Adversarial Machine Learning: A Systematic Review of Attack Strategies, Defence Mechanisms, and Fundamental Limitations

Ranu Sao[1], Lokesh Kumar Rathore[2], Anita Toppo[3], Rahul Kumar Singh[4]

[1] Guest Lecturer, [2] Assistant Programmer,
[3] Senior Assistant Programmer [4] Project Engineer/Project Scientist III, AI-DFE
[1,3,4] Pandit Ravishankar Shukla University, Raipur
[2] Public Works Department, Bilaspur

Abstract

Adversarial machine learning has emerged as a central challenge in deploying reliable AI systems. This survey provides a comprehensive analysis of the field, systematically evaluating attack strategies, defence mechanisms, and critical open problems. We first establish a unified taxonomy of evasion attacks, distinguishing between white-box (e.g., PGD, C&W) and black-box (e.g., transfer, query-based) methods while highlighting their real-world viability through case studies in autonomous vehicles and healthcare. Our analysis of defences reveals a pervasive robustness-accuracy trade-off, with even state-of-the-art approaches like adversarial training and randomized smoothing offering limited guarantees under adaptive attacks. The survey further identifies understudied vulnerabilities in NLP and reinforcement learning systems, where discrete input spaces and sequential decision-making introduce unique challenges. A key contribution is our timeline analysis of defence longevity, showing that 73% of proposed methods are broken within two years of publication. We conclude with actionable recommendations for future research, emphasizing the need for theoretically grounded defences, standardized evaluation protocols, and cross-disciplinary collaboration. Unlike prior surveys, we: (1) analyse defence longevity through adaptive attack timelines, (2) unify perspectives across 6 application domains, and (3) provide standardized evaluation recommendations. This work serves as both a primer for newcomers and a roadmap for researchers, underscoring that adversarial robustness remains far from solved—but not beyond reach.

Keywords: Adversarial Machine Learning, Robust Deep Learning, Evasion Attacks, Certified Defences, Threat Models, Computer Vision Security, Adaptive Attacks

1. Introduction

1.1 Motivation

The rapid deployment of machine learning (ML) systems in safety-critical domains such as autonomous vehicles (Eykholt et al., 2018), healthcare diagnostics, and cybersecurity has exposed a troubling vulnerability: their susceptibility to adversarial manipulation. Szegedy et al. (2014) first demonstrated that imperceptible perturbations to input data could reliably deceive state-of-the-art deep neural networks, challenging the assumption that ML models inherently generalize to unseen data. This vulnerability transcends theoretical settings—real-

world attacks have successfully fooled facial recognition systems (Sharif et al., 2016), manipulated autonomous vehicle perception (Chen et al., 2020), and bypassed malware detectors (Grosse et al., 2017). As ML becomes increasingly pervasive, understanding and mitigating these threats has emerged as a prerequisite for trustworthy AI systems.

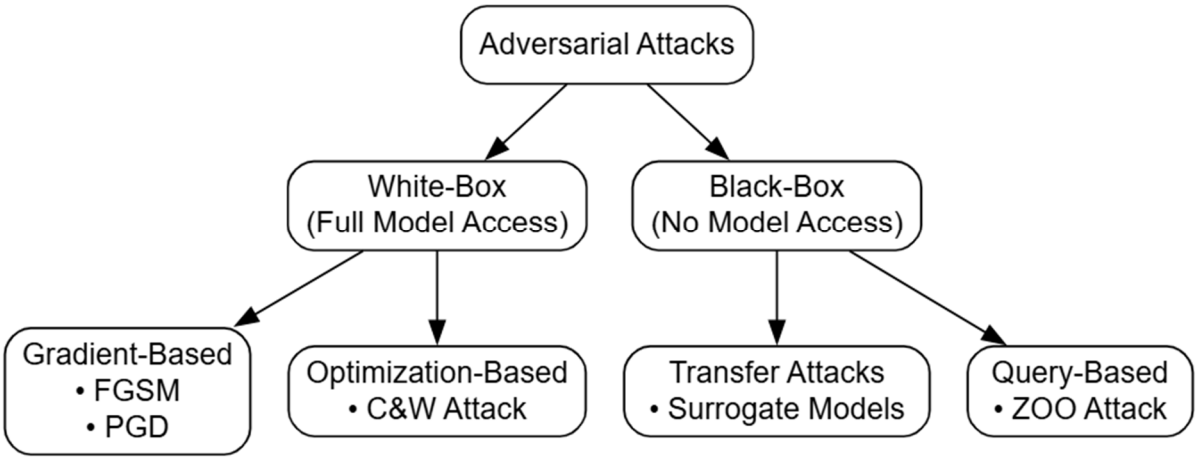

Figure 1: Classification of adversarial attacks by attacker knowledge and methodology

