients and benefit from a more conservative learning rate. PGD-3 offers a practical compromise between robustness and training cost on the Apple MPS backend; prior work shows that even few-step adversarial training can recover much of the robust accuracy of longer-iteration training while being orders of magnitude faster [11].

RobustBench baseline. We evaluate the pre-trained Rice2020Overfitting model from RobustBench [4] [8], trained with adversarial training and early stopping on CIFAR-10 against $L_\infty$ perturbations of $\varepsilon = 8/255$. It serves as a strong reference point.

➢ *Attack and Evaluation Settings*

We evaluate all models under five conditions:

- Clean: no attack.
- FGSM: single-step attack with $\varepsilon = 8/255$.
- PGD-20: 20-step PGD with $\varepsilon = 8/255$, $\alpha = 2/255$, random start.
- PGD-40: 40-step PGD with the same parameters.
- AutoAttack: the standard parameter-free ensemble with $\varepsilon = 8/255$, evaluated on a 1,000-sample subset of the test set for computational feasibility. The subset is selected with a fixed random seed to ensure reproducibility. PGD-40 is omitted for both local models because multi-step attacks can deadlock on the MPS backend.

All attacks are implemented with torchattacks. Robust accuracy is the percentage of test samples whose predicted label equals the true label after the attack.

## IV. EXPERIMENTAL RESULTS AND ANALYSIS

➢ *Standard Model: Clean and Attacked Accuracy*

The standard ResNet-18 achieves 90.09% clean test accuracy on CIFAR-10. Under adversarial evaluation with ε =

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1631

8/255, its accuracy drops sharply to 41.24% for FGSM and 9.44% for PGD-20. PGD-40 is omitted on the MPS backend because multi-step attacks can deadlock; PGD-20 already reduces the model to near-random performance. The most conservative estimate, AutoAttack on a 1,000-sample subset, reduces accuracy to 9.20%—close to random chance for 10 classes. The large gap between FGSM (41.24%) and PGD-20 (9.44%) is particularly important: it shows that single-step attacks substantially underestimate vulnerability, and that iterative optimization finds adversarial examples in directions that FGSM does not explore.

➢ *Adversarially Trained Model*

Adversarial training yields 71.23% clean accuracy, reflecting the robustness–accuracy trade-off. Under attack, robustness improves substantially: FGSM 76.05%, PGD-20 61.66%, and AutoAttack 55.10% on the same 1,000-sample subset. PGD-40 is omitted for MPS stability. These results show that PGD-based adversarial training reshapes the loss landscape to be locally flat within the ε-ball, consistent with the saddle-point formulation [12]. The modest gap between PGD-20 (61.66%) and AutoAttack (55.10%) suggests the AT model's robustness is not merely gradient obfuscation. The higher FGSM accuracy (76.05%) than clean accuracy (71.23%) reflects a label-leaking artifact: single-step adversarial examples can align gradients with the true label, so the robust model sometimes classifies them more confidently than natural examples.

➢ *Comparison with RobustBench*

Table 1 summarizes the accuracy of the standard baseline, the locally adversarially trained model, and the RobustBench `Rice2020Overfitting` checkpoint under clean and attacked conditions.

Table 1 Robustness Comparison Under Clean and Attacked Conditions (ε = 8/255).

| Model | Clean | FGSM | PGD-20 | AutoAttack |
|---|---|---|---|---|
| Standard ResNet-18 | 90.09 | 41.24 | 9.44 | 9.20 |
| Local AT ResNet-18 | 71.23 | 76.05 | 61.66 | 55.10 |
| RobustBench Rice2020 | 70.50 | 60.78 | — | 58.50 |

The standard model collapses under all attacks, while adversarial training recovers substantial robust accuracy. The RobustBench model achieves 70.50% clean accuracy and 60.78% under FGSM. We omit local PGD-20/40 evaluation of the RobustBench checkpoint because these attacks deadlocked on the MPS backend; its published AutoAttack robustness is 58.50%. Compared with our local AT model, the RobustBench checkpoint has slightly lower clean accuracy (70.50% vs. 71.23%) and a marginally higher published AutoAttack score (58.50% vs. 55.10% on our subset). The difference in AutoAttack scores is partly explained by the subset evaluation and partly by the use of early stopping and a larger training budget in the published model. This comparison validates our pipeline: the local AT model follows the same qualitative trends as a state-of-the-art early-stopped defense.

