MANTECH Publications

Journal of Next-Gen Threat Intelligence and Cyber Defense
Volume 2, Issue 1, January-April 2026

Adversarial Machine Learning & Defensive Countermeasures

Anamika Rathore[1], Deepak Bisht[2], Sonia Singh[3]

Professor[1], Students[2, 3]

Department of Computer Science Engineering

MIT College of Engineering

Corresponding Author Email: *anamika.rathore683@yahoo.com[1]*

ABSTRACT

Machine learning models are increasingly deployed in critical applications such as healthcare diagnostics, autonomous vehicles, biometric authentication, fraud detection, and financial systems. While these systems show remarkable performance, they are vulnerable to adversarial attacks that manipulate input data or training processes to degrade model performance or cause incorrect decisions. Adversarial Machine Learning (AML) studies these vulnerabilities and proposes defensive mechanisms to make systems more robust and reliable. This paper provides a detailed review of adversarial threats including evasion attacks, poisoning attacks, model extraction, and backdoor attacks. It also discusses defensive countermeasures such as adversarial training, defensive distillation, input preprocessing, robust optimization, and detection-based approaches. Furthermore, emerging research trends including certified robustness and game-theoretic defense models are examined. The study aims to provide a structured overview for researchers and practitioners working toward secure and trustworthy AI systems.

KEYWORDS: *Adversarial attacks, Machine learning security, Evasion attack, Poisoning attack, Model robustness, Defensive countermeasures, Backdoor attacks, Certified defense*

INTRODUCTION

Machine learning (ML) has transformed numerous domains due to its ability to learn patterns from large datasets. However, despite their high accuracy, ML models can be surprisingly

Page 1-16 © MANTECH PUBLICATIONS 2026. All Rights Reserved

fragile. Small, carefully crafted perturbations to input data—often imperceptible to humans—can cause models to misclassify with high confidence. This vulnerability was first systematically demonstrated in deep neural networks and has since been observed across various model architectures.

Adversarial Machine Learning is a research area focused on studying these vulnerabilities and designing methods to counteract them. Unlike traditional cybersecurity threats targeting system infrastructure, adversarial attacks target the model's learning mechanism itself. These attacks may occur during training (poisoning) or during inference (evasion), making the entire ML lifecycle potentially vulnerable.

With AI systems being deployed in safety-critical applications, understanding adversarial threats is no longer optional. If an autonomous vehicle misinterprets a stop sign due to adversarial noise, or a medical diagnostic tool is fooled into misclassifying a tumor, the consequences can be severe. Therefore, robust and secure ML models are essential for real-world deployment.

BACKGROUND OF ADVERSARIAL MACHINE LEARNING

Adversarial Machine Learning (AML) emerges from the convergence of machine learning, optimization theory, and cybersecurity. Traditional machine learning assumes that training and testing data follow similar distributions and are not intentionally manipulated. However, in adversarial environments this assumption no longer holds. Attackers may intentionally craft inputs or modify training data to exploit weaknesses in learning algorithms. As ML systems become widely deployed in security-sensitive and economically valuable applications, they become attractive targets for malicious actors.

The vulnerability of modern ML models, particularly deep neural networks, was highlighted when researchers demonstrated that adding very small, carefully designed perturbations to input samples can lead to incorrect predictions with high confidence. These perturbations are often imperceptible to human observers but significantly affect the model's internal decision boundaries. This reveals that many high-performing models rely on fragile patterns that are not always semantically meaningful.

Page 1-16 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Mathematically, an adversarial example can be expressed as:
x′=x+δx' = x + \deltax′=x+δ

Where:

- xxx represents the original clean input,
- δ\deltaδ is a small but intentionally crafted perturbation,
- x′x'x′ is the adversarial input that leads to misclassification.