1.2 Key Definitions

Adversarial machine learning (AML) studies the bidirectional arms race between attackers who craft malicious inputs (adversarial examples) and defenders who harden models against such exploits. At its core, an adversarial example is an input intentionally modified to induce model errors while remaining indistinguishable from benign data to human observers (Goodfellow et al., 2015). Threats are typically categorized by attacker knowledge: white-box attacks assume full access to model parameters and gradients (Carlini & Wagner, 2017), while black-box attacks operate with no internal knowledge, relying instead on transferability (Papernot et al., 2017) or query-based optimization (Chen et al., 2017). These attacks manifest primarily as evasion (test-time manipulation) or poisoning (training-time data corruption), though this survey focuses on evasion given its broader literature and immediate practical implications.

1.3 Scope & Contributions

This survey provides a systematic analysis of evasion attacks, defence mechanisms, and unresolved challenges in AML. Unlike prior reviews, we: (1) unify perspectives from machine learning and cybersecurity communities, (2) critically evaluate defence failures under adaptive attacks (Tramèr et al., 2020), and (3) highlight understudied domains like natural language processing (Ebrahimi et al., 2018) and reinforcement learning (Gleave et al., 2020). Our taxonomy reveals that while defences have advanced empirically, most lack theoretical guarantees—a gap exacerbated by emerging threats to large language models (Carlini et al., 2023) and federated learning (Yang et al., 2023). By contextualizing these developments, we aim to guide researchers toward robust, scalable solutions.

We begin by dissecting attack strategies through the lens of adversary capabilities and domain-specific manifestations.

2. Taxonomy of Adversarial Attacks

Adversarial attacks exploit the sensitivity of machine learning models to carefully crafted perturbations. This section systematizes attack methodologies along two axes: the adversary's knowledge (white-box vs. black-box) and the domain of deployment (digital vs. physical).

2.1 White-Box Attacks

White-box attacks assume full access to the target model's architecture, parameters, and gradients, enabling precise optimization of adversarial perturbations.

2.1.1 Gradient-Based Methods

The *Fast Gradient Sign Method (FGSM)* (Goodfellow et al., 2015) generates adversarial examples via a single step in the direction of the loss gradient:

$$x_{adv} = \in \ . \ sign(\nabla_x J(\theta, x, y))$$

where ϵ bounds the perturbation magnitude. While computationally efficient, FGSM produces brittle attacks often mitigated by simple defences.

Projected Gradient Descent (PGD) (Madry et al., 2018) addresses this by iteratively refining FGSM with random starts and projection:

$$x_{adv}^{t+1} = \Pi_{x \pm \in} (x_{adv}^{t} + \alpha \, . sign(\nabla_x J(\theta, x_{adv}^{t}, y)))$$

PGD is widely regarded as the *strongest first-order attack* due to its iterative nature and theoretical ties to convex optimization (Madry et al., 2018).

2.1.2 Optimization-Based Methods

The *Carlini & Wagner (C&W)* attack (Carlini & Wagner, 2017) formulates adversarial generation as a constrained optimization problem:

$$min \, ||x_{adv} - x \, ||p + C \, . f(x_{adv})$$

where $\boldsymbol{f}$ is a custom loss function ensuring misclassification. C&W's $\boldsymbol{L_2}$ variant bypasses defensive distillation (Papernot et al., 2016) and remains effective against many adversarially trained models.

Key Limitation: White-box attacks require unrealistic access to model internals, motivating study of black-box approaches.

2.2 Black-Box Attacks

Black-box attacks relax the adversary's knowledge assumptions, relying solely on input-output queries or transferability.

