

Review

Adversarial Machine Learning for Secure and Explainable AI Systems: A Comprehensive Review

Hajar Ouazza ^{1,*} , Fadoua Khennou ²  and Abderrahim Abdellaoui ¹ ¹ Engineering Sciences Laboratory, Ibn Tofail University, Kenitra 14000, Morocco; abderrahim@uit.ac.ma² Perception, Robotics and Intelligent Machines Lab, Computer Science Department, Université de Moncton, Moncton, NB E1A 3E9, Canada; fadoua.khennou@umoncton.ca

* Correspondence: hajar.ouazza@uit.ac.ma

Abstract

Adversarial machine learning (AML), reinforcement learning (RL), and explainable artificial intelligence (XAI) are increasingly studied as separate problems, yet their interactions under realistic threat conditions remain poorly understood. This review addresses that gap through a systematic analysis of 207 studies selected from 4447 records following the PRISMA 2020 guidelines, covering work published between 2020 and 2026 across cybersecurity and computer vision. A taxonomy of adversarial attacks is constructed across training and inference phases, defense mechanisms are examined with attention to their documented failure modes, and robustness evaluation practices are assessed across the surveyed literature. RL is analyzed in both offensive and defensive roles. Attack agents using RL achieve evasion rates of 74–97% against ML-based detectors, while RL-based defenses report robustness gains of up to 3× over static baselines under comparable threat conditions. XAI receives particular attention because the field treats it almost exclusively as a transparency mechanism, whereas the reviewed evidence shows that it also functions as an attack surface. Attribution methods such as LIME, SHAP, and Grad-CAM produce unreliable explanations under adversarial perturbation, and no system in the reviewed literature certifies that attribution properties are maintained when inputs are manipulated. The review concludes with an analysis of open problems and research directions for building systems that are robust against adaptive adversaries, interpretable under operational constraints, and auditable in environments where AI accountability is a legal requirement.

Keywords: adversarial machine learning; adversarial attacks; defense techniques; reinforcement learning; explainable artificial intelligence



Academic Editors: Tao Zhang,
Xiangyun Tang, Jiacheng Wang,
Chuan Zhang and Jiqiang Liu

Received: 15 June 2026

Revised: 18 July 2026

Accepted: 25 July 2026

Published: 7 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Machine learning and deep learning now drive automated decisions in autonomous systems, healthcare, finance, and cybersecurity. As these systems take on real operational weight, questions about their security and reliability have stopped being theoretical. The same performance gains that make modern AI attractive in high-stakes settings come with a structural problem: small carefully constructed input changes that are often invisible to human observers can push models into producing wrong or harmful outputs [1–3].

The scale of the problem is now documented. The AI Incident Database recorded 362 documented AI-related incidents in 2025, up from 233 in 2024, which had itself risen 56.4% over the previous year [4]. ENISA analyzed 4875 cybersecurity incidents across the European Union between July 2024 and June 2025 and characterized the period as the first

in which AI reshaped the threat landscape rather than merely appearing within it, with AI-supported phishing accounting for more than 80% of the observed social engineering activity by early 2025 and model poisoning listed among the techniques adversaries now use operationally [5]. The consequences reach deployment decisions. In Stanford's 2026 survey, 62% of organizations identified security and risk as the primary barrier to scaling agentic AI, ahead of technical limitations and regulatory uncertainty [4]. These figures describe a field where adversarial vulnerability has moved from a laboratory concern to a constraint on what organizations are willing to deploy.

AML developed precisely to address this gap. At its core, the field studies how attackers locate and exploit weaknesses in learning systems and what it takes to close them. This involves a set of interconnected concepts spanning attack construction, defense design, and the ongoing tension between the two. Nailing down the key terminology early on matters here because the rest of the discussion builds directly on it.

In this context, an adversary refers to the attacker, i.e., an entity that intentionally seeks to mislead a machine learning model in order to produce incorrect or unexpected outcomes. To achieve this, the adversary introduces small and deliberate modifications, known as perturbations, to input data, exploiting the model's sensitivity to minor variations. A threat model defines the adversary's capabilities and knowledge about the system, specifying what components can be accessed or manipulated as well as the attacker's objectives. Within this framework, an adversarial attack refers to the process of generating manipulated inputs that deceive the model, either during training or at inference time. The resulting inputs, known as adversarial examples, often appear indistinguishable from normal data to humans while successfully causing the model to produce incorrect predictions, as illustrated in Figure 1.

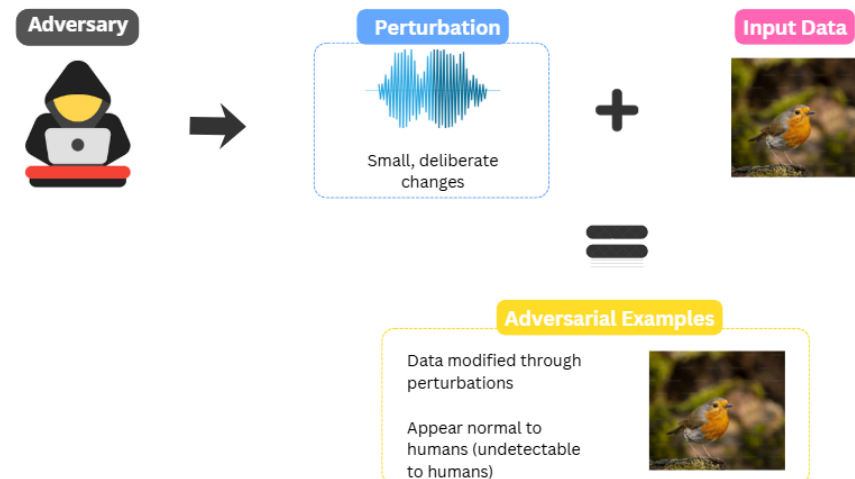


Figure 1. Threat scenarios for adversarial example generation.