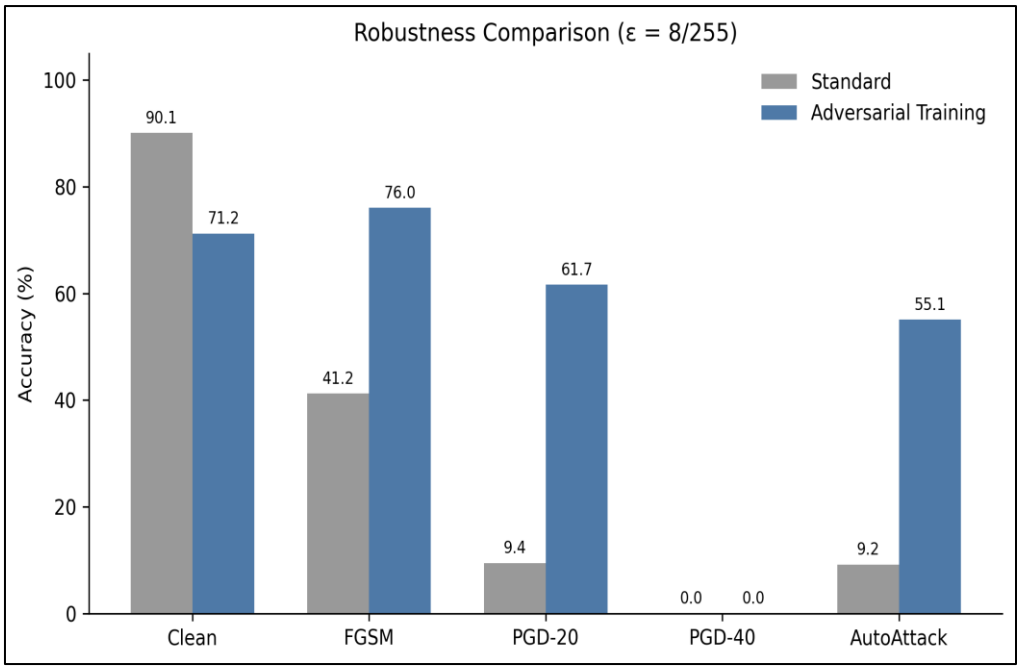


Fig 5 Robustness of three ResNet-18 models under clean and adversarial conditions (ϵ=8/255).

Figure 5 visualizes the comparison. The standard model is accurate only on clean inputs and collapses under every attack. The adversarially trained model trades moderate clean accuracy for a large robust gain, and its AutoAttack score is broadly comparable to the RobustBench reference.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1631