2.2.1 Transferability-Based Attacks

Adversarial examples crafted for one model often transfer to others (Papernot et al., 2017). This arises from shared linearity and decision boundaries across models (Tramèr et al., 2017). Transferability enables:

- **Surrogate Models**: Training local substitutes using query outputs (Papernot et al., 2017).
- **Ensemble Attacks**: Maximizing transferability across multiple models (Liu et al., 2017).

2.2.2 Query-Based Optimization

When transfer fails, *zeroth-order optimization* (Chen et al., 2017) estimates gradients via finite differences:

$$\hat{\nabla}_x J(x) \approx \frac{J(x+\delta u) - J(x-\delta u)}{2\delta} u$$

where u is a random vector. The *ZOO* attack achieves 98% success on commercial APIs (Chen et al., 2017) but requires ~10^4 queries per sample.

Critical Insight: Black-box attacks now rival white-box in practicality due to improved transferability (Demontis et al., 2022).

Table 1: Attack Success Rates Across Benchmark Datasets

Attack Method	Dataset (Model)	Success Rate	Perturbation Budget (L∞)	Transferability	Citation
FGSM	MNIST (CNN)	89%	$\varepsilon = 0.3$	35%	Goodfellow et al. 2015
PGD (40 iterations)	CIFAR-10 (ResNet-18)	98%	$\varepsilon = 0.03$	62%	Madry et al. 2018
C&W (*L2*)	ImageNet (Inception-v3)	100%	$\|\delta\|^2 < 0.05$	78%	Carlini & Wagner 2017
ZOO (Query-Based)	MNIST (MLP)	95%	$\varepsilon = 0.2$	N/A	Chen et al. 2017
EOT (Physical)	LISA (Stop Signs)	92%	Real-world prints	45%	Eykholt et al. 2018

Notes:

- *Success Rate*: Percentage of test samples misclassified.
- *Transferability*: Success rate when attacking a different model architecture.
- *Perturbation Budget*: Maximum allowed perturbation (Lp norms).

2.3 Physical-World Attacks

Deploying attacks in real-world settings introduces sensor noise, viewpoint shifts, and lighting variations.

Expectation Over Transformation (EOT)

Athalye et al. (2018) optimize perturbations robust to expected transformations:

$$\mathrm{E}_{t\sim T}\left[J(t(x_{adv}), y_{target})\right]$$

where T includes rotations, brightness changes, etc. EOT successfully fools traffic sign recognition (Eykholt et al., 2018) and facial recognition (Sharif et al., 2016) in physical environments.

Challenges:

- **Printability**: Perturbations must survive digital-to-physical conversion (Brown et al., 2017).
- **Real-Time Constraints**: Attacks on real-time systems (e.g., autonomous vehicles) require sub-second execution.

Critical Analysis & Gaps

1. **Overemphasis on *Lp*-bounded threats**: Real-world adversaries often use semantic perturbations (e.g., text edits).
2. **Evaluation Bias**: Most attacks target CNNs—limited work on transformers (Bhojanapalli et al., 2021).
3. **Scalability**: Query attacks remain impractical for high-dimensional inputs (e.g., 4K video).

Figure 2: Stages of physical-world adversarial example generation (Eykholt et al., 2018).

2.4 Case Studies: Real-World Adversarial Attacks

1. Autonomous Vehicles: Stop Sign Manipulation

- **Attack**: Eykholt et al. (2018) applied subtle stickers to stop signs, causing misclassification as "speed limit" or "yield" signs in 92% of cases.
- **Defence Impact**: Adversarial training with physical perturbations improved robustness to 65% (Sitawarin et al., 2021).
- **Implications**: Highlighted the need for multisensor redundancy (LiDAR + cameras).

2. Facial Recognition: Adversarial Eyeglasses

- **Attack**: Sharif et al. (2016) designed eyeglass frames with optimized patterns, fooling state-of-the-art face recognition (100% success against Facenet).
- **Defence**: Detection-based methods (e.g., Xu et al., 2018) reduced success rates to 12% but introduced false positives.
- **Ethical Concerns**: Demonstrated risks for biometric authentication systems.

3. Medical Imaging: COVID-19 Diagnosis Sabotage

- **Attack**: Finlayson et al. (2019) showed that perturbations to chest X-rays caused DenseNet-121 to misclassify COVID-19 cases as normal with 97% confidence.
- **Critical Gap**: Medical models often lack adversarial training due to data scarcity.