Existing studies commonly distinguish between training-time attacks, such as data poisoning, and inference-time attacks, such as evasion attacks [6,7]. Importantly, these threats are no longer purely theoretical. Prior research has demonstrated their impact in practical settings, including malware detection, facial recognition, autonomous driving, and network intrusion detection systems [8,9]. Considerable research effort has therefore focused on understanding adversarial behavior and improving model robustness. This has produced attack taxonomies, evaluation benchmarks, and a range of defense mechanisms. Adversarial training, input transformation, randomization-based defenses, gradient masking, and certified robustness methods [10–13] each improve resistance against specific attacks. None offers complete protection. Against adaptive adversaries that keep adjusting

their strategies, the gap between a defense and its defeat tends to be short. Reinforcement learning (RL) has attracted real attention here. RL systems can themselves be attacked through manipulated observations, rewards, or policies, and they also offer a practical framework in which adaptive defense agents can learn protective behaviors through direct environmental interaction [14,15]. Recent work puts RL in both roles: modeling intelligent attackers and building robust responses to them [16,17]. XAI is a separate but related thread. XAI techniques make complex models more interpretable by surfacing how they reach their outputs [18,19]. In adversarial settings this matters because transparency can expose abnormal behavior and support security analysis. It also turns out that explanation methods are not immune to attack but can be manipulated to produce misleading or unstable interpretations when perturbations are introduced [20,21]. That has pushed work toward combining XAI with AML, trying to build defenses that are both effective and legible [22,23]. Taken together, AML, reinforcement learning, and XAI form a productive intersection for secure AI design. Robustness, adaptability, and interpretability are not independent goals. Systems that resist attacks, adjust to new threats, and explain themselves clearly to human users require all three. This survey focuses on cybersecurity and computer vision, two domains where adversarial threats are serious and recent progress has been fast.

1.1. Scope of the Survey

The survey covers AML through the lens of RL and XAI, with cybersecurity and computer vision as the primary domains. On the cybersecurity side: malware detection, intrusion detection systems, fraud detection, and threat intelligence. In computer vision: image classification, object detection, facial recognition, and autonomous perception. Across both, the survey examines poisoning and evasion attacks, common robustness evaluation methods, how RL has been used for both attack and defense, and the growing role of XAI in detecting attacks and auditing models. The focus is deliberately narrow depth over breadth, although many of the findings carry over to other domains with similar adversarial dynamics.

1.2. Related Surveys

1.2.1. Cybersecurity

Several surveys have examined AML in cybersecurity, although most target narrow subproblems. Mohammed [24] frames AI-powered cyberattacks as an emerging offensive capability but stays on attack scenarios and does not reach intelligent defense, interpretability, or automated response. Ling et al. [25] review adversarial attacks against Windows PE malware detectors with careful attention to preserving executability, which is technically solid but limited to PE malware. McCarthy et al. [26] study functionality-preserving attacks in malware and IDS contexts, again centering on evasion with no treatment of RL or explainability. Broader surveys follow similar lines. Zhou et al. [27] organize adversarial attacks through a cyber kill-chain lens. Macas et al. [28] classify attacks and defenses across security applications. Alotaibi and Rassam [29] narrow to AML threats against intrusion detection. Rosenberg et al. [30] propose a cross-domain taxonomy. Martins et al. [31] document recurring problems, outdated datasets, unrealistic evaluations across malware, and intrusion scenarios. Vitorino et al. [32] go further, arguing that network perturbations need to respect protocol semantics and traffic validity. What these studies share, beyond their individual contributions, is a limited engagement with reinforcement learning for adaptive defense and little attention to explainability as a tool for analyst trust. More recent reviews do not change the picture much. Ke et al. [33] introduce AML as a security concern but stay introductory. Khaleel et al. [34] apply PRISMA methodology to cybersecurity defense and

flag data scarcity without exploring adversarial explainability risks or autonomous learning strategies. Rane et al. [35] discuss resilient AI through adversarial training and model hardening but treat robustness as a broad engineering problem rather than an integrated AML, RL, and XAI question.

1.2.2. Computer Vision

AML surveys in computer vision are more extensive and tend to be disconnected from cybersecurity concerns. Liang et al. [36] review attack and defense methods for image models from a traditional robustness perspective. Stanly et al. [37] distinguish generative from non-generative attacks, focusing on image attack generation rather than secure deployment. Li et al. [38] extend to 2D and 3D models and cover physical-world attacks that are relevant to autonomous systems, although explainability and adaptive defense get less attention. Akhtar et al. [39] survey images, video, and spatial transformations broadly, with limited reference to RL settings. Costa et al. [40] position adversarial robustness as a central obstacle to trustworthy AI but do not integrate XAI or cybersecurity-specific operational constraints. As Table 1 shows, these themes are studied in isolation. Cybersecurity reviews concentrate on attacks, malware, and network robustness. Computer vision surveys deal in image perturbations and benchmark defenses. XAI work discusses interpretability without adversarial settings. RL research covers optimization and control without security-aware explainability. The interactions between these areas under realistic adversarial conditions remain largely unexplored. This survey takes a different approach, examining AML, reinforcement learning, and XAI as connected problems across cybersecurity and computer vision. We analyze how RL supports adaptive attack and defense strategies, how XAI both builds trust and opens new attack surfaces, and why robustness evaluation should cover perception and security tasks together. That joint perspective matters for the design of AI systems operating in dynamic safety-critical environments.

Table 1. Comparison of existing reviews on AML.

