Journal of Intelligent Learning Systems and Applications, 2026, 18(2), 167-180
https://www.scirp.org/journal/jilsa
ISSN Online: 2150-8410
ISSN Print: 2150-8402

# Adversarial Machine Learning: Taxonomy, Threat Models, and Mitigation Strategies in Deep Neural Networks

**Taghreed Alqaisi**

Applied College, Taibah University, Madinah, Saudi Arabia
Email: tqesi@taibahu.edu.sa

**How to cite this paper:** Alqaisi, T. (2026) Adversarial Machine Learning: Taxonomy, Threat Models, and Mitigation Strategies in Deep Neural Networks. *Journal of Intelligent Learning Systems and Applications*, **18**, 167-180.
https://doi.org/10.4236/jilsa.2026.182011

**Received:** April 20, 2026
**Accepted:** May 26, 2026
**Published:** May 29, 2026



## Abstract

Machine learning models, and deep neural networks in particular, have achieved remarkable performance across a wide spectrum of tasks from image classification and natural language processing to autonomous navigation and medical diagnostics. However, alongside their successes, these models have exhibited a deep vulnerability: their susceptibility to adversarial examples, carefully crafted inputs designed to deceive a model into producing incorrect outputs. This paper presents a comprehensive survey and taxonomy of adversarial machine learning attacks and defenses, examining both the theoretical underpinnings and practical implications for real-world systems. We systematically classify attacks along four primary dimensions: knowledge of the target model (white-box versus black-box), the attack stage (training-time or inference-time), the attacker's goal (targeted versus untargeted), and the perturbation modality (pixel-level, semantic, or physical). For each category, we review representative algorithms, analyze their strengths and limitations, and discuss the threat models they inhabit. On the defense side, we evaluate adversarial training, certified robustness methods, detection-based approaches, and input preprocessing techniques. We conduct a critical assessment of the arms race dynamic between attackers and defenders and propose a unified evaluation framework for comparing defense mechanisms across attack types. We also discuss open research challenges, including the transferability problem, robustness-accuracy trade-offs, and adversarial robustness in the physical world. Our goal is to provide a rigorous, accessible foundation for researchers and practitioners working at the intersection of machine learning, security, and systems design.

## Keywords

Adversarial, Deep Learning, Robustness, Threat Models, Adversarial Training

## 1. Introduction

The extensive deployment of machine learning technologies across domains raises important questions about their reliability and security. Applications such as facial recognition technology used in airport security, self-driving cars in complex traffic, and natural language processing used to automate loan decisions, and the use of deep learning to assist radiologists in identifying tumors, all assume that the systems being used are robust against both naturally occurring variations and deliberately designed adversarial perturbations. In many of these use cases, the potential consequences of such systems failing are severe. The deployed nature of these systems also raises a complementary concern beyond robustness itself: when failures or attacks do occur, organizations need digital forensic readiness to investigate what happened, a need that is particularly acute for ML-driven IoT and edge environments where evidence is distributed and ephemeral [1].

The first systematic demonstration of neural networks' vulnerability to adversarial examples was conducted by Szegedy *et al.* [2]. They demonstrated that, for example, small perturbations made to images, which are normally imperceptible to human observers, can be made to fool a state-of-the-art image classifier to predict with a high degree of certainty that the perturbed image is some other class. This example is alarming for several reasons, particularly because it highlights the fundamental brittleness of a deep learning model for a specific class. In contrast to human perception systems (which tend to be reasonably robust to small perturbations and variations), neural networks (and neural network classifiers) are designed to be sensitive to specific patterns and perturbations.

In 2015, Goodfellow *et al.* [3] introduced a new dimension to this line of work by arguing that neural networks' adversarial vulnerability stems from their linearity in high-dimensional spaces. They introduced the Fast Gradient Sign Method (FGSM), which asserted that an adversary only needs to compute the gradient once to create a meaningful adversarial example. This suggested that the issue at hand was ubiquitous rather than incidental. Numerous studies have confirmed that this phenomenon is not limited to image classification and is present in models of text, audio, video, graph-based data, and even code.

In adversarial machine learning, the threat landscape from a security perspective can be organized temporally. From the time the model is trained until the time it is deployed, a threat to a model can be organized by when it occurs. During model training, threats can be categorized as poisoning attacks. These attacks are designed to alter a model's behavior by introducing corrupted data into the training set. Once the model is deployed, threats are categorized as evasion attacks. These attacks attempt to fool the model with crafted inputs. The two scenarios present unique threat models and, in turn, different defensive approaches, as well as the need to redefine the boundaries of trust in machine learning systems.