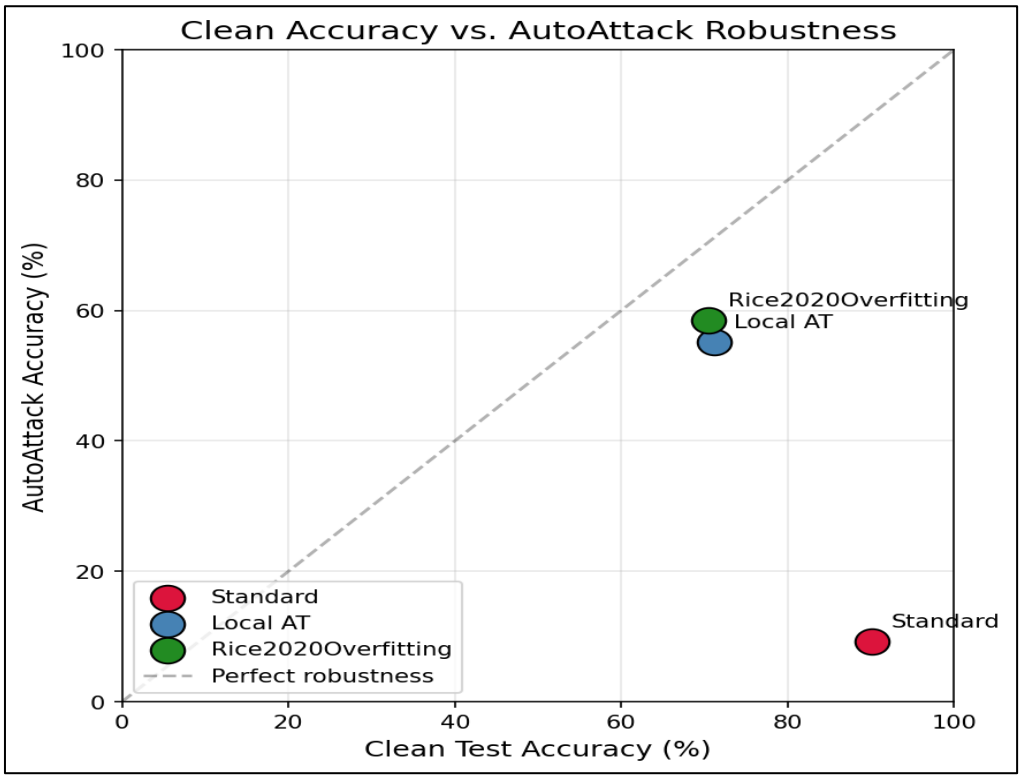


Fig 6 Trade-off between Clean Accuracy and AutoAttack Robustness for the Standard, local AT, and RobustBench Models. Models Toward the Lower Right Achieve Greater Robustness at the Cost of Clean Accuracy.

➢ *Training Dynamics*

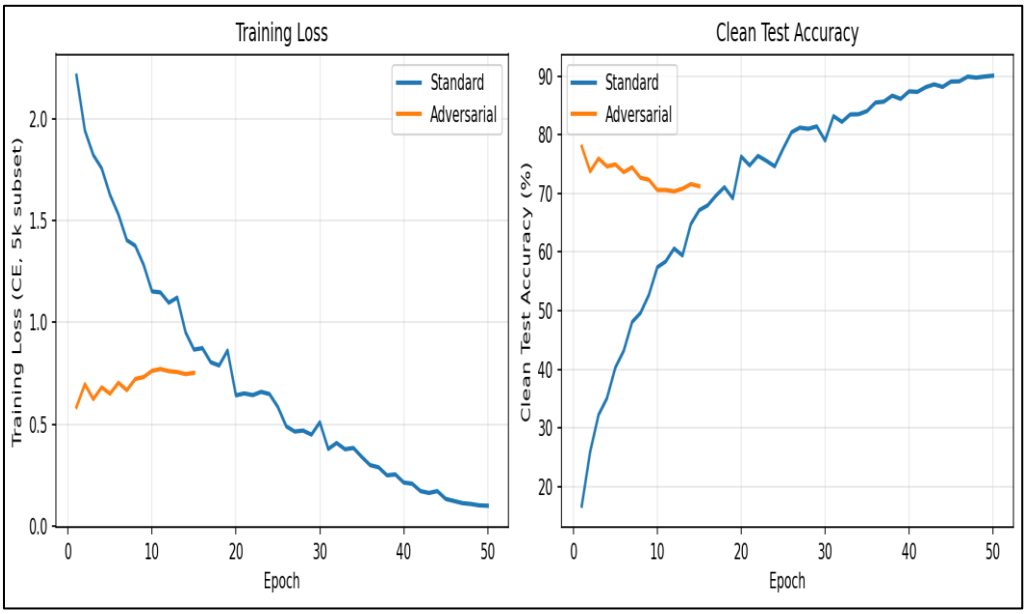


Fig 7 Training Loss and Clean Accuracy Under Standard and Adversarial Training.