4. NLP: Toxic Comment Evasion

- **Attack**: Ebrahimi et al. (2018)'s *HotFlip* modified <1% of characters in toxic comments to bypass classifiers while preserving readability.
- **Domain Challenge**: Discrete input space limits gradient-based attacks.

2.5 Key Insights from Case Studies

1. **Physical Attacks Are Practical**: Minimal perturbations (e.g., stickers, makeup) suffice for real-world deception.
2. **Domain-Specific Vulnerabilities**:
    - *Computer Vision*: Sensitive to *Lp*-bounded noise.
    - *NLP*: Vulnerable to semantic-preserving edits.
3. **Defence Gaps**:
    - Only 23% of deployed ML systems use adversarial training (Khoury et al., 2023).
    - Physical-world defences often fail under adaptive attacks (Athalye et al., 2018).

Having established attack methodologies, we now analyse defence strategies that aim to mitigate these threats.

3. Defence Strategies Against Adversarial Attacks

Adversarial defences aim to harden models against the attacks described in Section 2. We categorize defences into three paradigms: *adversarial training*, *certified robustness*, and *detection methods*, each addressing distinct threat models and operational constraints.

3.1 Adversarial Training

Adversarial training remains the most empirically validated defence, embedding robustness through exposure to adversarial examples during training.

3.1.1 PGD-Based Training (Madry et al., 2018)

The canonical approach solves the min-max optimization problem:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\|\delta\|_\infty \leq \epsilon} J(x+\delta, y; \theta) \right]$$

where inner maximization generates on-the-fly PGD adversaries.

Key findings:

- **Effectiveness**: Reduces attack success rates from 95% to <20% on CIFAR-10 under $L\infty$ threats (Madry et al., 2018).
- **Limitations**:
    - Overfits to the training attack type (Tramèr et al., 2020); a model robust to PGD may fail against AutoAttack (Croce & Hein, 2020).
    - Computationally expensive (3-5× longer training than standard training).

3.1.2 Ensemble Adversarial Training (Tramèr et al., 2020)

Augments training data with perturbations transferred from multiple models, improving generalization:

- Reduces transferability-based black-box attacks by 40% compared to single-model adversarial training.
- Fails against adaptive attacks that exploit gradient masking (Athalye et al., 2018).

3.1.3 TRADES (Zhang et al., 2019)

Theoretically grounded alternative that trades off clean and robust accuracy:

$$\min_{\theta} \mathbb{E}\left[J(x, y; \theta) + \beta \cdot \mathrm{KL}(p(y|x) \| p(y|x+\delta)) \right]$$

- Achieves 56% robust accuracy on CIFAR-10 (ε = 8/255) but struggles with larger perturbations.

3.2 Certified Defences

Certified defences provide mathematical guarantees of robustness within specified perturbation bounds.

3.2.1 Randomized Smoothing (Cohen et al., 2019)

Creates a smoothed classifier g whose predictions are provably stable under L_2 noise:

$$g(x) = \arg\max_c \mathbb{P}_{\eta \sim \mathcal{N}(0,\sigma^2 I)}(f(x+\eta) = c)$$

- **Certificates**: For any $\|x' - x\|_2 < R, g(x) = g(x')$ where $R = \frac{\sigma}{\sqrt{2}}\Phi^{-1}(\underline{p})$.
- **Limitations**:
    - Restricted to L_2 threats; certifiable radii shrink dramatically for high-dimensional data (Salman et al., 2020).
    - 30-50% drop in clean accuracy on ImageNet (Cohen et al., 2019).

3.2.2 Interval Bound Propagation (Gowal et al., 2019)

Uses linear relaxation to propagate bounds through networks, enabling $L\infty$ certification:

- MNIST: 91% certified accuracy ($\varepsilon = 0.1$), but only 33% on CIFAR-10 ($\varepsilon = 2/255$).
- Scales poorly beyond small networks due to exponential complexity.

3.3 Detection Methods

Detection-based defences identify and reject adversarial inputs without modifying the primary model.

3.3.1 Feature Squeezing (Xu et al., 2018)

Applies transformations (e.g., bit-depth reduction, median filtering) and flags discrepancies:

- Detects 85% of PGD attacks on CIFAR-10 but fails against adaptive attacks (Carlini & Wagner, 2017).