These vulnerabilities have significant practical implications. As Carlini and Wagner [4] showed, the previously suggested defenses have been proven ineffective against stronger, adaptive adversaries. This has been replicated by many so-called

defense mechanisms. Adversarial attacks in the physical world have targeted and succeeded in stopping sign recognition [5], face recognition [6], and object detection [7] systems, showing that the threat is not just digital. At the same time, the randomized smoothing [8] and interval bound propagation [9] methods, which have provided provable robustness for certain certified defenses, are indeed progress, even if not all the way there.

The remaining sections are organized as follows. Section 1 describes the review method used to compile this survey. Section 2 provides a formal framework for the threat model. Section 3 presents the taxonomy of adversarial attacks that we propose. Section 4 discusses defense strategies and their shortcomings. Section 5 looks at the evaluation methodology. Section 6 presents the new problems and those yet to be solved. Finally, Section 7 concludes the paper.

### Review Method

To ensure the survey's comprehensiveness, we adopted a structured review process. Literature was retrieved from major scholarly databases, including IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, and Google Scholar, supplemented by preprints from arXiv (cs.LG, cs.CR, cs.CV) and proceedings of the major venues that publish work on adversarial machine learning, namely NeurIPS, ICML, ICLR, CVPR, IEEE S&P, USENIX Security, and ACM CCS. The search keywords combined the core terms "adversarial example", "adversarial attack", "adversarial robustness", and "adversarial training" with deep-learning qualifiers such as "deep neural network", "image classification", and "large language model". The primary time window covered work published from 2014, the year of the first systematic demonstration of adversarial examples by Szegedy *et al.* [2], to 2024.

Studies were selected for inclusion when they 1) were peer-reviewed papers in top-tier security or machine-learning venues, or widely cited preprints from established research groups, 2) introduced an attack, defense, or evaluation methodology that is now considered foundational or has shaped subsequent work, and 3) reported empirical or theoretical results that materially influenced the taxonomy presented here. We deliberately favored representative and widely replicated studies over exhaustive enumeration. Works that were superseded by stronger follow-up results, or that were shown to rely on obfuscated gradients or non-adaptive evaluation, were retained only when they remained pedagogically useful for explaining the evolution of the field.

## 2. Formal Threat Model Framework

To make the taxonomy easier to apply consistently across the remainder of the paper, we summarize the threat model (**Figure 1**) used in this survey as a compact schema with five components, each of which is expanded in the subsections below:

**1) Adversary knowledge:** White-box, gray-box, or black-box (hard-label/soft-label) access to the target model.

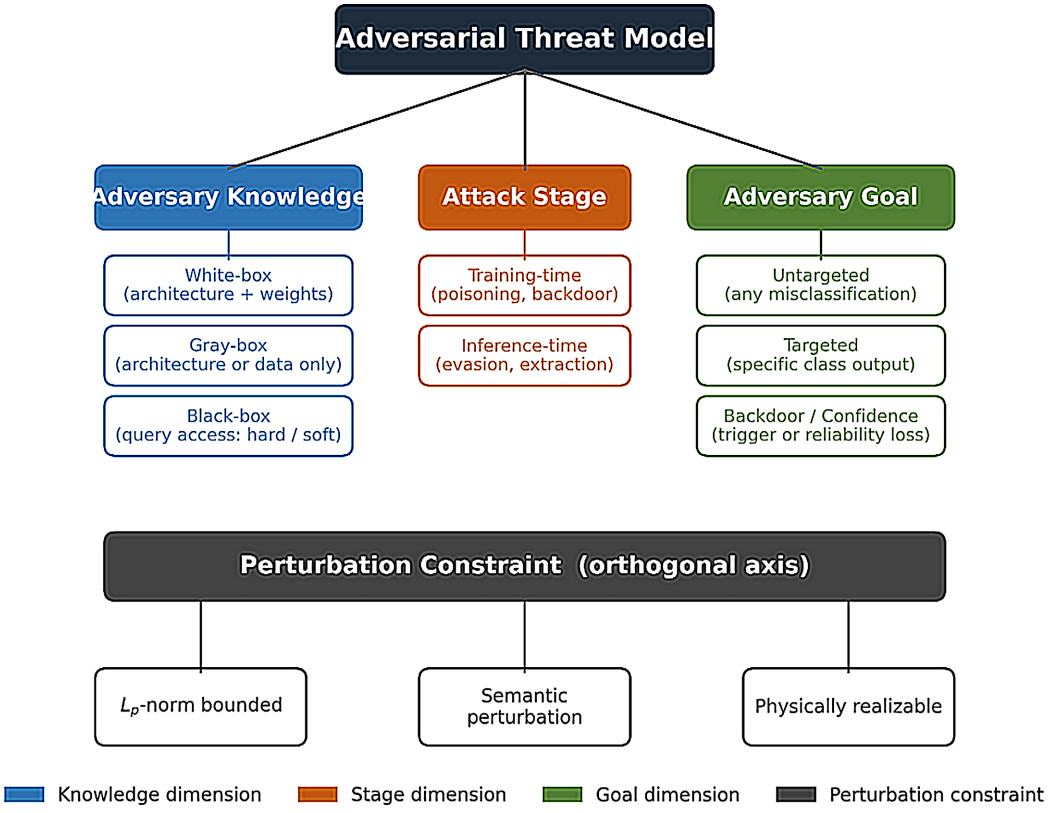


**Figure 1.** Adversarial threat model taxonomy.

**2) Adversary capability:** The operations available to the adversary, such as querying the model, modifying inputs, injecting training samples, or manipulating model weights.