The perturbation δ\deltaδ is typically constrained under a norm bound such as:
||δ||p≤ϵ||\delta||_p \leq \epsilon||δ||p≤ϵ
Where ppp may be L0L_0L0, L2L_2L2, or L∞L_{\infty}L∞ norm, and ϵ\epsilonϵ controls the maximum allowed perturbation magnitude. Even when ϵ\epsilonϵ is very small, optimization techniques are used to compute δ\deltaδ in a direction that maximizes model loss:
δ=arg⁡max⁡δJ(θ,x+δ,y)\delta = \arg\max_{\delta} J(\theta, x+\delta, y)δ=argδmax J(θ,x+δ,y)

Here, JJJ denotes the loss function and θ\thetaθ represents model parameters.

This formulation shows that adversarial example generation is essentially an optimization problem. The attacker searches for a small perturbation that causes maximum damage to prediction accuracy. From a defender's perspective, robustness can also be formulated as a min–max optimization problem, where the goal is to minimize the worst-case loss under bounded perturbations.

1. Evolution of the Field

Initially, adversarial attacks were studied mainly in image classification tasks. However, research has expanded to other domains including:

- Natural language processing (text perturbation attacks),
- Speech recognition (audio adversarial examples),
- Reinforcement learning systems,
- Graph neural networks,
- Malware detection systems.

Page 1-16 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Physical-world adversarial attacks have also been demonstrated, where perturbations printed on objects or embedded in physical environments can mislead real-world AI systems. This shows that AML is not limited to digital settings only.

2. Threat Model

A well-defined threat model is essential in adversarial machine learning research. Without clearly specifying attacker capabilities and goals, evaluating defenses becomes meaningless.

The threat model describes:

1. The attacker's knowledge,
2. The attacker's capability,
3. The attacker's objective.

a) Attacker Knowledge

White-box attacks:

In this scenario, the attacker has complete access to the model, including architecture, parameters, gradients, and training data distribution. This is the strongest adversarial assumption and is often used to evaluate worst-case robustness. Gradient-based attacks such as FGSM and PGD are typically white-box attacks.

Black-box attacks:

Here, the attacker does not know internal model details. They can only query the model and observe outputs (labels or confidence scores). Despite limited information, attackers can approximate gradients through query-based methods or exploit transferability, where adversarial examples generated on one model also fool another.

Gray-box attacks:

In gray-box settings, partial knowledge is available. For example, the attacker may know the training dataset but not the exact parameters. This scenario often represents practical real-world attacks.

b) Attacker Capability

Capability refers to what the attacker can modify:

Page 1-16 © MANTECH PUBLICATIONS 2026. All Rights Reserved

- Training data manipulation (Poisoning)
- Test-time input perturbation (Evasion)
- Model parameter tampering
- API query access for extraction

Capability constraints determine the feasibility of attack execution.

c) Attacker Goals

Adversarial objectives can be categorized into three primary security dimensions:

Integrity Violation

The attacker aims to cause specific misclassification without disrupting overall system functionality. For example:

- Making malware appear benign,
- Causing a stop sign to be classified as speed limit sign.

Integrity attacks are often targeted attacks.

Availability Attack

The goal is to degrade overall system performance. This may involve:

- Causing widespread misclassification,
- Forcing denial-of-service through excessive queries.

Availability attacks attempt to make the system unreliable or unusable.

Confidentiality Breach

Here the attacker tries to extract sensitive information from the model, such as:

- Reconstructing private training data (model inversion),
- Rebuilding proprietary models (model extraction),
- Membership inference (determining whether a specific data sample was part of training).

These attacks raise serious privacy and intellectual property concerns.

3. Transferability Property

One important phenomenon in adversarial ML is transferability. Adversarial examples crafted for one model often fool another model even with different architectures. This property enables

Page 1-16 © MANTECH PUBLICATIONS 2026. All Rights Reserved

black-box attacks, because an attacker can train a substitute model locally and generate adversarial samples that transfer to the target model.

Transferability suggests that vulnerabilities are not just implementation flaws but may be inherent to high-dimensional decision boundaries in modern ML models.

4. Robustness vs Accuracy Trade-off

Research shows there is often a trade-off between robustness and standard accuracy. Models trained to resist adversarial perturbations sometimes experience slight reductions in clean-data accuracy. Understanding this trade-off remains an active research area.