Review	Year	Domain	Adversarial Attacks	Defense Methods	RL	XAI
[36]	2022	Computer vision	✓	✓	✗	✗
[37]	2023	Computer vision	✓	✗	✗	✗
[38]	2024	Computer vision	✓	✗	✗	✗
[24]	2025	Cybersecurity	✓	✓	✗	✗
[39]	2021	Computer vision	✓	✓	✓	✗
[25]	2023	Cybersecurity (Windows PE malware)	✓	✓	✓	✗
[27]	2022	Cybersecurity	✓	✓	✗	✗
[28]	2024	Cybersecurity	✓	✓	✓	✗
[29]	2023	Cybersecurity (IDS)	✓	✓	✗	✗
[30]	2020	Cybersecurity	✓	✓	✗	✗
[31]	2020	Cybersecurity (IDS & malware detection)	✓	✓	✗	✗
[33]	2025	Cybersecurity	✓	✓	✗	✗
[26]	2022	Cybersecurity (IDS)	✓	✓	✗	✗
[34]	2024	Cybersecurity	✓	✓	✗	✗
[32]	2023	Cybersecurity (NIDS)	✓	✓	✗	✗
[35]	2025	Cybersecurity	✗	✓	✗	✗
This review	–	Cybersecurity & Computer vision	✓	✓	✓	✓

✓ = topic covered in the review. ✗ = topic not addressed.

1.3. Contributions

This survey unifies three areas that are typically studied apart, adversarial robustness, reinforcement learning, and XAI, into a single review of AML. Rather than limiting the

discussion to attack-versus-defense comparisons, we examine how RL supports adaptive security mechanisms and how XAI improves threat detection and system transparency. We also discuss evaluation metrics that go beyond attack success rates and model robustness to include adaptability, resilience, and interpretability under adversarial conditions. We document the major limitations in the existing approaches: scalability issues, transferability challenges, and the vulnerability of explanation techniques to adversarial manipulation. Finally, we identify research directions for AI systems that are secure, transparent, and better prepared for evolving threats.

1.4. Structure of This Survey

The paper is organized into six sections, as illustrated in Figure 2. The overall structure of the survey, along with the objective of each section, is presented below.

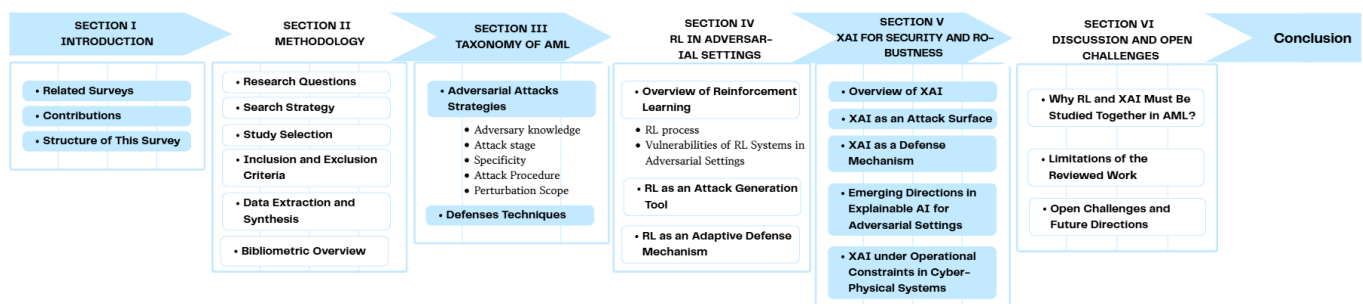


Figure 2. Review outline.

- **Section I—Introduction:** Presents the motivation, scope, and significance of studying the intersection of AML, RL, and XAI. It also defines the objectives of this survey and includes a brief review of the related work to position the study within the existing literature.
- **Section II—Literature Search Methodology:** Describes the systematic review protocol, including the search strategy, inclusion and exclusion criteria, and the PRISMA-based selection process used to ensure transparency and reproducibility. It also presents a bibliometric overview of the selected corpus, summarizing publication trends, domain distribution, method families, and dataset usage.
- **Section III—Taxonomy of Adversarial Machine Learning:** Presents a structured taxonomy of adversarial attack strategies and their characteristics and discusses the existing defense strategies and robustness techniques against adversarial attacks.
- **Section IV—Reinforcement Learning in Adversarial Settings:** Analyzes the role of RL in both attack generation and defense optimization within adversarial environments.
- **Section V—Explainable Artificial Intelligence for Security and Robustness:** Explores how XAI contributes to improving interpretability, detection, and trust in AML systems.
- **Section VI—Discussion and Open Challenges:** Synthesizes the key findings, identifies the research gaps, and discusses the limitations in the current AML, RL, and XAI integration.
- **Conclusion:** Summarizes the main contributions of the survey and presents directions for future research.

2. Literature Search Methodology

In this review, we adopt a structured and transparent methodology to examine the interplay between AML, RL, and XAI. The overall process is guided by the principles of the PRISMA 2020 framework [41], ensuring a systematic approach to study identification,

screening, eligibility assessment, and synthesis. To ensure reliability, we designed a reproducible process with clear research questions, defined selection criteria, and a consistent screening procedure. Efforts were made to minimize bias and ensure coverage of the most relevant and rigorous work published between 2020 and 2026. This approach enables a balanced view of how AML, RL, and XAI interact in adversarial settings, emphasizing their combined potential rather than treating them separately, as illustrated in Figure 3.

