Volume 11, Issue 6, June – 2026 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1744

A Comprehensive Survey of Adversarial Attacks and Defense Strategies in Artificial Intelligence Security

P. Mahalakshmi[1]; Dr. V. Jayalakshmi[2]

[1]Ph,D., Research Scholar,
School of Computing Science, VISTAS.
Chennai, Tamilnadu, India -600117

[2]Professor,
School of Computing Science, VISTAS.
Chennai, Tamilnadu, India -600117

Publication Date: 2026/07/04

Abstract: The growing deployment of deep learning models in safety-critical domains has exposed the artificial intelligence landscape to a widening array of adversarial threats, where imperceptible input perturbations reliably induce severe misclassifications. This survey provides a comprehensive synthesis of adversarial attacks and defense mechanisms in modern AI security. It introduces a structured taxonomy categorizing attacks into evasion, poisoning, and model inversion strategies, evaluated across varying levels of attacker knowledge. Correspondingly, current defense techniques—including adversarial training, anomaly detection, and gradient masking—are critically reviewed for their resilience against adaptive, real-world adversaries. The survey further examines robustness benchmarking and success rate analysis frameworks, emphasizing the gap between theoretical guarantees and practical deployment. By consolidating recent advances and persistent limitations, this work identifies open research challenges and outlines emerging directions toward provably secure and trustworthy AI systems for real-world applications.

Keywords: *Adversarial Machine Learning; Evasion Attacks; Poisoning Attacks; Model Inversion; Adversarial Training; Gradient Masking; Robustness Benchmarking; Defense-In-Depth; Trustworthy AI; Deep Learning Security.*

How to Cite: P. Mahalakshmi; Dr. V. Jayalakshmi (2026) A Comprehensive Survey of Adversarial Attacks and Defense Strategies in Artificial Intelligence Security. *International Journal of Innovative Science and Research Technology*, 11(6), 2458-2463. https://doi.org/10.38124/ijisrt/26jun1744

I. INTRODUCTION

As machine learning increasingly powers critical applications like autonomous vehicles and financial monitoring, developers must urgently transcend a narrow focus on predictive accuracy to prioritize the development of fundamentally secure and resilient systems. This repositioning of research focus directly confronts the structural weaknesses inherent in deep learning architectures, where even slight, maliciously engineered inputs can trigger severe operational failures. Such vulnerabilities are particularly pernicious because these modifications remain invisible to human observers, yet they exploit the complexities of high-dimensional feature spaces to compel incorrect classifications. Accordingly, the quest for model resilience has crystallized as a critical research imperative, seeking to bridge the widening gap between the intrinsic fragility of neural networks and the stringent safety demands of practical deployments.

This pursuit demands an interdisciplinary methodology that unites foundational principles from cybersecurity with rigorous empirical testing of model resilience, accounting for adversaries operating under both full architectural access and limited-query conditions. Through a systematic organization of these vulnerabilities, this survey maps the adversarial machine learning landscape, delineating boundaries between anticipatory defensive strategies designed to pre-empt exploits and responsive detection mechanisms that identify threats as they emerge. Central to this analysis is the persistent and troubling finding that empirical evidence consistently demonstrates no single defensive mechanism can withstand the full spectrum of adversarial tactics, with proposed safeguards routinely circumvented once subjected

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1744

to adaptive adversaries explicitly tailored against them (Tramèr et al., 2020; Wong & Kolter, 2017). The resulting arms race has repeatedly shown that defenses once regarded as robust, including distillation and gradient masking techniques, often lose their protective value shortly after publication as new attack formulations emerge to violate their underlying assumptions (Olowononi et al., 2020; Wong & Kolter, 2017). Such recurring failures definitively establish that adversarial robustness cannot be treated as a one-time engineering objective but must instead be cultivated as a continuously evolving property, demanding layered combinations of anticipatory hardening, runtime detection, and provably grounded guarantees in order to remain meaningful under realistic threat conditions (Olowononi et al., 2020; Tramèr et al., 2020).