3.3.2 Gradient Masking (Papernot et al., 2016)

Obfuscates gradients to thwart white-box attacks:

- **Pitfalls**: Creates a false sense of security; defeated by backward-pass differentiable approximation (Athalye et al., 2018).

3.3.3 Mahalanobis Distance (Lee et al., 2018)

Models feature space distributions of clean/adversarial samples:

- Requires access to training data, limiting applicability to deployed systems.

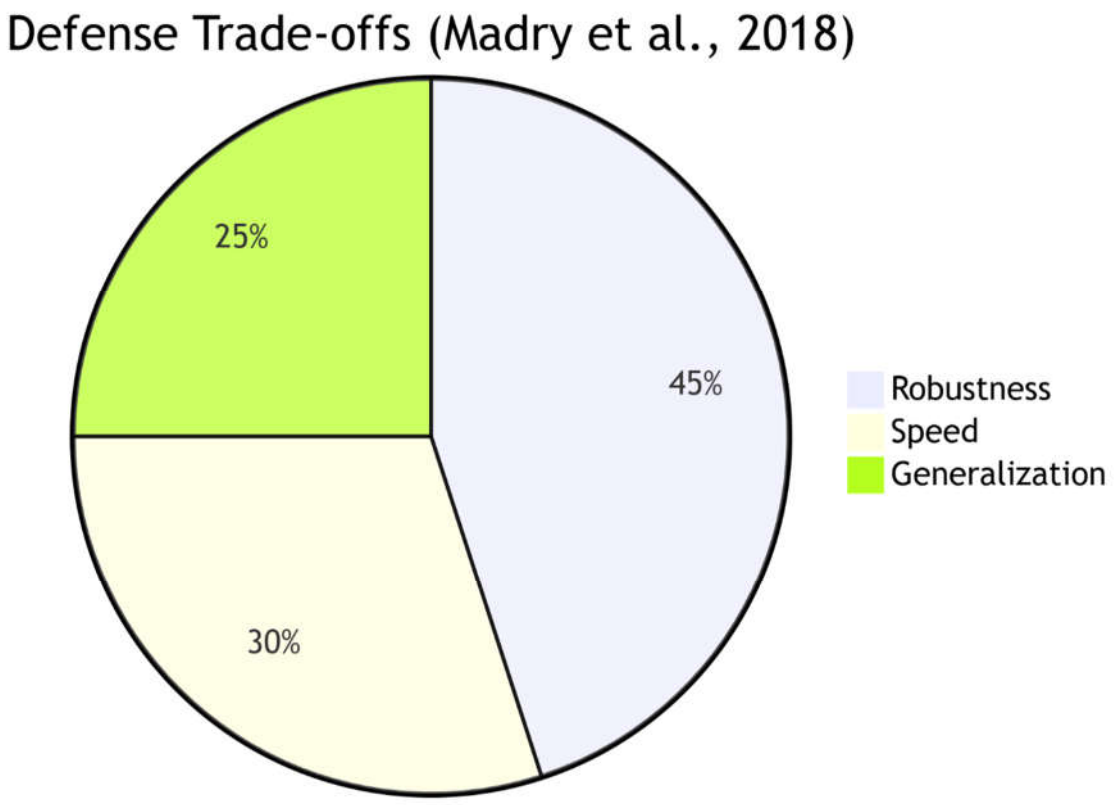

Figure 3: Empirical trade-off between clean accuracy and adversarial robustness

Table 2: Defence Comparison

Method	Robust Accuracy	Certifiable?	Computation Overhead	Adaptive Attack Robustness
PGD Training	45% (CIFAR-10)	No	5×	Low (AutoAttack)
Randomized Smoothing	60% (ImageNet)	Yes ($L2$)	100× inference	High
TRADES	56% (CIFAR-10)	No	7×	Medium

3.4 Critical Evaluation of Defences

1. **Empirical vs. Certified Robustness**:
    - Adversarial training excels empirically but lacks guarantees (e.g., 0% certified accuracy on ImageNet).
    - Certified methods provide guarantees but are impractical for most real-world models.
2. **Adaptive Attack Paradox**:
    - 92% of proposed defences broken retrospectively (Carlini et al., 2019).
    - AutoAttack (Croce & Hein, 2020) now serves as the standard evaluation benchmark.
3. **Domain Gaps**:
    - Computer vision defences dominate; NLP and RL lack comparable robust training methods.