**3) Attack stage:** Training-time (poisoning, backdoor) or inference-time (evasion, query-based).

**4) Adversary goal:** Untargeted misclassification, targeted misclassification, confidence reduction, or backdoor activation.

**5) Perturbation constraint:** Lp-norm bounded perturbation, semantic perturbation, or physically realizable perturbation.

Throughout the rest of this paper, an attack or a defense is described, where appropriate, by instantiating these five components, so that comparisons across categories remain consistent.

## 2.1. Adversary Capabilities and Knowledge

What an adversary knows before they begin an attack, what means they have at their disposal, and what their attack goal needs to be defined with utmost precision when building a thorough threat model. In the domain of adversarial machine learning, adversary knowledge is usually defined in a continuum from white-box to black-box. The white-box attack model is one in which an adversary is fully aware of the target model, including its architecture, parameters, training methodology, and the loss function used during training. This is the most formidable attack model

and is most used to define the highest standards of attack efficacy. In a black-box model, an adversary has access to the target model's predictions, which may be either the predicted label (hard-label black-box) or the predicted label's probability distribution (soft-label black-box), and has the freedom to choose the inputs (*i.e.*, queries) for which the predictions are obtained.

The gray-box model represents a scenario in which an adversary is in a more advantageous position than the black-box model. In a gray-box model, an adversary may know the model architecture, but not the model weights. They may also have access to the model's training data. The gray-box model represents the most realistic scenario, since in most cases, adversaries have access to model cards, API endpoints, or research publications that describe the model family.

### 2.2. Adversary Goals

In adversarial machine learning, adversary goals can be described using two axes: specificity and impact. Untargeted attacks are relatively "easy" in that the adversary's goals are any sort of misclassification. The adversary is successful if the model produces an erroneous output, with no restrictions on the output. Targeted attacks focus on an attacker's output, prompting the model to produce a specific output. Targeted attacks, compared to untargeted attacks, are harder to execute and, in many contexts, more damaging.

Beyond causing misclassification, adversaries also might have more subtle objectives with regards to the model that may include: causing a reduction in confidence of the model with no misclassification (*i.e.*, intentionally degrading the model's reliability as opposed to degrading the accuracy), inducing certain failure modes (e.g., an AI physician helper that always fails to recognize a certain disease), or, more subtly, causing the model to exhibit certain "backdoor" or "sleeper" behaviors that are triggered by the adversary. This heterogeneity of objectives illustrates the need for defenses and evaluations in multiple dimensions.

### 2.3. Perturbation Constraints

Adversarial attacks and the construction of adversarial examples are characterized by a specific type of perturbation constrained by specific constraints. The most common use of constraints, which are imposed by an adversary, is the so-called Lp-norm constraints. Adversarial perturbations that are L∞ (*i.e.*, bounds the maximum change to any single pixel), L2 (*i.e.*, Euclidean distance), or L1 (*i.e.*, sum of absolute differences, *i.e.*, Manhattan/taxicab distance) are the most used in the construction of adversarial examples. The use of these constraints is often a reflection of the model's behaviors of an "acceptable" or "imperceptible" modification to the model, and these norms lead to different model behaviors.

The manner of describing attacks using Lp-norm constraints has shifted toward semantic perturbation models to address a greater portion of the attack-surface realism. Semantic perturbations introduce edits and modifications whose semantics enable attacks, such as altering an object's color, making slight rotations, occluding

the object, or adjusting the lighting in such a way that it remains perceptually seamless to a human observer. Since such edits evade many standard Lp-norm metrics, the attacks become more challenging to defend.

## 3. Taxonomy of Adversarial Attacks

### 3.1. Evasion Attacks at Inference Time

Evasion attacks are the first and most popular research topic in adversarial machine learning. The FGSM [3] forms a perturbation by calculating the gradient of the loss with respect to the input, and then taking a single step that maximally increases the loss in the direction of a defined perturbation budget $\varepsilon$. While FGSM is computationally efficient, the single-step perturbations it produces have been shown to be substantially weaker than iterative attacks under standard $L\infty$ evaluation budgets, and adversarial training against PGD-style adversaries [10] typically reduces FGSM attack success markedly, although the resulting models are not necessarily robust against stronger adaptive attacks [4] [11]. The Projected Gradient Descent (PGD) attack [10] is an iterative extension of FGSM that performs multiple small gradient steps, then projects the perturbation back onto the Lp-norm ball after each step. PGD is considered a strong first-order attack and is the most used attack when evaluating adversarial training.