II. RESEARCH BACKGROUND

The rapid integration of deep learning into critical infrastructures has outpaced the development of necessary security frameworks, resulting in meaningful deficiencies in system dependability. Deep neural networks can be systematically fooled by inputs that appear normal yet drive models toward confidently incorrect outputs. These weaknesses necessitate a shift toward evaluating models based on adversarial resilience rather than conventional accuracy, prioritizing standardized robustness benchmarks and automated verification.

Open-source libraries and datasets empower practitioners to empirically test defenses against well-known threats such as the Fast Gradient Sign Method and Projected Gradient Descent, with assessments revealing stark performance gaps: misclassification rates climb from 42.2% under basic gradient-based attacks to 86.8% with advanced optimization strategies.

Recent research has therefore differentiated between two complementary paradigms: robust classification mechanisms that incorporate adversarial examples directly into training (such as adversarial training via PGD and ensemble-based training across multiple teacher models) and auxiliary anomaly detection systems that flag statistical deviations from clean data distributions via binary classifiers, outlier analysis, and feature-level denoising layers. Researchers increasingly conclude that neither paradigm alone is sufficient, advocating layered combinations—such as coupling an adversarial detector with an input reconstructor—to address blind-spot vulnerabilities.

Contemporary literature accordingly frames protection as a defense-in-depth posture mapping mitigations to relevant lifecycle stages, complemented by adaptive response mechanisms capable of reconfiguring defenses against newly identified attack vectors. This layered view reflects the acknowledgment that adversarial robustness cannot be fully established through static benchmarks alone and must be continuously re-evaluated. Establishing such an integrated framing is a necessary prerequisite for moving beyond reactive mitigations toward proactive architectures that anticipate adversarial exploitation before deployment to safety-critical environments.

➢ *Statement of the Problem.*

The core challenge lies in the discrepancy between the rapid adoption of machine learning in critical systems and the lack of consistent, standardized protocols for evaluating their resistance to adversarial attacks.

- This absence of a unified benchmarking framework complicates the comparative analysis of defensive efficacy, often leaving practitioners to rely on empirical mitigations with limited theoretical guarantees (Vassilev et al., 2024).
- Furthermore, the absence of universally accepted evaluation methods impedes the structured detection of potential security weaknesses across various operational settings, thereby exposing organizations to increasingly advanced and dynamic threats.
- Therefore, this study seeks to address these limitations by:
- Providing a thorough classification of adversarial attack surfaces.
- Systematically assessing the performance of current defensive measures against both established and newly emerging threats.
- Highlighting readily available resources and precise evaluation metrics that are crucial for verifying the resilience of machine learning models in intricate, real-world scenarios.
- By mapping mitigation strategies to specific threat vectors, this work enables a transition from reactive patching to proactive, robust architectures through:
- The creation of evaluation protocols that integrate domain-specific constraints.
- Incorporating operational safety requirements or data privacy sensitivities.
- Ensuring that model resilience is rigorously measured against realistic, threat-informed benchmarks.

➢ *Taxonomy of Attacks*

- Adversarial threats are organized according to three primary axes: the attacker's familiarity with the target system, the specific phases where exploitation is attempted, and the resulting influence on model operations.
- This structured classification illustrates the broad spectrum of potential security risks, underscoring the necessity for a comprehensive defense strategy to effectively address these diverse threats.
- The taxonomy adopts a multidimensional approach that integrates principles from signal processing, cryptography, and machine learning to categorize vulnerabilities.
- This approach differentiates between white-box, gray-box, and black-box access scenarios, highlighting how varying levels of insight into a model's architecture and training data dictate an adversary's capacity to perform evasion, poisoning, or extraction attacks.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1744

- The framework further incorporates the specific motivations and resources available to adversaries, allowing researchers to define realistic threat models that are critical for engineering precise security measures.
- The taxonomy differentiates between deliberate efforts aimed at manipulating specific inputs—such as adversarial evasion designed to force misclassification—and broader strategies intended to compromise the system's foundational operational integrity, such as poisoning the training pipeline or extracting sensitive model internals.
- By mapping these distinct threat classes against the target's lifecycle, the framework provides the necessary visibility for practitioners to evaluate their defenses against both localized perturbations and systemic structural weaknesses.
- Ultimately, this comprehensive categorization serves as a foundational roadmap for identifying and mitigating the diverse attack vectors that threaten the reliability, safety, and confidentiality of contemporary machine learning deployments.