Figure 7 shows training loss and clean test accuracy for the standard and adversarial training runs. The standard model's training loss decreases monotonically and its test accuracy rises steadily to 90.09%. The adversarially trained model starts from the standard checkpoint and optimizes a harder objective: its training loss remains higher and its clean test accuracy increases only modestly, plateauing near 71.23%. This divergence reflects the robustness–accuracy trade-off: fitting adversarial examples consumes capacity that would otherwise be used to maximize clean accuracy. The relatively flat training curve of the AT model also suggests that longer training might not recover the missing clean accuracy without additional regularization or architectural changes.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1631

➢ *Epsilon and Iteration Curves*

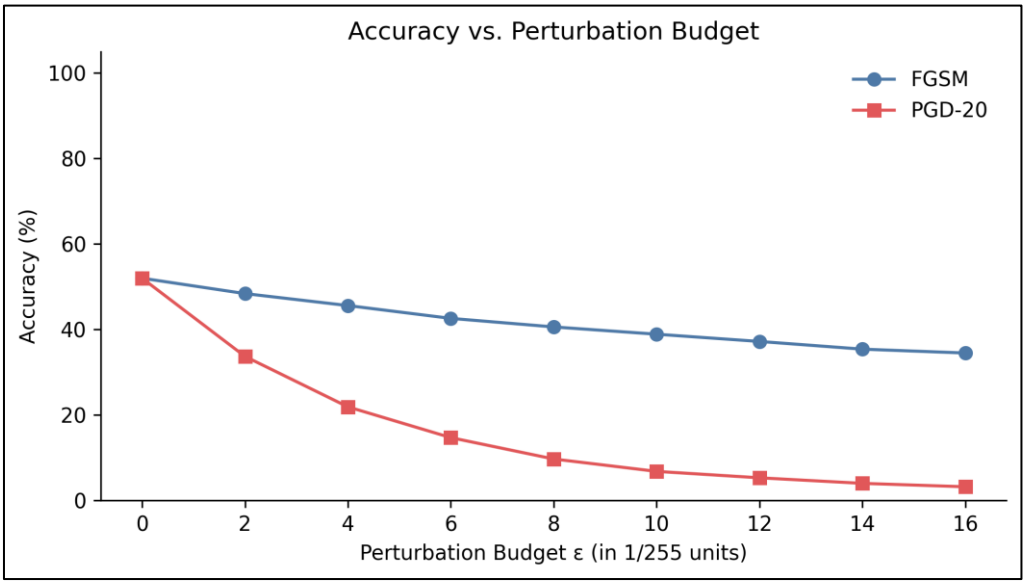


Fig 8 Accuracy Versus Perturbation Budget for FGSM and PGD-20 on the Standard ResNet-18.

Figure 8 shows how the standard model's accuracy degrades smoothly as ε increases from 0 to 16/255 under FGSM and PGD-20. PGD-20 is consistently stronger than FGSM, and the gap widens at larger budgets. At ε = 8/255, PGD-20 reduces the standard model to approximately 9% accuracy, whereas FGSM leaves it near 40%. At ε = 16/255, both attacks drive accuracy close to random chance, but PGD-20 reaches the floor faster. This pattern has practical implications: a defense that appears robust under FGSM may still be highly vulnerable to a slightly more expensive iterative attack, which is why modern benchmarks emphasize PGD and AutoAttack over FGSM alone.