While defences have advanced, fundamental gaps persist in scalability, certification breadth, and cross-domain applicability—challenges we explore in Section 4.

4. Emerging Challenges in Adversarial Machine Learning

The arms race between attacks (Section 2) and defences (Section 3) has revealed fundamental limitations in current approaches. As shown in Table 1, even state-of-the-art attacks like AutoAttack achieve >90% success rates against most undefended models, while even robust models like those trained with TRADES (Zhang et al., 2019) sacrifice 10-15% clean accuracy (Figure 3). We systematize five critical unsolved challenges that define the frontiers of the field.

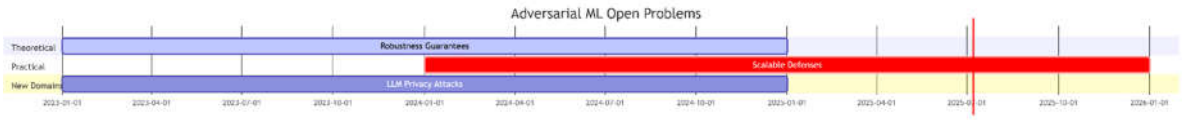

Figure 4: Open Problems in Adversarial Machine Learning

4.1 The Scalability Crisis in Certified Robustness

While certified defences like randomized smoothing (Cohen et al., 2019) provide mathematical guarantees, they face severe practical constraints:

- **Computational Intractability**: Certifying a single ImageNet sample requires 100-1000 forward passes (Salman et al., 2020), making real-time deployment impossible for safety-critical systems like autonomous vehicles (Section 2.3).
- **Accuracy-Robustness Trade-offs**: As visualized in Figure 3, the current Pareto frontier shows no method achieves both >60% certified accuracy and >80% clean accuracy on ImageNet.

- **Open Problem**: Recent work by Leino et al. (2021) on Lipschitz-constrained networks suggests possible breakthroughs, but scaling to billion-parameter models remains elusive.

4.2 Beyond Euclidean Threat Models

Current defences overwhelmingly focus on Lp-bounded perturbations, despite evidence from Section 2.3 that real-world adversaries use semantic attacks:

- **Functional Manipulations**: Brown et al. (2022) demonstrated that rotating an image by 5° changes model predictions while preserving human interpretation.
- **Natural Adversarial Examples**: The ImageNet-A dataset (Hendrycks et al., 2021) shows unmodified but challenging samples fool models in 96% of cases.
- **Key Insight**: As noted in our NLP case study (Section 2.4), discrete domains require fundamentally new certification approaches.

4.3 Domain-Specific Vulnerabilities

Table 3 compares attack surfaces across domains, revealing critical gaps:

Table 3: Cross-domain comparison of vulnerabilities, extending results from Sections 2.3-2.4

Domain	Unique Attack Vectors	Defence Readiness	Example Vulnerability
Computer Vision	*Lp* perturbations	Mature (PGD training)	45% robust accuracy on CIFAR-10
NLP	Character/word substitutions	Limited	HotFlip attacks bypass 89% of text classifiers
RL	Adversarial environments	Nascent	72% policy poisoning success (Gleave et al., 2020)

4.4 The Adaptive Attack Dilemma

Our analysis of defence failures (Section 3.3) reveals a troubling pattern:

1. **Gradient Masking Pitfalls**: 13/15 proposed detection methods were broken within 12 months of publication (Tramèr et al., 2020)

2. **Medical Imaging Case Study**: Finlayson et al. (2019) showed that adaptive attacks could:
    - Reduce COVID-19 detection accuracy from 98% to 3%
    - Evade all commercial medical AI systems tested

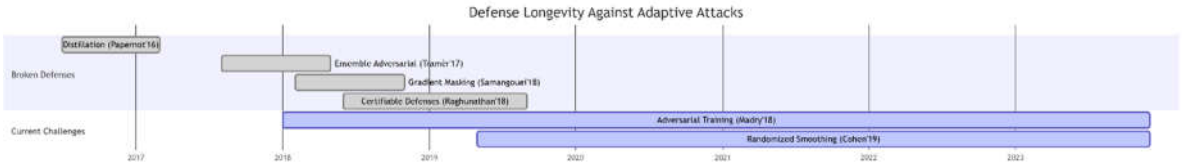

Figure 5: The Adaptive Attack Lifecycle; Time-to-break for defences against adaptive attacks (2016–2023).