➢ *Evasion Attacks*

Evasion attacks occur during the inference phase, involving the strategic introduction of often imperceptible perturbations to input data that induce incorrect model predictions. These methods range from foundational constrained optimization techniques such as L-BFGS to efficient single-step approaches including the Fast Gradient Sign Method and its iterative variants, all operating by maximizing loss on perturbed inputs within defined distance norms to preserve data plausibility. Algorithms such as Projected Gradient Descent and Carlini–Wagner further refine this paradigm by optimizing differentiable surrogate losses to achieve high-confidence misclassifications. By exploiting transferability, adversaries can frequently circumvent the need for direct gradient access to perform black-box evasion, selecting between targeted interventions that force specific misclassifications and untargeted efforts that induce any system failure, depending on the desired impact.

➢ *Poisoning Attacks*

During training, poisoning attacks involve the injection of compromised examples into the training data to deliberately undermine the model's learning process. By contaminating a subset of training samples, adversaries can embed dormant backdoors that become active only when triggered by specific input patterns at inference time. Because these corrupted entries remain statistically indistinguishable from legitimate data, they bypass standard validation routines and gradually erode the integrity of the classifier's learned decision boundaries. This degradation enables attackers to selectively redirect outputs toward predetermined targets while evading routine anomaly detection. Since corrupted gradient contributions persist within trained parameters long after contamination, removing an embedded backdoor typically requires full model retraining on a carefully audited dataset or, at minimum, comprehensive sanitization to eliminate residual traces of the manipulation.

➢ *Model Inversion*

This category of security vulnerability involves the extraction of sensitive training data or private attributes through the systematic exploitation of a model's prediction API. By observing the confidence scores or probability distributions returned by a target model, an adversary can perform iterative queries to reconstruct private features or even entire training samples. These privacy-breaching attacks, alongside membership inference techniques, facilitate the unauthorized exposure of data utilized in model development, effectively circumventing standard access controls (Patil & Zuber, 2023). These risks are further compounded by model extraction attacks, where adversaries query the system to replicate its hyperparameters or internal architecture, ultimately facilitating the construction of a surrogate model for offline exploitation (Roshan et al., 2023). By repeatedly querying the system, attackers strip away the value of proprietary intellectual property and turn a protected interface into an avenue for illicit replication. Obtaining a high-fidelity surrogate allows these actors to bypass standard usage fees and commercialize stolen functionality through rival interfaces at minimal expense, often requiring only a small number of queries to successfully reconstruct the target. Beyond the immediate loss of confidentiality, this cloned model serves as an effective staging ground for follow-up attacks, as techniques like membership inference and adversarial perturbation generation developed on the proxy often transfer seamlessly to the original system, further undermining its security and trustworthiness.

➢ *Defense Mechanisms*

Robust defense frameworks necessitate a layered approach, integrating both proactive training techniques and reactive monitoring systems to neutralize potential adversarial influence. Specifically, techniques such as adversarial training must be balanced with anomaly detection systems that monitor query frequency and output distribution to identify suspicious extraction attempts in real-time (Zhang et al., 2021). Moreover, deploying differential privacy methods introduces calibrated noise into the system's responses, making it difficult for unauthorized parties to infer private training information or reconstruct key model parameters. Furthermore, increasing the cost of extraction for adversaries requires a multi-layered approach that integrates API rate limiting and response obfuscation with principled output perturbation—such as query unlearning—and continuous statistical monitoring of query distributions, ultimately ensuring that the resources required for illicit replication exceed the value of the stolen asset.