From a geometric viewpoint, robust models tend to learn smoother decision boundaries, whereas highly optimized but non-robust models may create sharp and complex boundaries that are easy to exploit.

5. Security Evaluation Metrics

To measure robustness, several evaluation metrics are used:

- **Robust accuracy:** Accuracy under adversarial attack.
- **Attack success rate:** Percentage of successful adversarial manipulations.
- **Perturbation norm magnitude:** Strength of adversarial noise.
- **Certified robustness radius:** Guaranteed safe perturbation bound.

Proper evaluation must consider adaptive attacks, because weak evaluation can give a false sense of security.

TYPES OF ADVERSARIAL ATTACKS

Adversarial attacks can be broadly categorized based on **when** they occur in the machine learning pipeline and **how** the attacker manipulates the system. Some attacks target the training phase, while others exploit vulnerabilities during inference. Additionally, certain attacks focus on extracting confidential information rather than simply degrading model accuracy.

Understanding these categories is important because defensive strategies often depend on the stage at which the attack occurs.

Page 1-16 © MANTECH PUBLICATIONS 2026. All Rights Reserved

1. Evasion Attacks (Inference-Time Attacks)

Evasion attacks occur during the deployment or inference phase. In this scenario, the model has already been trained, and the attacker modifies input samples at test time to induce incorrect predictions. Importantly, evasion attacks do not alter the training dataset or model parameters. Instead, they exploit weaknesses in the learned decision boundary.

These attacks are particularly effective against deep neural networks because such models operate in high-dimensional input spaces where small perturbations can significantly influence classification outcomes.

Fast Gradient Sign Method (FGSM)

The Fast Gradient Sign Method is one of the earliest and simplest adversarial attack techniques. It uses the gradient of the loss function with respect to the input to compute perturbations:
$x' = x + \epsilon \cdot sign(\nabla_x J(\theta, x, y))$

Where:

- x is the original input,
- ϵ controls perturbation magnitude,
- $J(\theta, x, y)$ is the loss function,
- ∇_x represents gradient with respect to input.

FGSM is computationally efficient because it requires only one gradient computation. However, since it performs a single-step update, it may be less powerful than iterative methods.

Projected Gradient Descent (PGD)

Projected Gradient Descent is an iterative extension of FGSM. Instead of applying one perturbation step, PGD applies multiple small steps and projects the perturbed sample back into the allowed perturbation region after each step.

Mathematically:
$x_{t+1} = \Pi_{\epsilon} \left(x_t + \alpha \cdot sign(\nabla_x J(\theta, x_t, y)) \right)$

Page 1-16 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Where:

- Πϵ\Pi_{\epsilon}Πϵ ensures perturbations remain within the constraint bound,
- α\alphaα is step size.

PGD is often considered a strong first-order adversary and is widely used for evaluating model robustness.

Carlini & Wagner (C&W) Attack

The C&W attack formulates adversarial generation as an optimization problem with carefully designed objective functions. It minimizes perturbation magnitude while ensuring misclassification.

The general form is:

min⁡δ||δ||p+c·f(x+δ)\min_{\delta} ||\delta||_p + c \cdot f(x+\delta)δmin||δ||p+c·f(x+δ)

This attack is powerful and capable of bypassing several early defense mechanisms, including defensive distillation.

Targeted vs Untargeted Evasion

- **Untargeted attack:** Cause misclassification into any incorrect class.
- **Targeted attack:** Force prediction into a specific target class.

Targeted attacks are generally harder but more dangerous in critical applications.

Real-World Evasion Examples

- Slightly modified traffic signs fooling autonomous vehicles.
- Perturbed audio commands misinterpreted by voice assistants.
- Adversarial patches placed on objects to mislead surveillance systems.

These examples show that evasion attacks are not only theoretical but also practical.

2. Poisoning Attacks (Training-Time Attacks)

Poisoning attacks occur during the training phase. In this case, the attacker manipulates the training dataset so that the model learns incorrect patterns. These attacks are especially dangerous in systems that rely on crowd-sourced or continuously updated datasets.