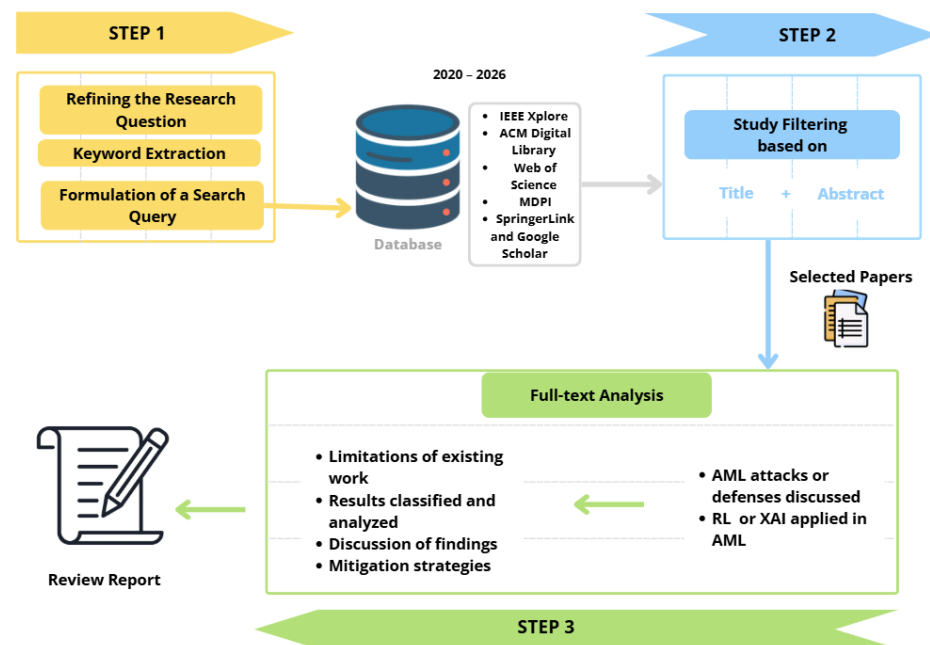


Figure 3. Methodology process.

2.1. Research Questions

Four research questions guided the review:

- RQ1: What challenges affect AI robustness under adversarial attacks, and how well do existing countermeasures hold up?
- RQ2: Can RL be used both to generate adversarial attacks and defend against them?
- RQ3: What role can XAI play in detecting and mitigating adversarial attacks?
- RQ4: What are the current gaps and open problems in securing AI models against adversarial threats?

These questions keep the review focused on the intersection of AML, RL, and XAI—and what that intersection actually means for building systems robust enough to withstand real attacks.

2.2. Search Strategy

Six databases were queried: IEEE Xplore, ACM Digital Library, Web of Science, MDPI, SpringerLink, and Google Scholar. Their selection reflects complementary coverage across machine learning, security, and systems research, spanning both theoretical foundations and applied adversarial work. A single Boolean query was constructed from three thematic clusters and applied uniformly across all platforms, with syntax adjusted to each interface.

The first cluster grouped attack-related terminology: “adversarial machine learning,” “AML,” “adversarial attacks,” “evasion attacks,” “poisoning attacks,” “white-box,” and “black-box.” The second cluster covered interpretability: “explainability in AI” and “XAI.” The third captured the learning paradigm: “reinforcement learning” and “RL.” The resulting query took the following form.

("adversarial machine learning" OR "AML" OR "adversarial attacks" OR "evasion attacks" OR "poisoning attacks" OR "white-box" OR "black-box") AND ("explainability in AI" OR "XAI") AND ("reinforcement learning" OR "RL")

Searches were conducted in January 2026. Results were restricted to publications from January 2020 onward, a boundary chosen because the consolidation of adversarial ML as a distinct research area largely post-dates that point. Only peer-reviewed journal articles and conference papers in English were retained. The query was applied using the standard search interface of each platform. Relevance was verified at the title-and-abstract screening stage for all retrieved records.

2.3. Study Selection

The search returned 4447 records in total across all six databases (IEEE Xplore: 902, ACM Digital Library: 726, Web of Science: 1090, MDPI: 1200, SpringerLink: 519, and Google Scholar: 10 articles from additional sources not covered by the primary databases). After removing 312 duplicates, 4135 unique records moved to title and abstract screening. At this stage, we excluded 3267 studies that were clearly outside the survey's scope, leaving 868 for full-text review. A further 661 were excluded: 264 lacked substantive AML content, 218 addressed unrelated topics, 151 applied XAI or RL in non-adversarial settings, 23 provided insufficient methodological detail, and 5 were extended abstracts without full empirical content.

Figure 4 illustrates the screening process, highlighting how studies were gradually filtered out at each stage.

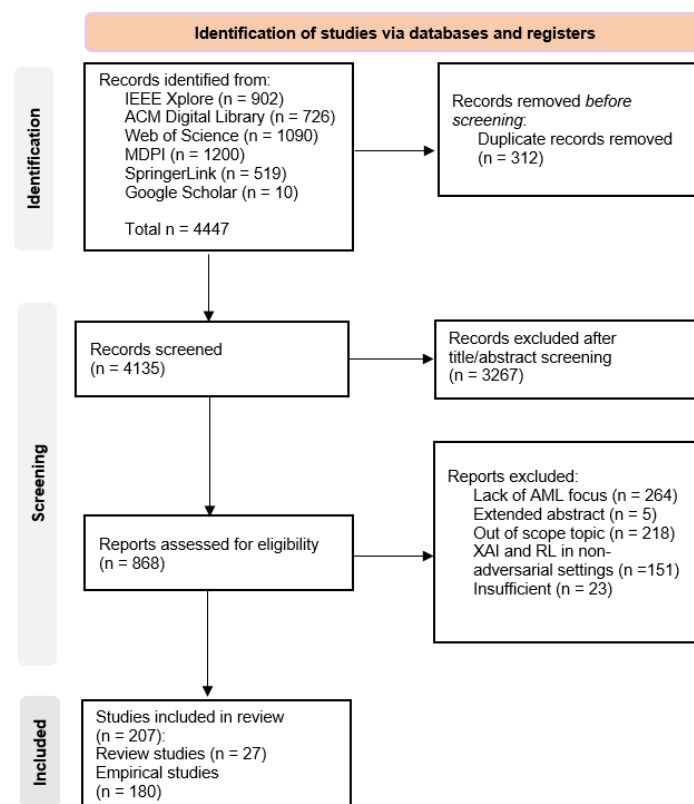


Figure 4. Flow diagram of study identification and selection using the PRISMA 2020 guidelines.

The final set includes 207 studies, with 180 empirical works and 27 review papers, all focusing on meaningful and well-documented links between adversarial robustness, explainability, or RL. The complete evidence matrix, which maps each of the 180 empirical

studies to its domain, threat model, attack or defense role, RL and XAI involvement, and the research questions it addresses, is reported in Table A1 in Appendix A.