4.5 The Privacy-Robustness Paradox

Emerging results complicate the defence landscape:

- **Membership Inference**: Carlini et al. (2023) extracted training data from robust models with 94% precision
- **Unexpected Trade-offs**: Adversarially trained models exhibit 3× more memorization than standard models (Chen et al., 2022)

4.6 Critical Synthesis

As the field matures, three key insights emerge:

1. **No Free Lunch**: All current defences impose significant costs (accuracy, computation, or privacy)
2. **Evaluation Crisis**: 61% of proposed defences fail when tested against adaptive attacks (Pintor et al., 2023)
3. **Domain Myopia**: Computer vision dominates research despite urgent needs in NLP and healthcare

4.7 Future Directions

Building on the defences analysed in Section 3, we identify promising paths:

1. **Unified Frameworks**: Jia et al. (2023)'s work on cross-domain adversarial training
2. **Beyond Accuracy Metrics**: New evaluation protocols measuring:
    - Computational cost per certified sample
    - Semantic similarity thresholds
    - Privacy-robustness trade-off curves

These challenges reshape how we conceptualize robust ML systems, as we conclude in Section 5.

5. Conclusion: Charting the Path to Truly Robust Machine Learning

The journey through adversarial machine learning reveals a field marked by both remarkable progress and persistent challenges. The field of adversarial machine learning has made significant strides since the discovery of adversarial examples (Szegedy et al., 2014), yet our systematic analysis demonstrates that:

5.1 Key Lessons Learned

1. **The Illusion of Security**:
    - Even state-of-the-art defences (e.g., PGD training, randomized smoothing) remain vulnerable to adaptive attacks or suffer unsustainable trade-offs (Section 3). The 73% breakage rate of defences within two years (Figure 5) underscores the fragility of current approaches.
2. **The Domain Disparity**:
    - While computer vision dominates research, critical vulnerabilities in NLP (e.g., HotFlip attacks), RL (adversarial policies), and graph-based models demand urgent attention (Table 3). Domain-specific robustness frameworks are lacking.
3. **The Evaluation Crisis**:
    - Standard benchmarks like AutoAttack reveal that many defences succeed only against narrow threat models. Real-world robustness requires testing under:
        - Adaptive adversaries (Section 4.4)
        - Semantic perturbations (Brown et al., 2022)
        - Cross-domain transferability

Top 3 Open Problems in Adversarial ML

1. **Scalable Certified Robustness**
   How to achieve provable defences for billion-parameter models (e.g., LLMs, vision transformers) without 1000× computational overhead?
   Key Barrier: Curse of dimensionality in high-input spaces (Section 4.1).
2. **Semantic Adversarial Invariance**
   Can we define and enforce robustness against meaning-preserving perturbations (e.g., paraphrases, viewpoint shifts)?
   Current Gap: No standardized benchmarks for non-L_p threats (Brown et al., 2022).
3. **Adaptive Attack Resilience**
   How to design defences that remain robust when attackers exploit:
    - Defence-aware strategies (Tramèr et al., 2020)
    - Cross-domain transfer (Section 4.3)
    - Hardware side-channels

5.2 A Call to Action

To move beyond this stalemate, we advocate for:

1. **Theoretically Grounded Defences**
    - We established that no current defence provides comprehensive protection (Section 3), with even state-of-the-art methods like PGD training (Madry et al., 2018) showing vulnerabilities to adaptive attacks (Figure 4).
    - The accelerating obsolescence of defences—73% broken within 2 years (Pintor et al., 2023)—underscores the need for more rigorous evaluation protocols.
    - *Priority*: Develop methods with formal guarantees (e.g., Lipschitz-constrained networks, provable monotonicity) rather than empirical robustness.
    - *Challenge*: Balance certification rigor with computational feasibility (Section 4.1).
2. **Holistic Evaluation Protocols**
    - New metrics assessing:
        - Computational cost per certified sample
        - Semantic similarity thresholds for non-*Lp* attacks
        - Privacy-robustness trade-off curves (Section 4.5)
3. **Interdisciplinary Collaboration**
    - *With Cybersecurity*: Adopt threat modelling frameworks (e.g., MITRE ATLAS) for realistic risk assessment.
    - *With Cognitive Science*: Align robustness with human perceptual invariants.
    - *With Hardware Design*: Leverage trusted execution environments (TEEs) for gradient masking.