The Carlini and Wagner (C&W) attack [4] introduced the generation of adversarial examples as an optimization problem, employing a well-defined objective function that attempts to minimize a trade-off between the size of the perturbation and the confidence of the adversarial classification. In the targeted context, the C&W attack is considerably more potent than gradient-based attacks and was pivotal in illustrating the futility of numerous proposed countermeasures, including distillation defenses. The Auto-Attack framework [11] merges several complementary attacks to address the incomplete evaluation problem, which has led to overconfidence in the robustness of proposed defenses, and thus provides a uniform, parameter-free evaluation.

Decision-based black-box attacks [12] [13] operate in the most constrained black-box scenario, where only the final predicted label is available. The Boundary Attack begins with a significant perturbation that successfully deceives the model, then gradually reduces its magnitude while remaining adversarial, akin to performing a random walk along the decision boundary. These attacks show that even with extremely limited knowledge of the target model, adversarial fragility persists.

### 3.2. Poisoning and Backdoor Attacks

Poisoning and backdoor (or Trojan) attacks are examples of training-time attacks against machine learning models. While data poisoning attacks aim to degrade a model's performance, backdoor attacks implant malicious behavior that is triggered by specific inputs. Backdoor attacks were first demonstrated by Chen *et al.* [14] by training a model using a small number of poison images containing a trigger (e.g., a small square of pixels) that caused the model to incorrectly classify any image

containing the trigger into an attacker-specified class, while the model performed correctly on all clean images.

In the context of Machine Learning as a Service (MLaaS), where users customize and tune models via third-party training services with little to no knowledge of the underlying training data, backdoor attacks are a particularly prominent threat. Recent studies show that backdoor attacks can be achieved through poisoning the model architecture [15], manipulating the fine-tuning process of publicly available pre-trained models, and supply-chain attacks on model weights in open-source repositories.

### 3.3. Model Extraction and Inference Attacks

**Scope note.** The attacks discussed in this subsection model extraction, membership inference, and model inversion—target the confidentiality of a deployed model or its training data rather than the integrity of its predictions on a single input. They therefore sit at the boundary between adversarial machine learning and a broader machine-learning privacy and security literature. We include them here because they share the threat-model vocabulary developed in Section 2 (white-box vs. black-box access, query-based capability, inference-time stage), and because real-world deployments must defend against both prediction-integrity and confidentiality threats simultaneously. Readers interested in attacks whose primary goal is to manipulate prediction outcomes should treat Sections 3.1 and 3.2 as the core of the adversarial-ML taxonomy, and Section 3.3 as a closely related privacy/security category that is described here for completeness (Figure 2).