The attacker's objective may be:

Page 1-16 © MANTECH PUBLICATIONS 2026. All Rights Reserved

- To degrade overall model performance,
- To implant specific vulnerabilities,
- To bias predictions against certain classes.

Label Flipping

Label flipping is one of the simplest poisoning techniques. The attacker changes labels of selected training samples. For example, malicious emails may be labeled as legitimate, causing the spam classifier to learn wrong associations.

Even small percentages of flipped labels can significantly reduce accuracy, especially in sensitive classification tasks.

Gradient Manipulation

In distributed or federated learning systems, attackers may manipulate gradient updates before they are aggregated. By sending malicious gradient values, attackers can influence global model parameters.

This is particularly relevant in federated learning environments where training happens across decentralized devices.

Data Injection Attacks

Attackers may inject specially crafted malicious samples into the training set. These samples are optimized to influence decision boundaries in a targeted way.

For instance:

- Injecting malicious data to cause consistent misclassification of one particular class.
- Inserting outlier samples to distort clustering algorithms.

Impact of Poisoning Attacks

Poisoning attacks can:

- Reduce model accuracy significantly,
- Create targeted vulnerabilities,
- Introduce systematic bias,

Page 1-16 © MANTECH PUBLICATIONS 2026. All Rights Reserved

- Compromise fairness.

Unlike evasion attacks, poisoning affects the model globally and can persist long after training.

3. Backdoor Attacks

Backdoor attacks (also called Trojan attacks) are a specialized form of poisoning. In this case, the attacker inserts a hidden trigger into training samples. The model learns to associate the trigger pattern with a specific label.

During normal evaluation, the model performs well on clean data. However, when the trigger is present, it consistently produces attacker-chosen outputs.

Working Mechanism

1. Add trigger pattern to subset of training samples.
2. Change their labels to attacker's target class.
3. Train model normally.
4. At inference, attach trigger to input → model misclassifies.

Example:

- A small sticker placed on a stop sign causes it to be classified as a speed limit sign.

Why Backdoor Attacks Are Dangerous

- Hard to detect using standard accuracy metrics.
- Remain dormant until triggered.
- Can be implanted during outsourced model training.

Backdoor attacks are particularly concerning in supply-chain ML scenarios, where pre-trained models are downloaded from external sources.

4. Model Extraction and Inversion Attacks

Not all adversarial attacks focus on accuracy degradation. Some aim at confidentiality breaches.

Page 1-16 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Model Extraction Attacks

In model extraction attacks, the attacker repeatedly queries a deployed model and uses input-output pairs to train a surrogate model that approximates the original one.

Steps typically include:

1. Query model with carefully chosen inputs.
2. Record outputs (labels or probability scores).
3. Train local model using collected data.

Consequences include:

- Intellectual property theft,
- Financial losses,
- Increased vulnerability to further attacks.

These attacks are common in ML-as-a-Service APIs.

Model Inversion Attacks

Model inversion aims to reconstruct sensitive training data from model outputs.

For example:

- Recovering face images used to train a recognition system.
- Reconstructing medical records.

In some cases, attackers exploit confidence scores to infer sensitive attributes of individuals.

Membership Inference

A related attack is membership inference, where attackers determine whether a specific sample was used during training. This can leak private information in sensitive domains such as healthcare.

Table 1: Comparison of Major Adversarial Attacks

Attack Type	Stage	Knowledge Required	Impact	Example Method
Evasion	Inference	White/Black box	Misclassification	FGSM, PGD
Poisoning	Training	Partial access	Model corruption	Label flipping

Page 1-16 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Attack Type	Stage	Knowledge Required	Impact	Example Method
Backdoor	Training	Data injection	Triggered errors	Trojan attack
Model Extraction	Inference	Black box	IP theft	Query-based
Inversion	Inference	Black box	Privacy leakage	Reconstruction

DEFENSIVE COUNTERMEASURES

Defense strategies aim to increase robustness without degrading performance.

1. Adversarial Training

Adversarial training augments the training dataset with adversarial examples.