2.4. Inclusion and Exclusion Criteria

We followed clearly defined inclusion and exclusion criteria to ensure the relevance, consistency, and methodological rigor of the study selection process. Table 2 summarizes these criteria.

Table 2. Inclusion and exclusion criteria.

Inclusion Criteria	Exclusion Criteria
Studies focusing on AML, including its integration with RL and/or XAI.	Studies lacking a clear focus on AML or adversarial settings.
Research addressing adversarial settings (attacks, defenses, robustness, or adaptive learning).	Works involving RL or XAI only in non-adversarial contexts.
Applications in relevant domains such as cybersecurity, computer vision, or safety-critical systems.	Out-of-scope topics not aligned with the objectives of this survey.
Peer-reviewed articles or high-quality studies with empirical or theoretical contributions.	Extended abstracts, posters, or papers with insufficient technical depth.
Studies with clear methodology, experimental validation, and well-reported results.	Studies with insufficient methodological detail or weak experimental support.

Each study was evaluated throughout the selection process based on the clarity of its objectives, methodological rigor, soundness of experimental design, and relevance to AML, RL, and XAI. Only studies that satisfied these criteria were included.

2.5. Data Extraction and Synthesis

Data were extracted from 207 studies using a structured template capturing key elements such as research objectives, methods, adversarial setting (attack or defense), use of RL/XAI, datasets, evaluation metrics, and main findings. To ensure consistency, the extraction process involved iterative verification and cross-checking. A narrative synthesis was then conducted, organized around the main research questions. The analysis focuses on adversarial attacks, defense strategies, the role of RL in adversarial environments, and the contribution of XAI to interpretability and robustness. This approach enables a clear comparison of existing work while highlighting key trends, limitations, and research gaps.

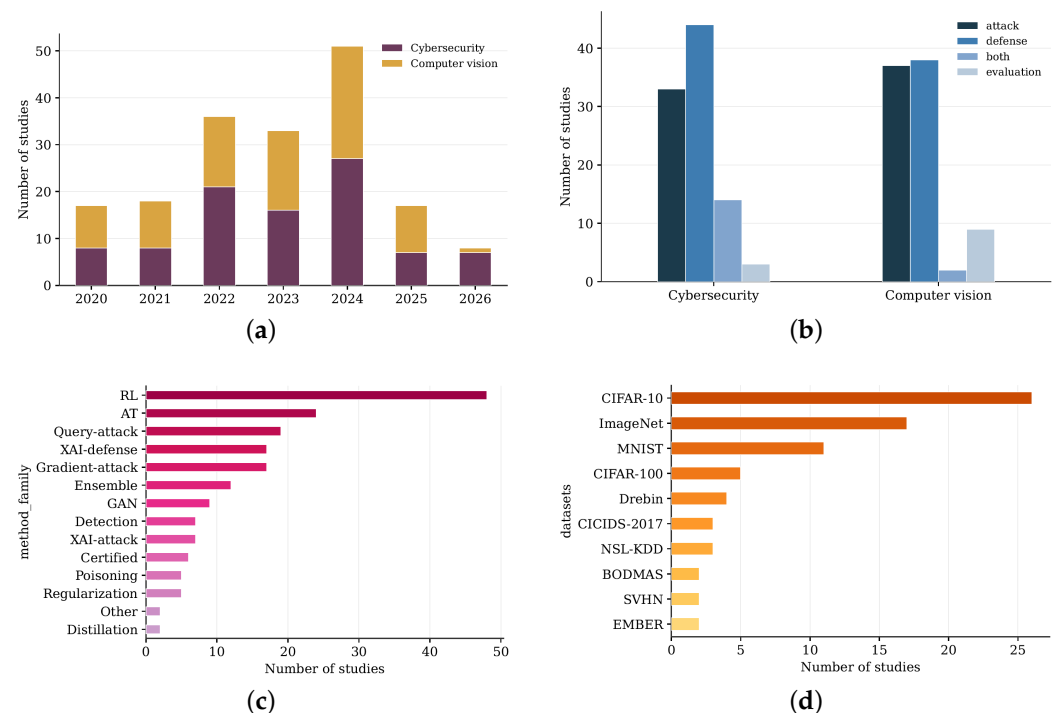
2.6. Bibliometric Overview

The 180 empirical studies retained after screening were examined along four dimensions: publication trend, study role, method family, and dataset usage. These figures describe the reviewed corpus rather than the adversarial machine learning literature as a whole since the selection criteria defined in Section 2.4 shape what enters the sample. Read with that qualification, they indicate where the reviewed evidence is concentrated and where it is thin. Table 3 reports the distribution of these studies across the four research questions and the two application domains.

Table 3. Distribution of the 180 empirical studies across the four research questions and the two application domains. A study may address more than one research question.

RQ	Focus	Total	Cyber	Vision
RQ1	Robustness challenges and countermeasures	168	91	77
RQ2	RL for attack generation and defense	53	42	11
RQ3	XAI for detection and mitigation	30	15	15
RQ4	Gaps and open problems	12	3	9

Annual output within the corpus (Figure 5a) rises from 17 studies in 2020 to 51 in 2024, with cybersecurity and computer vision contributing in near-equal measure (94 and 86 studies overall). The lower counts for 2025 and 2026 reflect incomplete indexing at the time of the search rather than a measurable decline in activity.

**Figure 5.** Bibliometric overview of the 180 empirical studies included in this review. (a) Publications per year by domain; (b) study roles by domain; (c) method families across the corpus; (d) most frequently used datasets.