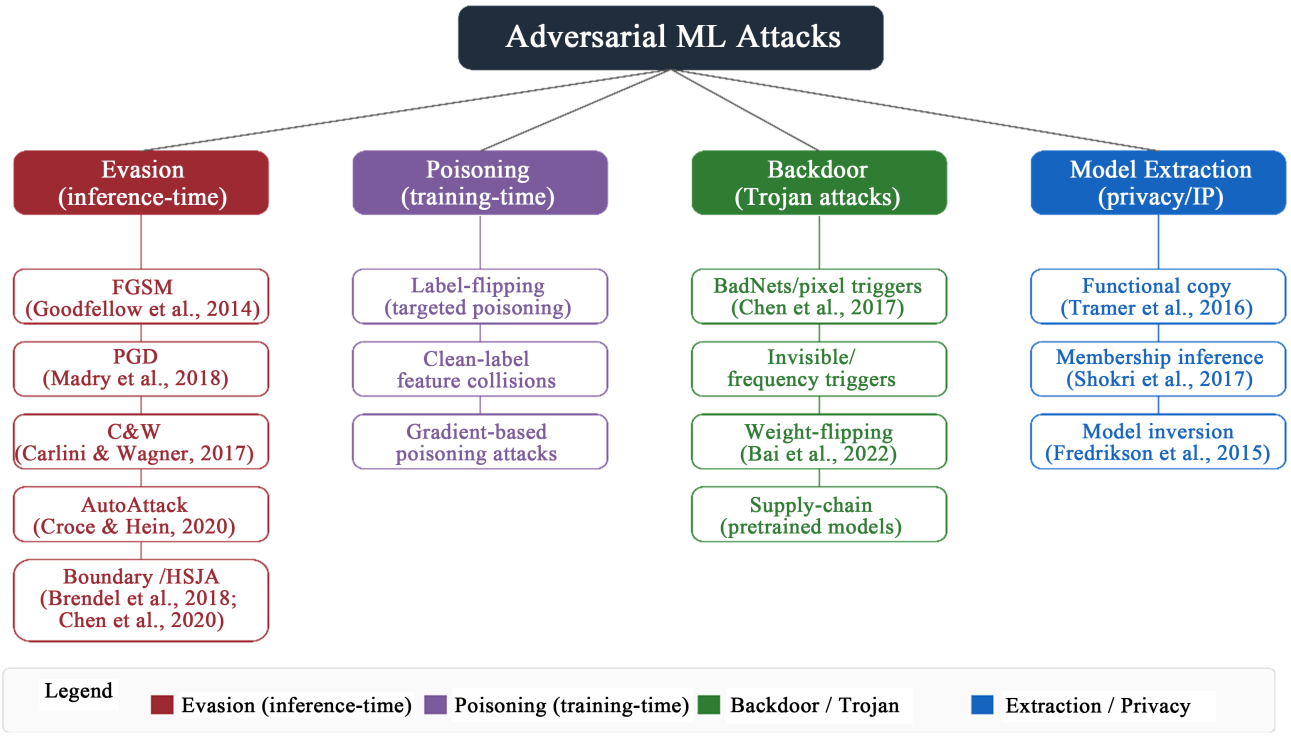


**Figure 2.** Attack taxonomy tree (hierarchical).

In addition to concerns about adversarial attacks on prediction outcomes, attacks against the model itself, as a target of proprietary concerns or as a holder of sensitive information, have arisen. Model extraction attacks [16] use black-box machine learning API query instantiation to build a functionally equivalent copy of the target model, thereby compromising the target model without the original training data. Using a minimal yet well-engineered subset of input queries, the adversary can model the target's behavioral responses with remarkable accuracy, thereby eliminating the significant investment required to develop a bespoke rival model.

Membership inference attacks [17] pose a direct privacy threat to the training dataset, potentially exposing sensitive information by assessing whether a particular data point is part of the model's training corpus. Similarly, model inversion attacks [18] pose a risk by attempting to reconstruct representative examples from the training set by exploiting the model's parameters or output probabilities, thereby recovering sensitive personal data such as facial images and medical records.

## 4. Defense Mechanisms and Their Limitations

### 4.1. Adversarial Training

Adversarial training of Madry *et al.* [10] is currently the most accepted and studied form of defense against adversarial examples. Adversarial training augments the training set with PGD-generated adversarial examples. As a result, the model becomes more robust to such perturbations during training. Adversarial training is a saddle-point training approach. In the defense case, the inner maximization step yields the most adversarial example. The outer minimization step uses defense-focused parameter updates. The result may lead to a model that is more difficult to attack.

Despite the improvements and effectiveness of this training, several shortcomings remain. The first is related to the overall computation time required by the inner optimizations of PGD. The training time is typically a factor of (7 - 10) above that of the standard optimizations. There is also a trade-off between a model's accuracy and its robustness. Attack-theoretic analysis was conducted for adversarial examples, and accuracy was empirically shown to typically decline during model training. Adversarial training lacks the somewhat simpler case of model generalization in a multi-attack environment. For example, a model robust to Lp attacks may not be robust to Lq attacks.

### 4.2. Certified Defenses

While empirical defenses can be defeated by increasingly sophisticated adaptive adversaries, certified defenses aim to provide formal assurances of model robustness. The most scalable certified defense applicable to large neural networks is randomized smoothing [8]. Starting with a classifier f, the smoothed classifier g uses f to classify a noisy version of the input (after adding Gaussian noise) and takes a majority vote after multiple noisy samples. The central result is that the smoothed

classifier possesses a provable Lipschitz continuity, which places a limit on the change in output probabilities resulting from Lp-norm perturbations.

Interval Bound Propagation (IBP) and related abstract interpretation methods [9] [19] take an alternative approach by training models whose predictions can be verified by propagating either symbolic intervals or zonotopes through the layers of the network. Although for small networks these techniques provide a tight bound, they become increasingly conservative (looser) with larger network depth and width, which in turn limits the use of larger, more powerful architectures (Figure 3).

| Defense Method | Clean Accuracy | Robust Accuracy | Formal Guarantee | Compute Cost | Scalability | Attack Coverage |
|---|---|---|---|---|---|---|
| **Adversarial Training** *(PGD-AT, Madry+ 2018)* | ●●●●○ ~85% (CIFAR-10) | ●●●○○ ~56% (ε=8/255) | **Empirical** no formal proof | ●●○○○ 7–10× standard | ●●●●○ any architecture | **$L_p$-norm attacks** trained distribution |
| **Randomized Smoothing** *(RS, Cohen+ 2019)* | ●●●○○ ~75% (acc. loss) | ●●●○○ ~43% certified | **Certified** $L_2$ guarantee | ●●○○○ many fwd passes | ●●●●○ model-agnostic | **$L_2$-norm attacks** Gaussian noise |
| **Interval Bound Propagation** *(IBP, Mirman+ 2018)* | ●●○○○ ~70% small nets | ●●●○○ tight for small ε | **Certified** $L_p$-norm guarantee | ●●●○○ faster than RS | ●●○○○ loose on large nets | **$L_\infty/L_2$ attacks** bounded perturbation |
| **Adversarial Detection** *(e.g., LID, Lee+ 2018)* | ●●●●● no accuracy loss | ●○○○○ broken by adaptive | **Heuristic** no formal proof | ●●●●○ lightweight | ●●●●○ model-agnostic | **non-adaptive only** fails vs adaptive |
| **Input Preprocessing** *(Feature Squeezing, Xu+ 2018)* | ●●●●○ minor acc. loss | ●●○○○ moderate resistance | **Empirical** no formal proof | ●●●●● very low overhead | ●●●●● any model | **weak attacks** bypassed by strong $L_p$ |