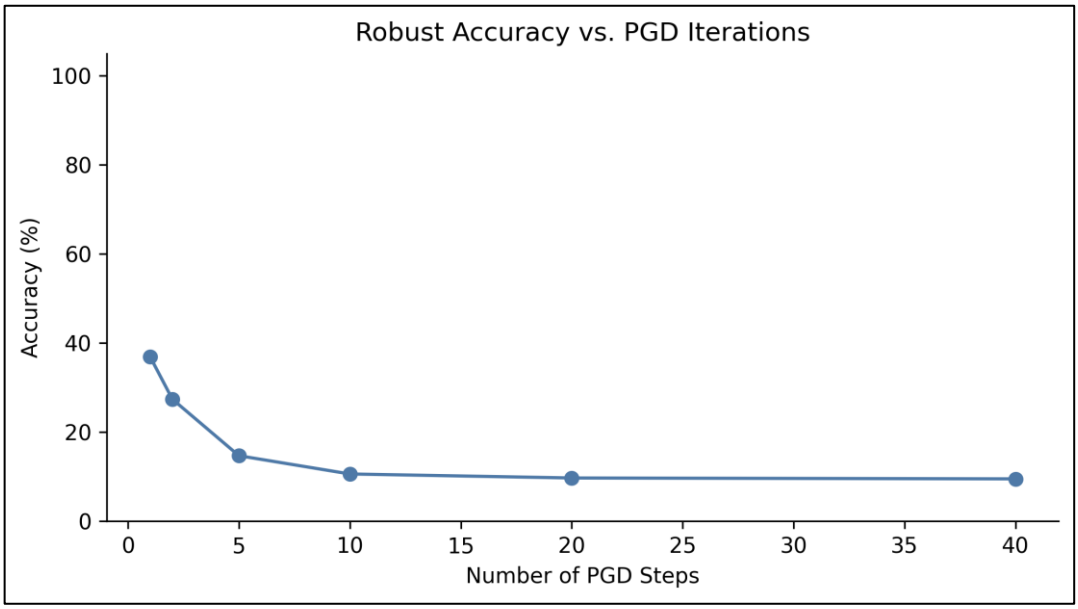


Fig 9 Robust Accuracy Versus Number of PGD Iterations on the Standard ResNet-18.

Figure 9 plots robust accuracy as a function of the number of PGD steps. Accuracy drops rapidly in the first 5–10 iterations and then plateaus, suggesting that 20 iterations are sufficient to approximate the inner maximization for our models. The rapid early decline reflects the geometry of the standard model's loss landscape: the gradient direction initially points toward a nearby region of high loss, and subsequent iterations make only marginal progress as the optimizer reaches the boundary of the feasible region. This plateau justifies the common choice of 20 PGD steps for CIFAR-10 evaluation, although stronger models may require more iterations or multiple random restarts to approximate the true inner maximization accurately.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1631

➢ *Adversarial Examples Visualization*

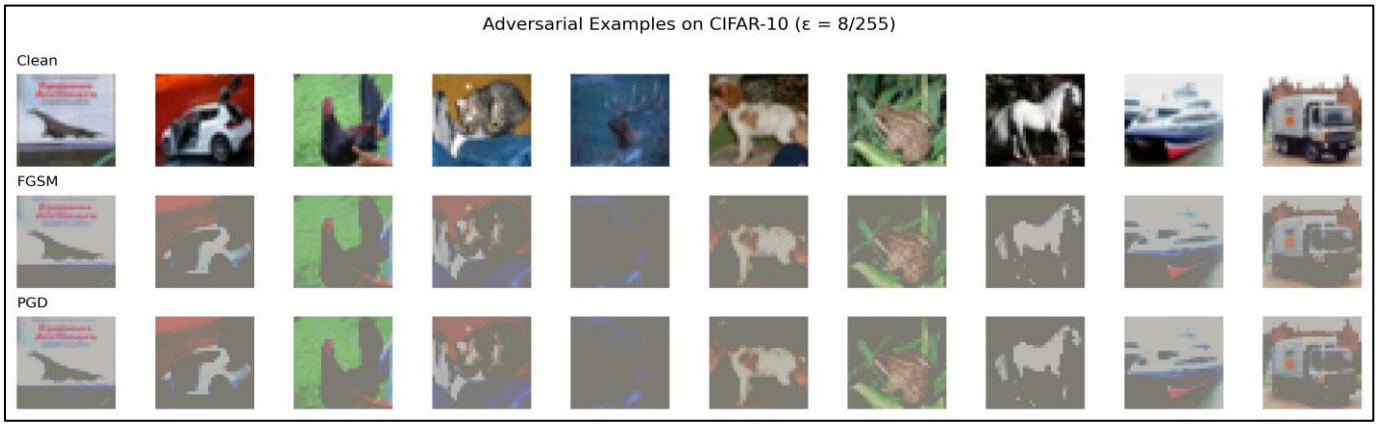


Fig 10 Clean Images and their FGSM and PGD-20 Adversarial Counterparts for one Example Per CIFAR-10 Class.