➢ *Adversarial Training*

This technique involves augmenting the training process with perturbed samples to enhance the model's resilience against adversarial examples that exploit small, non-perceptual input modifications. By proactively incorporating these adversarial inputs into the training set, the model learns to generalize more robustly, effectively narrowing the gap between clean and perturbed decision boundaries. This integration forces the neural network to prioritize invariant features over noise-sensitive artifacts, effectively regularizing the latent space to prevent classification errors

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1744

triggered by adversarial perturbations (Oseni et al., 2021). Adversarial training must nevertheless reconcile the near-unavoidable trade-off—formalized as a theoretical lower bound on the sum of clean and robust errors—between fortified resistance to specific perturbations and degraded accuracy on pristine inputs, a tension that is compounded by robust overfitting in which the gap between training and testing robustness widens as epochs progress, so that state-of-the-art recipes on benchmarks like CIFAR-10, SVHN, and ImageNet typically reach their peak robust accuracy at intermediate stages rather than at convergence, prompting researchers to pursue adaptive perturbation budgets, curriculum-based sample selection, and asymmetric loss formulations as ways to recover clean-data fidelity without surrendering defensive potency.

➢ *Detection Methods*

Operating as a reactive line of defense, these approaches screen incoming samples by inspecting intermediate neuron activations or measuring statistical departures from the expected feature distributions, enabling the system to flag or filter out suspicious inputs prior to reaching the final classification stage. For instance, feature squeezing mitigates adversarial noise by projecting inputs into restricted, low-dimensional subspaces, yet like other detection mechanisms grounded in statistical invariants or local consistency, these approaches are frequently subverted by sophisticated attacks designed to operate within learned boundaries, compelling defenders into a relentless cycle of continuous adaptation against adversaries that adeptly circumvent static filters.

➢ *Gradient Masking*

- This defense strategy undermines gradient-based attacks by concealing or distorting the model's gradient information, making it harder for adversaries to apply optimization-driven methods that generate adversarial perturbations.
- By disrupting the calculation of these signals, the defense forces adversaries to rely on less efficient estimation techniques or surrogate-based strategies.
- However, this approach is inherently limited. It typically produces models that are robust only against white-box attacks while remaining susceptible to black-box transferability, in which adversaries use surrogate models to bypass the obscured gradients.
- Moreover, these methods often produce shattered or stochastic gradients, creating unstable, non-differentiable landscapes that adversaries bypass through gradient-free optimization or expectation-over-transformation techniques.
- Papernot and colleagues provided an early demonstration that defenses based on infinitesimal perturbations could be circumvented at rates of around seventy percent by transferring adversarial examples constructed on a locally trained substitute, establishing that gradient masking conceals vulnerability rather than eliminates it.
- Athalye and coauthors later systematized this concern by dismantling seven of nine defenses published at ICLR 2018, attributing their collapse to recurring failures such as shattered, stochastic, and vanishing gradients, which adaptive adversaries routinely exploit through techniques including backward-pass differentiable approximation and expectation-over-transformation.
- These findings expose a persistent gap between apparent and substantive robustness, showing that the field's reliance on static white-box evaluations produces an illusionary sense of security that adaptive attackers routinely puncture.
- Advancing the discipline beyond this impasse therefore requires rigorous adaptive evaluation procedures, in which adversaries are granted full knowledge of the defense and freedom to tailor their strategies, before any claim of certified robustness can be responsibly advanced.
- Such adaptive testing should be complemented by ablation of common obfuscation patterns, inspection of loss curvature near decision boundaries, and cross-validation against transfer-based and gradient-free attack variants, ensuring that reported resilience reflects genuine invariance rather than artifacts of the defense mechanism.
- This evidentiary standard compels the community to treat published defenses as provisional hypotheses awaiting adversarial falsification rather than as settled security guarantees, an orientation that will gradually steer research toward mechanisms with mathematically provable robustness rather than resilience merely observed under narrow threat models.

III. EVALUATION METRICS

To accurately verify the security of an artificial intelligence system, one must establish a detailed threat model that defines the potential attacker's capabilities, knowledge, and operational constraints. Such a framework must explicitly specify the adversary's access level—whether white-box, grey-box, or black-box—together with the permissible magnitude of perturbations and the target domain. Ambiguous threat assumptions often yield robustness metrics that fail to generalize to realistic scenarios or to account for modern adaptive adversaries. Conversely, clearly bounding the attacker's computational resources and knowledge supports standardized benchmarking capable of distinguishing genuine structural resilience from superficial performance gains.