*Rating: ●●●●● strongest → ● weakest  Formal guarantee: Certified = provable bound; Empirical = strong but not proven; Heuristic = bypassable by adaptive adversary.*

**Key insight: no single defense dominates across all six dimensions. Adversarial training offers the best empirical robustness; certified methods offer formal guarantees at a cost in clean accuracy and scalability.**

**Figure 3.** Defense strategy comparison.

## 4.3. Detection-Based Approaches

Rather than making the classifier itself robust, detection-based defenses aim to identify adversarial inputs before they reach the classifier and either reject them or route them to additional scrutiny. Feature squeezing [20] applies input transformations such as bit-depth reduction and spatial smoothing, and flags inputs whose classifier outputs change significantly after transformation as potentially adversarial. Mahalanobis distance detection [21] computes the distance between an input's intermediate feature representations and class-conditional Gaussian distributions fitted on clean training data, flagging inputs that are far from the training data manifold.

A fundamental limitation of detection-based defenses is that a sophisticated adversary can often craft adversarial examples that simultaneously fool the classifier

and evade the detector, an attack known as adaptive or dynamic adversarial perturbation. Carlini and Wagner [4] demonstrated this vulnerability for many proposed detectors, and subsequent work has consistently shown that detectors evaluated only against non-adaptive attacks provide an overly optimistic picture of security. Evaluating defenses against adaptive adversaries who are aware of and specifically target the defense mechanism is now considered an essential methodology in the field.

## 5. Evaluation of Methodology and Benchmarking

### 5.1. The Problem of Overestimated Robustness

In adversarial machine learning, overestimation of robustness—where defenses appear effective against weak attacks but fail against stronger, more sophisticated ones—has been a problem plaguing this domain of research. Carlini *et al.* [22] pointed out some of the blunders made in evaluation, including the use of weak attacks as baselines, the failure to evaluate against adaptive adversaries, the use of non-standard metrics, and obfuscated gradients as a defense. Obfuscated gradients, where the so-called defense would make computing the gradients difficult but would not contribute to any improvements in the actual defenses, have been noted and cataloged by Athalye *et al.* [23], who showed that out of the nine defenses that were accepted at ICLR 2018, 7 of them could be defeated by differentiating through the obfuscation.

After these challenges, the community has largely agreed on minimum requirements for evaluating robustness. The AutoAttack benchmark [11] is a pillar of this standard, as it provides a parameter-free evaluation process that comprises 4 attacks (2 black-box) for a more reliable assessment of model robustness. RobustBench [24] is a leaderboard that tracks the latest state-of-the-art robustness on given benchmarks, providing a standard for comparison and discouraging one-off evaluations of specific attack ensembles or configurations.

### 5.2. Benchmark Datasets and Standardized Protocols

CIFAR-10 and ImageNet under L∞ perturbation budgets of $\varepsilon = 8/255$ and $\varepsilon = 4/255$, respectively, have emerged as the primary benchmarks for image classification robustness. While these benchmarks have been valuable for driving progress, they have also been criticized for their narrow focus on a single perturbation type and their distance from real-world threat scenarios. Recent benchmarks have expanded the evaluation scope to include multiple perturbation types simultaneously (union robustness), semantic perturbations, distribution shifts, and robustness to naturally occurring corruptions and perturbations [25].

### 5.3. A Unified Evaluation Checklist

Building on the lessons from Sections 5.1 and 5.2, and consistent with the threat-model schema introduced in Section 2, we propose the following operational checklist for reporting and comparing the robustness of a defense. Any robustness claim

should specify, at a minimum, each of the following items:

**1) Threat access.** State the assumed adversary knowledge (white-box, gray-box, or black-box with hard- or soft-label access) and the assumed adversary capability (query budget, ability to inject training data, ability to modify model weights).

**2) Perturbation budget.** Report the perturbation type (Lp-norm, semantic, or physical) and the exact budget (e.g., $\varepsilon = 8/255$ for $L\infty$ on CIFAR-10).

**3) Clean accuracy.** Report the standard test accuracy of the defended model on unperturbed inputs, so that the robustness-accuracy trade-off is visible.