Study roles differ between the two domains (Figure 5b). Defense work is the largest category in both (44 studies in cybersecurity; 38 in vision), but studies that address attack and defense jointly are concentrated in cybersecurity, which accounts for 14 of the 16 such works in the corpus. Vision studies more often isolate either attack or defense as a standalone contribution. Whether this pattern generalizes beyond the sample cannot be established from these counts alone, but, within the reviewed evidence, it means that co-designed threat-and-countermeasure evaluations are available almost exclusively for security tasks.

Method families (Figure 5c) are led by reinforcement learning with 48 studies, ahead of adversarial training (24) and query-based black-box attacks (19). This ordering is a direct consequence of the inclusion criteria, which required RL or XAI involvement, and should not be read as a measure of the relative prevalence of these methods in the wider field. It does show, however, that the RL-based work meeting those criteria is substantial enough to support the analysis in Section 4. Certified defenses appear in only 6 of the 180 studies, which limits what this review can say about provable robustness and is itself part of the gap discussed in Section 6.

Dataset usage within the corpus is narrow (Figure 5d). Across 110 dataset mentions, CIFAR-10, ImageNet, and MNIST account for 49%, while security-specific benchmarks such as NSL-KDD, CICIDS-2017, and BODMAS appear only a handful of times each. Because the corpus spans cybersecurity and vision in near-equal proportion, this imbalance is not explained by domain composition alone, and it gives quantitative form to the benchmark-narrowness limitation discussed in Section 6.2.

One cross-cutting count motivates the review itself. Reinforcement learning appears in 53 of the 180 studies and XAI in 30, yet only one study combines the two. Since both were inclusion criteria, this near-absence is not a sampling artifact. It reflects the reviewed literature.

3. Taxonomy of Adversarial Machine Learning

3.1. Adversarial Attack Strategies

Researchers commonly classify adversarial attacks according to the attacker's knowledge, objectives, capabilities, and the nature of the applied perturbations, as illustrated in Figure 6.



Figure 6. Adversarial attack strategies. Arrows indicate hierarchical relationships between attack categories and their subcategories.

3.1.1. Adversary Knowledge

White-box: In the white-box setting, the adversary has complete knowledge of the target model, including its architecture, parameters, and gradients. This level of access enables highly optimized perturbations and therefore represents one of the strongest threat models in AML. Recent studies show that white-box research has evolved along two parallel directions: improving attack efficiency and strengthening model robustness. For instance, ref. [42] introduced a dynamic programming-based method that achieves successful misclassification by modifying only a limited number of pixels, highlighting that effective attacks do not necessarily require large perturbations. Similarly, ref. [43] demonstrated that vulnerabilities may also emerge from preprocessing pipelines, where visually benign inputs can become adversarial after common image transformations such as scaling or sharpening. These findings indicate that white-box risk extends beyond the classifier itself to the broader data-processing chain.

On the defense side, several works focused on improving intrinsic robustness. The EMSF²Net framework proposed in [44] combined multi-stage feature fusion with regularization and reported stronger resistance to gradient-based attacks while preserving clean-data accuracy. In cybersecurity contexts, similar conclusions were reported by [45], where adversarial training and heuristic hardening methods significantly restored the performance of deep-learning-based IDSs under attack. Likewise, ref. [46] showed that adversarial retraining

can substantially improve malware detection robustness while also reducing false alarms. Taken together, these studies suggest that, although white-box attacks remain powerful, carefully designed training and architectural strategies can meaningfully reduce their impact.

Black-box: In contrast, black-box attacks assume no access to model internals, forcing the adversary to rely on queries, predicted labels, or confidence scores. Because of this restricted access, recent research has largely centered on query efficiency, transferability, and gradient-free optimization. Swarm intelligence and evolutionary search remain particularly effective in this setting. Studies based on artificial fish swarm optimization [47], artificial bee colony search [48], and evolutionary optimization [49] showed that high ASR can be achieved without explicit gradients. This demonstrates that complex models may still be vulnerable even when internal information is completely hidden.

A second major trend is surrogate-model learning. Instead of attacking the target directly, adversaries train substitute models and transfer the generated perturbations. FE-DaST [50] showed that this can be done even without access to the original training data, while meta-learning-based simulators further improved attack transferability and efficiency [51]. Partial retraining strategies also reduced query requirements in constrained scenarios [52]. Collectively, these works indicate that black-box restrictions can often be bypassed once an accurate surrogate approximation is obtained.

Generative models have also expanded the realism of black-box attacks. For example, attackGAN [53] generated malicious network traffic that successfully evaded IDSs, while DDSG-GAN [54] improved cross-model transferability through adversarial generation. In security-oriented domains, RL has emerged as another practical attack paradigm. Prior studies used RL to generate evasive malware variants [55], bypass XSS detection mechanisms [56], and automate SQL injection evasion against learning-based detectors [57]. Similar vulnerabilities were reported for binary function detection [58], Android malware classifiers [59], and confidence-score-based malware detectors [60]. These results confirm that black-box attacks are no longer limited to image classification but represent a broad threat across cybersecurity applications.

Defensive research in the black-box setting remains more challenging. Query-monitoring frameworks such as MalDBA [61] attempt to identify suspicious interaction patterns before successful exploitation occurs. However, large-scale evaluations in [62] found that many proposed defenses degrade significantly when tested against adaptive attackers. Meanwhile, decision-based optimization methods such as boundary attacks and their improved variants continue to reduce query costs and increase practicality [63,64]. Even minimal manipulations, such as structured pixel rearrangements, were shown to strongly affect both RGB and infrared models [65]. Furthermore, transferability studies in cybersecurity datasets confirmed that adversarial examples crafted on one model frequently generalize to others [66]. Overall, the literature suggests that black-box security remains an open challenge, particularly because practical constraints do not necessarily prevent highly effective attacks.

Gray-box: Between these two extremes lies the gray-box setting, where the adversary possesses partial knowledge of the model, training data, or deployment pipeline. Although less studied, this scenario is often more realistic in operational environments. For example, ref. [67] showed that semi-supervised learning systems relying on publicly shared labeled data remain vulnerable to gray-box attacks across image classification, object detection, and semantic segmentation tasks. This suggests that even limited external knowledge can substantially increase adversarial capability, making gray-box threats highly relevant for real-world deployments. We summarize the main characteristics of the reviewed studies in Table 4.