$\min_{\theta} \mathbb{E}_{(x,y)} \left[\max_{\delta} J(\theta, x+\delta, y) \right]$

It improves robustness but increases computational cost.

Advantages:

- Effective against known attacks

Limitations:

- May not generalize to unseen attacks
- Computationally expensive

2. Defensive Distillation

Defensive distillation trains a secondary model using softened probabilities from the first model.

Goal: Reduce sensitivity to input perturbations.

However, later studies showed that strong attacks can still bypass distillation.

3. Input Transformation Techniques

These include:

- Feature squeezing
- Image compression
- Randomization
- Noise filtering

Page 1-16 © MANTECH PUBLICATIONS 2026. All Rights Reserved

While simple to implement, adaptive attackers may overcome them.

4. Detection-Based Approaches

Instead of preventing attacks, these methods detect adversarial inputs.

Techniques include:

- Statistical anomaly detection
- Confidence thresholding
- Neural fingerprinting

Detection systems often suffer from high false positive rates.

5. Robust Optimization

Robust optimization formalizes adversarial defense as a min-max optimization problem.

This theoretical framework provides stronger guarantees but is computationally heavy.

6. Certified Defenses

Certified defenses provide mathematical proof that a model is robust within a certain perturbation bound.

Methods include:

- Randomized smoothing
- Interval bound propagation
- Lipschitz constraints

Certified defenses are promising but limited to small perturbation radii.

Table 2: Comparison of Defensive Techniques

Defense Method	Robustness Level	Computational Cost	Works Against Adaptive Attack?
Adversarial Training	High	High	Partially
Distillation	Medium	Medium	Weak against strong attacks
Input Transformation	Low-Medium	Low	Often bypassed
Detection Systems	Medium	Medium	Depends on threshold

Page 1-16 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Defense Method	Robustness Level	Computational Cost	Works Against Adaptive Attack?
Certified Defense	High (bounded)	Very High	Strong within bounds

ADVERSARIAL ATTACKS IN REAL-WORLD APPLICATIONS

1. Autonomous Vehicles

Perturbed traffic signs can cause incorrect detection. Physical-world adversarial examples are particularly challenging.

2. Healthcare Systems

Medical image classifiers can be manipulated to misdiagnose conditions.

3. Biometric Authentication

Face recognition systems are vulnerable to adversarial glasses or patterns.

4. Financial Fraud Detection

Attackers may craft transactions that evade fraud detection algorithms.