**4) Robust accuracy.** Report accuracy under a strong, standardized attack such as PGD with sufficient iterations or, preferably, the AutoAttack ensemble [11], which combines white-box and black-box attacks.

**5) Adaptive-attack testing.** Report results against an adaptive adversary that is aware of the defense and explicitly designs attacks to bypass it, following the guidelines of Carlini *et al.* [22]. Defenses that have been evaluated only against fixed, non-adaptive attacks should be reported as such, and any robustness claim should be qualified accordingly.

**6) Computational cost and reproducibility.** Report training and evaluation cost relative to a standardized baseline, and provide code, hyperparameters, and pre-trained weights to support independent verification on RobustBench [24] or similar leaderboards.

Treating these six items as a minimum reporting standard makes defenses comparable across attack types and operationalizes the unified evaluation framework promised earlier in this paper.

## 6. Emerging Challenges and Open Research Problems

### 6.1. Robustness in Large Language Models

Large Language Models (LLMs) are now the prevailing technology for natural language processing. The opportunity for new types of adversarial machine learning is large, given the limited research in this area. Jailbreaking via prompt construction is one example. Adversarial suffixes [26] and other prompt construction techniques showcase the ability of an LLM to be trained to provide responses that are harmful, biased, and/or factually incorrect. The involvement of language models introduces significant difficulty in constructing a prompt, as the example and counterexample will be natural-language statements and statements under the Lp-norm constraint.

Zou *et al.* [26] introduced universal adversarial suffixes, which can be appended to any prompt to induce harmful outputs. This is a serious concern given the rapid global deployment of large language models. Defenses to this technique are limited. Approaches such as response filtering, active monitoring, and Reinforcement Learning from Human Feedback (RLHF) have been employed to reduce the success rate of jailbreaking attempts, and have been reported to be partially effective against many manually crafted prompts; however, recent studies have shown that automatically optimized adversarial suffixes [26] and related prompt-construc-

tion techniques can still bypass these defenses on aligned models, suggesting that current mitigations should be regarded as risk-reducing rather than fully effective against adaptive adversaries. The intersection of adversarial machine learning and AI safety contains some of the most interesting and important open research problems.

### 6.2. Physical-World Adversarial Robustness

The physical world has its own unique adversarial attacks, distinct from digital attacks. Each physical attack is subject to unique defenses in the physical world, namely changes in perspective, distance, light, and print quality. There have been demonstrations of physical attacks in the real world, particularly with road sign recognition, facial verification, and autonomous driving. Brown *et al.* [27] demonstrated a prominent physical adversarial attack (the adversarial patch), which is more concerning than purely digital attacks.

Novel attacks typically inspire novel defenses, which is especially true in this situation. The common approach to defending against digital adversarial attacks is ineffective in this case. The challenge of distributing physical transformations exemplifies this concern. Multi-view training, in which data are collected from different viewpoints, and sensor fusion with more constrained joint adversarial manipulation have both been proposed to address this concern, although they have not been fully validated.

## 7. Conclusion

This paper has presented a comprehensive survey of adversarial machine learning, covering the theoretical foundations of adversarial vulnerability, a systematic taxonomy of attack methods, a critical evaluation of existing defenses, and the open research challenges that remain. The field has made substantial progress since its foundational work: certified defenses now offer provable guarantees for certain settings, adversarial training provides strong empirical robustness on standard benchmarks, and the evaluation methodology has become significantly more rigorous. Nevertheless, important challenges remain. The robustness-accuracy trade-off is not yet fully understood theoretically, and it is unclear whether it is an intrinsic limitation or an artifact of current training algorithms. The generalization of defenses across attack types and modalities is limited. Physical world adversarial robustness remains an active area of research with significant practical stakes. And the emergence of large-scale generative models has introduced entirely new attack surfaces that existing frameworks are ill-equipped to address. We believe that addressing these challenges will require tight integration between the adversarial machine learning community and adjacent fields, including formal verification, statistical learning theory, robust optimization, and human factors research. The security of machine learning systems is ultimately a systems problem, requiring defenses that are robust not only in isolation but in the complex, adaptive, adversarial environments in which deployed models operate.

## Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

## References


[1] Kebande, V.R., Ikuesan, R.A., Karie, N.M., Alawadi, S., Choo, K.R. and Al-Dhaqm, A. (2020) Quantifying the Need for Supervised Machine Learning in Conducting Live Forensic Analysis of Emergent Configurations (ECO) in IoT Environments. *Forensic Science International: Reports*, **2**, Article 100122. https://doi.org/10.1016/j.fsir.2020.100122

[2] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R. (2014) Intriguing Properties of Neural Networks. arXiv: 1312.6199.

[3] Goodfellow, I., Shlens, J. and Szegedy, C. (2015) Explaining and Harnessing Adversarial Examples. arXiv: 1412.6572.