Table 4. Comprehensive summary of adversarial attacks (white-box, black-box, and gray-box).

Ref.	Year	Task	Type	Method	Model	Dataset	Key Results
[42]	2024	Computer vision	White-box	DPSPSM	VGG-16 CNN	CIFAR-10, CIFAR-100	ASR 30.45% (CIFAR-10), 29.30% (CIFAR-100)
[44]	2022	Computer vision	White-box	FGSM, PGD, MI-FGSM, CW	EMSF ² Net	CIFAR-10	Acc 92.7% clean; 42.7% FGSM; 53.8% PGD
[43]	2022	Computer vision	White-box	Content Camouflage Attack	ResNet-18, CNN	CIFAR-10, FER-2013	100% success; ~90% accuracy drop
[45]	2023	NIDS	White-box	FGSM, JSMA, PGD, CW	Deep NIDS	NSL-KDD, CICIDS-2017	Acc >20% drop; up to 94% recovery
[46]	2025	Malware detection	White-box	FGSM, PGD, BIM	DNN, CNN-GRU-Att	BODMAS	Acc 46–59% drop; up to 99.8% recovery
[67]	2024	Computer vision	Gray-box	GAAS	SSL models	CIFAR-10, CIFAR-100, SVHN	Up to 95% accuracy drop
[64]	2021	Computer vision	Black-box	DBBA	DNNs	CIFAR-10, ImageNet	4596 queries (efficient attack)
[68]	2023	Image classification	Black-box	SAGM	ResNet18, AlexNet	MNIST, CIFAR-10	100% ASR outperforming baselines
[48]	2022	Image classification	Black-box	ABCAttack	CNN, RNN, ResNet	MNIST, CIFAR-10, ImageNet	Up to 100% ASR (MNIST)
[49]	2022	Image classification	Black-box	QuEry Attack	Inception-v3, ResNet-50	CIFAR-10, MNIST	100% success, low queries
[51]	2022	Computer vision	Black-box	Simulator Attack+	CNN-based models	CIFAR-10, CIFAR-100	100% SR, 29–133 queries
[53]	2024	IDS	Black-box	attackGAN	IDS models	NSL-KDD	81.37% ASR
[62]	2021	Computer vision	Black-box	Adaptive attack	CNNs	CIFAR-10, Fashion-MNIST	Defense improves robustness <25%
[56]	2022	IDS	Black-box	Soft Q-learning attack	XSS models	XSS dataset	>85% evasion success
[66]	2024	IDS	Black-box	Transfer attack	DNNs	CICDDoS-2019	60–70% transfer success
[58]	2023	Malware detection	Black-box	Boundary perturbation	Neural analyzers	Binary dataset	>90% success
[63]	2024	Computer vision	Black-box	GOBA	CNNs	MNIST, ImageNet	+13.78% ASR gain
[60]	2023	Malware detection	Black-box	Confidence-guided attack	MalConv	PE files	90.8% evasion
[54]	2023	Computer vision	Black-box	DDSG-GAN	CNNs	MNIST, CIFAR-10	100% MNIST success
[47]	2024	Computer vision	Black-box	EFSAttack	CNNs	CIFAR-10, MNIST	100% ASR (few hundred queries)
[69]	2022	Computer vision	Black-box	EA vs. BIM	ResNet, VGG	ImageNet	0–5% transfer rate
[59]	2024	Android malware	Black-box	EvadeDroid	ML detectors	Android datasets	80–95% evasion
[50]	2022	Computer vision	Black-box	FE-DaST	CNNs	CIFAR, MNIST	95–100% MNIST success
[55]	2021	Malware detection	Black-box	MAB-Malware	ML/AV systems	EMBER dataset	74–97% ML evasion
[61]	2023	Malware detection	Black-box	MalDBA	MalConv	VirusShare	88% → 0% ASR
[52]	2020	Computer vision	Black-box	Substitute model attack	CNNs	ImageNet, MNIST	<6000 queries successful
[70]	2022	Computer vision	Black-box	NP-Attack	CNNs	CIFAR-10, ImageNet	100% ASR with low queries
[65]	2023	Computer vision	Black-box	Pixel rearrangement	CNNs	CIFAR-10, ImageNet	Cross-domain transferability

Table 4. *Cont.*

Ref.	Year	Task	Type	Method	Model	Dataset	Key Results
[57]	2023	Web security	Black-box	SSQLi	ML models	Web datasets	97.39% bypass success
[71]	2021	Malware detection	Black-box	A3CMal (RL)	DNN classifiers	Drebin	Misclassification, functionality kept
[37]	2021	NIDS (IoT)	Black-box	Model extraction + saliency	Kitsune	IoT traffic	94.31% ASR, <0.005% bytes changed
[72]	2025	Malware detection	Black-box	GAME-RL (RL)	API-call detectors	Real-world malware	Improved evasion, low overhead
[38]	2020	NIDS (SDN)	Black-box	Hydra evasion	Neptune (ML-NIDS)	SYN-flood DDoS	Detection 100% → 0%
[73]	2026	Malware detection	Black-box	Evolutionary (EA)	AV engines	Metamorphic malware	Evades up to 94% of AV detectors

3.1.2. Attack Stage

Adversarial attacks are commonly categorized into training-time poisoning attacks and inference-time evasion attacks depending on whether the adversary targets the learning process or the deployed model, as shown in Figure 7.