➢ *Robustness Benchmarking*

To achieve truly rigorous evaluation, researchers must look beyond mere aggregate accuracy and instead conduct thorough ablation studies that systematically disable individual defense mechanisms to isolate their specific contributions to the overall resilience of the model. Moreover, applying a wider range of evaluation scenarios—such as testing against random noise and considering expanded threat environments—helps confirm that defenses are genuinely resilient rather than just overly optimized for particular adversarial perturbations. Finally, establishing clear and testable performance benchmarks under defined l_p-norm constraints facilitates consistent comparisons across diverse methodologies, helping to eliminate the ambiguity introduced by poorly defined threat models. Adhering to these methodological standards ensures that reported security gains are not merely artifacts of obfuscated

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1744

gradients but rather represent demonstrable, verifiable improvements in defensive posture (Athalye et al., 2018).

- *Success Rate Analysis*

This indicator measures how frequently an attacker manages to cause the model to misclassify inputs under a given threat model, thereby acting as a practical gauge for evaluating the system's baseline security level. Beyond simple error rates, analysts should further compute vulnerability scores to quantify how sensitive specific internal model components are to various input perturbations. Complementing these quantitative outcomes with finer-grained, data-oriented indicators—including measurements of activation coverage across internal neurons and targeted sensitivity profiling of individual architectural components—yields a far more granular diagnosis of how particular layers, feature detectors, or structural pathways precipitate misclassification, allowing practitioners to localize the latent instabilities that aggregated vulnerability scores necessarily average away. Furthermore, integrating resilience metrics such as the percentage drop in accuracy under adversarial conditions offers a holistic perspective on a system's capacity to maintain operational integrity when confronted with active, sophisticated threats. By contextualizing the severity of performance degradation against baseline efficacy, practitioners can establish more robust thresholds for acceptable failure, ensuring that the model's defensive posture is assessed not merely in terms of binary classification outcomes, but across a nuanced spectrum of operational reliability.

However, this expansion of diagnostic granularity is not without substantive critique. Critics argue that an overemphasis on component-level vulnerability profiling risks substituting architectural introspection for adversarial rigor, producing elaborate yet epistemically shallow evaluations that confuse mechanistic explainability with empirical robustness. Activation coverage and sensitivity indices, while informative in controlled settings, often fail to generalize across threat models because they remain tightly coupled to the specific topology and training regime of the evaluated model, thereby reproducing the very transferability blind spots that undermine white-box-centric assessment paradigms. Moreover, attempts to derive holistic resilience indices from percentage-based accuracy degradation can mask the non-uniform distribution of adversarial impact: a modest aggregate drop may obscure catastrophic failures concentrated in semantically critical classes, while conversely, a pronounced aggregate decline may reflect benign outputs rather than exploitable weaknesses. Equally problematic is the tendency for granular metrics to invite adversarial gaming, whereby defenders optimize internal indicators without correspondingly improving end-to-end robustness, a dynamic reminiscent of the metric-targeting pathologies documented in broader machine learning evaluation. Skeptics therefore contend that the field's progressive migration toward finer-grained diagnostic instruments, though intellectually appealing, may inadvertently divert attention and resources from the comparatively unglamorous but indispensable tasks of standardized benchmarking, cross-architecture transfer testing, and adaptive adversary elicitation, which collectively furnish the most credible evidence of genuine defensive improvement.