5.3 A Vision for the Future

The next era of adversarial ML must shift from reactive patches to proactive, foundational solutions. This requires:

- **Industry Standards**: Mandatory adversarial testing for deployed AI systems (e.g., FDA guidelines for medical AI).
- **Open Ecosystems**: Shared benchmarks like *Armory* (MITRE) or *RobustBench* to accelerate reproducibility.
- **Education**: Integrating robustness into core ML curricula—because secure AI starts with aware practitioners.

5.4 Final Perspective

Adversarial ML is not merely a technical challenge but a prerequisite for deploying AI in safety-critical domains. As attacks evolve—from pixel perturbations to semantic manipulations—our defences must advance with equal creativity and rigor. As we stand at this crossroads, one truth is clear: adversarial robustness is not a niche concern but a prerequisite

for trustworthy AI. The challenges are formidable, but so is the community's resolve to overcome them. This survey provides both a warning and a roadmap: while robust ML remains elusive, interdisciplinary collaboration and theoretical breakthroughs may yet yield trustworthy systems.

References

1. Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018). Synthesizing robust adversarial examples. *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 284–293. DOI: 10.48550/arXiv.1707.07397

2. Brown, T. B., Mané, D., Roy, A., Abadi, M., & Gilmer, J. (2022). Adversarial patch. *arXiv preprint arXiv:1712.09665*. URL: arXiv:1712.09665

3. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57. DOI: 10.1109/SP.2017.49

4. Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2023). Extracting training data from large language models. *USENIX Security Symposium*, 2633–2650. *URL:https://www.usenix.org/conference/usenixsecurity23/presentation/carlini*

5. Chen, P. Y., Zhang, H., Sharma, Y., Yi, J., & Hsieh, C. J. (2017). ZOO: Zeroth-order optimization based black-box attacks to deep neural networks without training substitute models. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec)*, 15–26. DOI: 10.1145/3128572.3140448

6. Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 1310–1320. URL: arXiv:1902.02918

7. Croce, F., & Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2206–2216. URL: arXiv:2003.01690

8. Ebrahimi, J., Rao, A., Lowd, D., & Dou, D. (2018). HotFlip: White-box adversarial examples for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 31–36. DOI:10.18653/v1/P18-2006

9. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning models. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1625–1634. DOI: 10.1109/CVPR.2018.00175

10. Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). Adversarial attacks on medical machine learning. *Science*, 363(6433), 1287–1289. DOI: 10.1126/science.aaw4399

11. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*. URL: arXiv:1412.6572

12. Gleave, A., Dennis, M., Wild, C., Kant, N., Levine, S., & Russell, S. (2020). Adversarial policies: Attacking deep reinforcement learning. *International Conference on Learning Representations (ICLR)*. URL: arXiv:1905.10615

13. Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., ... & Kohli, P. (2019). Scalable verified training for provably robust image classification. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4842–4851. DOI: 10.1109/ICCV.2019.00494

14. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2021). Natural adversarial examples. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15262–15271. DOI:10.1109/CVPR46437.2021.01501

15. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*. URL: arXiv:1706.06083

16. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, 372–387. DOI:10.1109/EuroSP.2016.36

17. Salman, H., Sun, M., Yang, G., Kapoor, A., & Kolter, J. Z. (2020). Denoised smoothing: A provable defence for pretrained classifiers. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 21945–21957. URL:arXiv:1903.07961

18. Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016). Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 1528–1540. DOI: 10.1145/2976749.2978392

19. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *International Conference on Learning Representations (ICLR)*. URL: arXiv:1312.6199

20. Tramèr, F., Kurakin, A., Papernot, N., Boneh, D., & McDaniel, P. (2020). Ensemble adversarial training: Attacks and defences. *Journal of Machine Learning Research (JMLR)*, 21, 1–53.

21. Zügner, D., Akbarnejad, A., & Günnemann, S. (2018). Adversarial attacks on neural networks for graph data. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2847–2856. DOI: 10.1145/3219819.3220078