[4] Carlini, N. and Wagner, D. (2017) Towards Evaluating the Robustness of Neural Networks. 2017 *IEEE Symposium on Security and Privacy* (*SP*), San Jose, 22-26 May 2017, 39-57. https://doi.org/10.1109/sp.2017.49

[5] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., *et al.* (2018) Robust Physical-World Attacks on Deep Learning Visual Classification. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 1625-1634. https://doi.org/10.1109/cvpr.2018.00175

[6] Sharif, M., Bhagavatula, S., Bauer, L. and Reiter, M.K. (2016) Accessorize to a Crime. *Proceedings of the* 2016 *ACM SIGSAC Conference on Computer and Communications Security*, Vienna, 24-28 October 2016, 1528-1540. https://doi.org/10.1145/2976749.2978392

[7] Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L. and Yuille, A. (2017) Adversarial Examples for Semantic Segmentation and Object Detection. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 1378-1387. https://doi.org/10.1109/iccv.2017.153

[8] Cohen, J., Rosenfeld, E. and Kolter, J. Z. (2019) Certified Adversarial Robustness via Randomized Smoothing. arXiv: 1902.02918.

[9] Mirman, M., Gehr, T. and Vechev, M. (2018) Differentiable Abstract Interpretation for Provably Robust Neural Networks. *Proceedings of the* 35*th International Conference on Machine Learning*, Stockholm, 2018, 3578-3586.

[10] Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A. (2018) Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv: 1706.06083.

[11] Croce, F. and Hein, M. (2020) Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-Free Attacks. *International Conference on Machine Learning* (*ICML*), Online, 13-18 July 2020 2206-2216.

[12] Brendel, W., Rauber, J. and Bethge, M. (2018) Decision-Based Adversarial Attacks. arXiv: 1712.04248.

[13] Chen, J., Jordan, M.I. and Wainwright, M.J. (2020) Hopskipjumpattack: A Query-Efficient Decision-Based Attack. 2020 *IEEE Symposium on Security and Privacy* (*SP*), San Francisco, 18-21 May 2020, 1277-1294. https://doi.org/10.1109/sp40000.2020.00045

[14] Chen, X., Liu, C., Li, B., Lu, K. and Song, D. (2017) Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. arXiv: 1712.05526.

[15] Bai, J., Wu, B., Zhang, Y., Li, Y., Li, Z. and Xia, S.T. (2021) Targeted Attack against Deep Neural Networks via Flipping Limited Weight Bits. arXiv: 2102.10496.

[16] Tramer, F., Zhang, F., Juels, A., Reiter, M.K. and Ristenpart, T. (2016) Stealing Machine Learning Models via Prediction APIs. *Proceedings of the* 25*th USENIX Security Symposium*, Austin, 10-12 August 2016, 601-618.

[17] Shokri, R., Stronati, M., Song, C. and Shmatikov, V. (2017) Membership Inference Attacks against Machine Learning Models. 2017 *IEEE Symposium on Security and Privacy* (*SP*), San Jose, 22-26 May 2017, 3-18. https://doi.org/10.1109/sp.2017.41

[18] Fredrikson, M., Jha, S. and Ristenpart, T. (2015) Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures. *Proceedings of the* 22*nd ACM SIGSAC Conference on Computer and Communications Security*, Denver, 12-16 October 2015, 1322-1333. https://doi.org/10.1145/2810103.2813677

[19] Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., *et al.* (2019) Scalable Verified Training for Provably Robust Image Classification. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October 2019-2 November 2019, 4841-4850. https://doi.org/10.1109/iccv.2019.00494

[20] Xu, W., Evans, D. and Qi, Y. (2018) Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. *Proceedings* 2018 *Network and Distributed System Security Symposium*, San Diego, 18-21 February 2018, 1-15. https://doi.org/10.14722/ndss.2018.23198

[21] Lee, K., Lee, K., Lee, H. and Shin, J. (2018) A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. arXiv: 1807.03888.

[22] Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A. and Kurakin, A. (2019) On Evaluating Adversarial Robustness. arXiv: 1902.06705.

[23] Athalye, A., Carlini, N. and Wagner, D. (2018) Obfuscated Gradients Give a False Sense of Security. arXiv: 1802.00420.

[24] Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P. and Hein, M. (2021) RobustBench: A Standardized Adversarial Robustness Benchmark. NeurIPS 2021 Datasets and Benchmarks Track.

[25] Hendrycks, D. and Dietterich, T. (2019) Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. arXiv: 1903.12261.

[26] Zou, A., Wang, Z., Kolter, J.Z. and Fredrikson, M. (2023) Universal and Transferable Adversarial Attacks on Aligned Language Models. arXiv:2307.15043.

[27] Brown, T.B., Mane, D., Roy, A., Abadi, M. and Gilmer, J. (2018) Adversarial Patch. arXiv: 1712.09665.