- *Future Research Directions*

Future efforts should concentrate on building a comprehensive theoretical foundation that harmonizes the diverse approaches to attack and defense across varied model architectures, including both convolutional neural networks and attention-based transformers. This theoretical framework must move beyond current techniques tailored to specific architectures, instead addressing the fundamental differences in learning dynamics and feature extraction that constrain the portability of defensive strategies. Equally important will be the cultivation of genuinely transferable defensive strategies that can move fluidly between fundamentally different model paradigms, since the current paradigm of bespoke, architecture-specific solutions severely constrains real-world deployment at scale. Equally critical moving forward is the broader research community's obligation to subject every newly proposed defensive mechanism to genuinely demanding threat scenarios that impose no artificial handicaps upon would-be attackers, since constrained evaluations routinely produce inflated impressions of security that collapse the moment a determined adversary exploits the omitted capabilities. Rigorous testing protocols must therefore explicitly incorporate adaptive opponents who possess full knowledge of the safeguard under examination, given that prior investigations have consistently demonstrated how non-adaptive benchmarking tends to obscure exploitable weaknesses that only surface when attackers consciously tailor their strategies against the very mechanism intended to repel them. The construction of shared community repositories, openly accessible leaderboards, and reproducible software infrastructures that aggregate the strongest contemporary attacks across diverse perturbation norms would furnish a much-needed common ground for cross-study comparison, exposing the actual empirical boundaries of current defensive practice and separating substantiated advances from overstated claims of progress.

IV. CONCLUSION

Ultimately, advancing the field of adversarial machine learning requires transitioning from reliance on empirical, model-specific defenses to establishing foundational, universally verifiable guarantees of robustness. Realizing this objective requires a comprehensive shift toward exploring the theoretical foundations of adversarial interactions by employing advanced methods from topology, statistical analysis, and learning theory to bypass the boundaries of existing empirical research. This endeavor must also be complemented by the development of sophisticated threat models that assess operational scenarios where machine learning deployments are uniquely vulnerable. Equally essential is the pursuit of mathematically verifiable limits of resilience through formal verification methods such as interval bound propagation and abstract interpretation, which permit the derivation of provable safety assurances that hold under any admissible adversarial perturbation, thereby

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jun1744

substituting the unpredictability of heuristic defense evaluation with a principled framework anchored in deductive inference rather than architecture-dependent empirical benchmarking. Such a transformation is poised to close the divide between brittle, empirically crafted safeguards and durable systems that preserve dependable behavior even when confronted with adversarial techniques never encountered during their design or validation.

REFERENCES

[1]. Athalye, A., Carlini, N., & Wagner, D. T. (2018). Obfuscated Gradients Give a False Sense of Security: Circumventing\n Defenses to Adversarial Examples. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.1802.00420

[2]. Olowononi, F. O., Rawat, D. B., & Liu, C. (2020). Resilient Machine Learning for Networked Cyber Physical Systems: A Survey for Machine Learning Security to Securing Machine Learning for CPS. *arXiv (Cornell University)*, *23*(1), 524–552. https://doi.org/10.1109/comst.2020.3036778

[3]. Oseni, A., Moustafa, N., Janicke, H., Liu, P., Tari, Z., & Vasilakos, A. V. (2021). Security and Privacy for Artificial Intelligence: Opportunities and\n Challenges. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2102.04661

[4]. Patil, C., & Zuber. (2023). A Reinvent Survey on Machine Learning Attacks. *E3S Web of Conferences*, *453*, 1016–1016. https://doi.org/10.1051/e3sconf/202345301016

[5]. Roshan, K., Zafar, A., & Haque, S. B. U. (2023). Untargeted white-box adversarial attack with heuristic defence methods in real-time deep learning based network intrusion detection system. *arXiv (Cornell University)*, *218*, 97–113. https://doi.org/10.1016/j.comcom.2023.09.030

[6]. Tramèr, F., Carlini, N., Brendel, W., & Mądry, A. (2020). On Adaptive Attacks to Adversarial Example Defenses. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2002.08347

[7]. Vassilev, A., Oprea, A., Fordyce, A. J., & Anderson, H. S. (2024). *Adversarial machine learning :* https://doi.org/10.6028/nist.ai.100-2e2023

[8]. Wong, E., & Kolter, J. Z. (2017). Provable defenses against adversarial examples via the convex outer adversarial polytope. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.1711.00851

[9]. Zhang, X., Chen, C., Xie, Y., Chen, X., Zhang, J., & Xiang, Y. (2021). Privacy Inference Attacks and Defenses in Cloud-based Deep Neural Network: A Survey. In *arXiv (Cornell University)*. Cornell University. https://doi.org/10.48550/arxiv.2105.06300