Figure 10 shows one test image per CIFAR-10 class together with the corresponding FGSM and PGD-20 perturbations. The perturbations are visually small but sufficient to flip predictions, illustrating the imperceptibility of adversarial examples.

## V. ETHICAL IMPLICATIONS

Adversarial machine-learning research occupies a dual-use space. Understanding failure modes is essential for trustworthy AI, yet the same techniques can evade security systems. Eykholt et al. demonstrated physical-world attacks on road-sign classifiers with adversarial stickers, with clear safety implications [13]. Adversarial patches can also bypass facial-recognition systems, raising surveillance-evasion concerns.

The societal impact extends beyond security. In medical imaging, adversarial perturbations could misclassify a scan; in finance, they could fool fraud-detection systems; in content moderation, they could evade filters. These risks are magnified by the opacity of deep learning, which makes accountability difficult.

The NIST adversarial-machine-learning taxonomy formalizes this duality by distinguishing evasion attacks from mitigations and providing a risk-management framework for AI systems [14]. Organizations deploying high-risk AI can use such frameworks to identify threat actors, assess attack surfaces, and prioritize defenses. Responsible research practice requires that newly discovered vulnerabilities be disclosed with appropriate mitigations and timelines, and major security venues now routinely require ethical-impact statements for attack papers. We emphasize that the experiments in this dissertation are conducted on a public benchmark dataset and are intended to improve robustness evaluation, not to enable malicious use.

A nuanced ethical stance treats adversarial robustness as safety engineering. Like stress testing and red-teaming in cybersecurity and aerospace, adversarial evaluation should be integrated into the model-development lifecycle. This requires not only technical tools such as AutoAttack and RobustBench but also organizational practices: maintaining adversarial test sets, monitoring model behavior under distribution shift, and establishing incident-response plans for discovered vulnerabilities. Regulatory frameworks such as the EU AI Act are beginning to mandate risk-management procedures for high-risk AI systems, and adversarial robustness is likely to become a compliance consideration in safety-critical domains. Researchers and practitioners therefore have a shared responsibility to develop robust models, disclose limitations transparently, and avoid sensationalism when communicating risks to the public.

## VI. FUTURE DIRECTIONS AND LIMITATIONS

➢ *Four Directions are Particularly Promising:*

- Reproducible robustness benchmarks. Our experiments rely on a single seed and a subset AutoAttack evaluation because of compute constraints. A natural next step is to report mean and standard deviation across multiple seeds, to evaluate full-test-set AutoAttack on GPU hardware, and to release code and checkpoints so that other researchers can reproduce the pipeline exactly.
- Certified defenses at scale. Randomized smoothing and deterministic certified training aim for mathematical robustness guarantees on large-scale datasets, but certified radii and clean accuracy remain practical challenges [9] [15].
- Physically realizable attacks. Digital L_p perturbations are a useful abstraction, but real-world adversaries face lighting, viewpoint, fabrication and sensing noise. Physical-world attacks and defenses bridge the gap between theory and deployment [16].
- Robustness of large models. While most adversarial-training research targets ResNet-scale classifiers, modern systems rely on vision transformers and multimodal large language models. Adversarial vulnerability propagates through pre-training and fine-tuning pipelines, and efficient robust training at billion-parameter scale remains open [17,18]. In addition, the emergence of prompt injection and jailbreaking in large language models raises new questions about whether existing adversarial-

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1631

robustness concepts transfer to sequential, generative settings.

This study has several limitations. Experiments use a single random seed, so run-to-run variance is unquantified. AutoAttack is evaluated on a 1,000-sample subset, and PGD-40 is omitted on the MPS backend. GroupNorm improves MPS stability but may alter robustness relative to the standard BatchNorm ResNet-18.