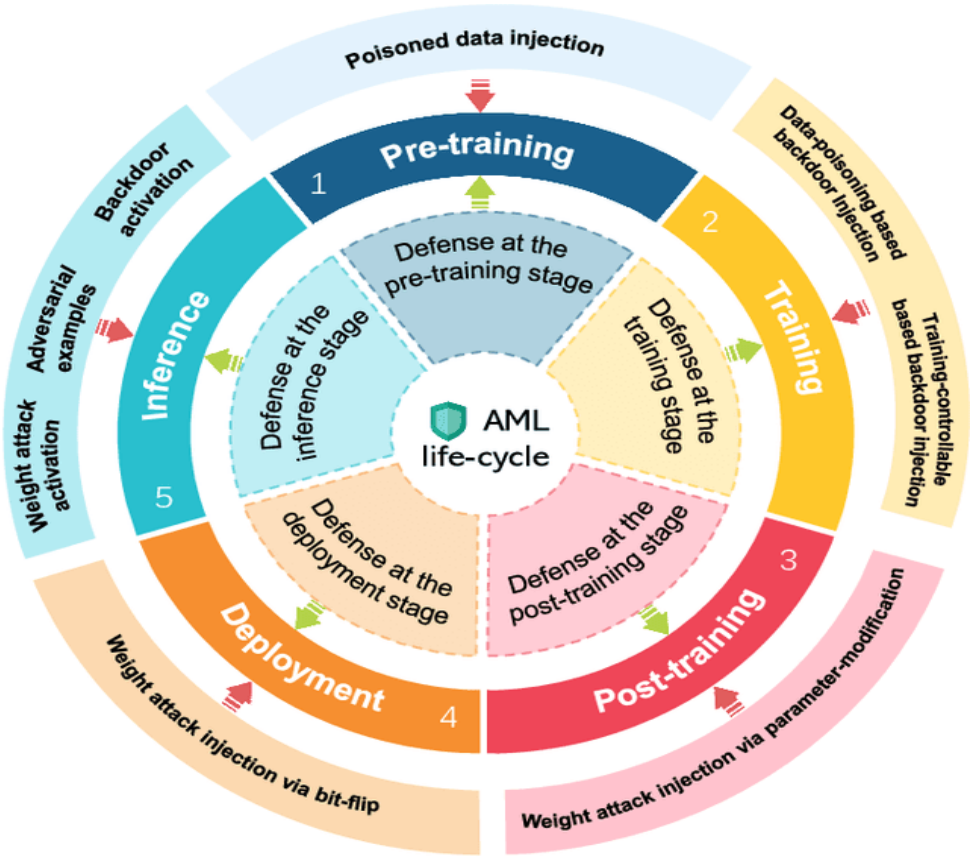


Figure 1: Adversarial ML Lifecycle

Page 1-16 © MANTECH PUBLICATIONS 2026. All Rights Reserved

6. Game-Theoretic Perspective

Adversarial ML can be modeled as a game between attacker and defender.

- Attacker: Maximizes loss
- Defender: Minimizes worst-case risk

Game theory provides equilibrium strategies but practical implementation remains complex.

EMERGING RESEARCH DIRECTIONS

1. Self-Healing Models

Models that adapt after detecting adversarial inputs.

2. Explainable AI for Defense

Explainability tools help identify unusual decision boundaries.

3. Federated Learning Security

AML threats in decentralized learning environments are growing.

4. Hardware-Level Protection

Secure enclaves and trusted execution environments.

CHALLENGES AND OPEN PROBLEMS

- Balancing robustness and accuracy
- Scalability of certified defenses
- Defense against unknown attacks
- Standardized robustness benchmarks
- Energy-efficient robust training

It is still difficult to create a model that is fully robust without sacrificing performance.

CONCLUSION

Adversarial Machine Learning presents a serious challenge to modern AI systems. From small perturbations in images to large-scale data poisoning, adversarial attacks threaten the reliability and trustworthiness of ML models. While several defensive countermeasures have been proposed, none provide complete security. Adversarial training and certified defenses show promising results but come with computational overhead.

Page 1-16 © MANTECH PUBLICATIONS 2026. All Rights Reserved

Future research must focus on adaptive and generalized defense strategies that work across diverse attack types. Integration of explainability, formal verification, and secure hardware may provide holistic solutions. As AI systems continue to expand into critical domains, ensuring robustness against adversarial threats is essential for safe deployment.

The field remains evolving, and collaboration between security researchers and ML engineers will be necessary to develop resilient AI systems.

REFERENCES

1. Szegedy, C., et al. (2014). Intriguing properties of neural networks. ICLR.
2. Goodfellow, I., et al. (2015). Explaining and harnessing adversarial examples. ICLR.
3. Madry, A., et al. (2018). Towards deep learning models resistant to adversarial attacks. ICLR.
4. Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after adversarial ML. Pattern Recognition.
5. Carlini, N., & Wagner, D. (2017). Towards evaluating robustness of neural networks. IEEE S&P.
6. Gu, T., et al. (2017). BadNets: Backdooring deep neural networks. arXiv.
7. Papernot, N., et al. (2016). Distillation as a defense to adversarial perturbations. IEEE S&P.
8. Tramèr, F., et al. (2016). Stealing machine learning models via prediction APIs. USENIX.
9. Cohen, J., et al. (2019). Certified adversarial robustness via randomized smoothing. ICML.
10. Yuan, X., et al. (2019). Adversarial examples: Attacks and defenses. IEEE TNNLS.

Page 1-16 © MANTECH PUBLICATIONS 2026. All Rights Reserved