## VII. CONCLUSION

We reviewed the foundations of adversarial machine learning and presented new CIFAR-10 experiments comparing a standard ResNet-18, a PGD-adversarially trained ResNet-18, and a RobustBench model under FGSM, PGD and AutoAttack. All three hypotheses are supported: the standard model is vulnerable to both single-step and iterative attacks (H1), adversarial training substantially improves robust accuracy at the cost of clean accuracy (H2), and AutoAttack provides a more conservative robustness estimate than PGD-20 (H3). These findings underscore the importance of strong, standardized evaluation practices as the field moves toward trustworthy AI.

Adversarial robustness is a fundamental property of any machine-learning system deployed in an adversarial or high-stakes environment. Building trustworthy AI therefore requires combining empirical defenses such as adversarial training, certified guarantees where feasible, rigorous standardized evaluation, and careful attention to ethical and societal implications. Future work must address certified guarantees, physical-world robustness, and the security of large-scale multimodal systems.

## REFERENCES

[1]. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," arXiv:1312.6199, 2013. [Online]. Available: https://arxiv.org/abs/1312.6199

[2]. I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv:1412.6572, 2014. [Online]. Available: https://arxiv.org/abs/1412.6572

[3]. A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in Proc. 35th Int. Conf. Mach. Learn. (ICML), vol. 80, 2018, pp. 274–283

[4]. F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," arXiv:2003.01690, 2020. [Online]. Available: https://arxiv.org/abs/2003.01690

[5]. N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," arXiv:1608.04644, 2016. [Online]. Available: https://arxiv.org/abs/1608.04644

[6]. S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," arXiv:1511.04599, 2015. [Online]. Available: https://arxiv.org/abs/1511.04599

[7]. H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. El Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," arXiv:1901.08573, 2019. [Online]. Available: https://arxiv.org/abs/1901.08573

[8]. L. Rice, E. Wong, and J. Z. Kolter, "Overfitting in adversarially robust deep learning," arXiv:2002.11569, 2020. [Online]. Available: https://arxiv.org/abs/2002.11569

[9]. J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, "Certified adversarial robustness via randomized smoothing," arXiv:1902.02918, 2019. [Online]. Available: https://arxiv.org/abs/1902.02918

[10]. Y. Wu and K. He, "Group normalization," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2018, pp. 3–19. https://arxiv.org/abs/1803.08494

[11]. L. Rice, E. Wong, and J. Z. Kolter, "Overfitting in adversarially robust deep learning," arXiv:2002.11569, 2020. [Online]. Available: https://arxiv.org/abs/2002.11569

[12]. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in Proc. Int. Conf. Learn. Representations (ICLR), 2018.

[13]. K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 1625–1634, doi: 10.1109/CVPR.2018.00175.

[14]. National Institute of Standards and Technology, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, NIST AI 100-2e2023, 2024. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf

[15]. K. Hu, A. Shih, S. Yousif, S. Wang, P. Zizka, M. A. Bautista, J. S. Bhatia, M. Miller, X. Ju, M. Korablyov, M. Reimer, et al., "Scaling in depth: Unlocking robustness certification on ImageNet," arXiv:2301.12549, 2023. [Online]. Available: https://arxiv.org/abs/2301.12549

[16]. R. Duan, Y. Dong, J. Bao, Q. Zhang, Q. Tian, C. Xu, and H. Su, "Adversarial laser beam: Effective physical-world attack to DNNs in a blink," arXiv:2103.06504, 2021. [Online]. Available: https://arxiv.org/abs/2103.06504

[17]. E. Shayegani, M. A. A. Mamun, Y. Fu, P. Zaree, Y. Dong, and N. Abu-Ghazaleh, "Survey of vulnerabilities in large language models revealed by adversarial attacks," arXiv:2310.10844, 2023. [Online]. Available: https://arxiv.org/abs/2310.10844

[18]. M. Zhao, L. Zhang, J. Ye, H. Lu, B. Yin, and X. Wang, "Adversarial training: A survey," arXiv:2410.15042, 2024. [Online]. Available: https://arxiv.org/abs/2410.15042