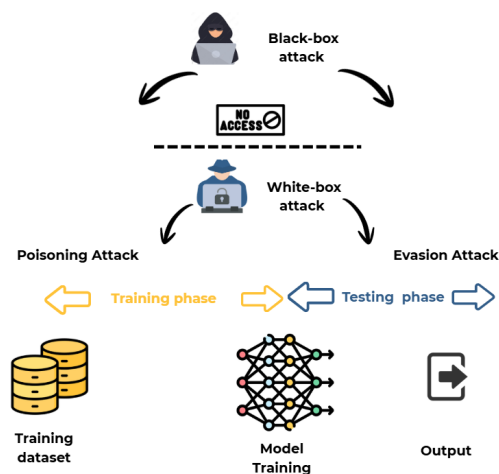


Figure 7. Knowledge-based attacks and attack-stage scenarios. Arrows indicate the progression from attacker knowledge level to the corresponding attack stage targeting either the training phase or the inference phase.

Poisoning attacks: Poisoning attacks manipulate the training data in order to corrupt decision boundaries, implant hidden behaviors, or reduce overall model reliability. Although this threat appears across many learning settings, recent studies emphasize its severity in safety-critical applications. For example, ref. [74] showed that clean-label poisoning can compromise YOLOv5-based maritime object detection systems by inducing misclassification of ship categories without altering ground-truth labels. This result is particularly concerning because it demonstrates that poisoning attacks may remain difficult to detect while still producing significant operational impact.

Evasion attacks: In contrast, evasion attacks are performed after training by crafting malicious inputs that exploit weaknesses in a fixed model during inference. This remains one of the most practical and widely studied attack scenarios. In malware security, reinforcement-learning-based strategies were shown to bypass multiple detection systems, motivating patching frameworks such as MalVPatch to restore resilience [75]. Similar concerns were reported in later work where DRL-generated malware variants, such as MalDQN, were used both to evaluate detector weaknesses and improve robustness through adversarial retraining [76].

Beyond malware analysis, evasion threats also affect networked and web-based systems. For instance, ref. [77] proposed mechanisms for detecting adversarial examples targeting NIDS systems, while ref. [78] examined attacks against web-phishing classifiers and highlighted the benefits of feature hardening and ensemble defenses. In ransomware detection, ref. [79] demonstrated that combining static and dynamic behavioral features with ensemble learning can significantly strengthen resistance to adversarial manipulation.

3.1.3. Specificity

This dimension classifies adversarial attacks according to the attacker's objective, namely whether the goal is to force a model toward a chosen label or simply to trigger any incorrect prediction.

Targeted attacks: In targeted attacks, the adversary seeks a specific output class, which provides greater control over the final decision but usually requires more precise optimiza-

tion. Considerable research has focused on improving the practicality and transferability of this setting. For example, ref. [80] proposed an efficient iterative framework that combines gradient-based optimization with FGSM-style updates to generate targeted universal adversarial perturbations capable of steering multiple inputs toward the same label. In cybersecurity, targeted manipulation has also been demonstrated against malware classifiers, where functionality-preserving modifications to PE files were used to redirect predictions toward attacker-selected malware families [81].

The challenge of transferring targeted attacks across unseen models has also received increasing attention. To address this issue, ref. [82] introduced a dynamic loss strategy that adaptively adjusts gradient directions in order to improve cross-model transferability. Similar progress has been reported in remote sensing and radar applications. In particular, ref. [83] designed a transferable targeted attack for SAR image recognition by combining feature alignment with gradient optimization, achieving strong black-box performance. Beyond effectiveness alone, realism and imperceptibility are also important considerations. In this regard, ref. [73] modeled targeted universal perturbations as structured noise, producing attacks that are visually more natural while maintaining strong success rates. Interestingly, targeted perturbations can also support defense objectives: ref. [84] showed that carefully crafted targeted universal perturbations can expose abnormal model behavior and help to identify backdoored networks.

Untargeted attacks: By contrast, untargeted attacks do not require a specific destination class. Their objective is simply to cause any incorrect prediction, making them generally easier to generate and often less computationally demanding. As demonstrated in [46], untargeted white-box attacks remain highly effective against malware detection systems, while adversarial retraining can significantly improve robustness. Similarly, ref. [45] showed that untargeted attacks such as FGSM, JSMA, PGD, and C&W can severely degrade intrusion detection performance if adequate defenses are not in place.

Overall, targeted attacks offer precise adversarial control but are typically harder to optimize, whereas untargeted attacks are easier to deploy and remain highly practical in real-world threat scenarios.

3.1.4. Attack Procedure

Another important distinction concerns how the perturbation is generated.

One-shot attacks: One-shot (single-step) attacks compute the perturbation in a single gradient update. Their low computational cost makes them attractive for real-time or large-scale scenarios, with FGSM being the most representative example [85].

Iterative attacks: Iterative attacks refine the perturbation across multiple optimization steps, usually leading to stronger and more reliable adversarial examples. Although more computationally expensive, they often outperform single-step methods in attack success and stealth. For instance, the iterative framework proposed in [80] demonstrated the effectiveness of repeated optimization for targeted universal attacks. Likewise, ref. [46] reported that multi-step methods such as PGD and BIM can be more damaging than FGSM in malware detection settings, highlighting the practical importance of iterative optimization.

3.1.5. Perturbation Scope

This dimension distinguishes whether the perturbation is crafted for a specific sample or designed to generalize across many inputs.

Input-specific perturbations: These perturbations are generated individually for each sample and are typically optimized to remain minimal while exploiting local model vulner-

abilities. Because they are tailored to a single instance, they often achieve high effectiveness with limited visible distortion.

Universal adversarial perturbations: In contrast, universal adversarial perturbations (UAPs) consist of a single reusable noise pattern capable of misleading many different inputs. Their broad applicability makes them particularly concerning in practical deployments. For example, ref. [80] demonstrated targeted universal perturbations that drive multiple samples toward the same chosen class. In another domain, ref. [86] extended the UAP concept to network traffic classification, showing that a single perturbation can mislead traffic detectors and even transfer across different models and feature spaces.

3.2. Defense Techniques

Recent studies increasingly indicate that adversarial robustness cannot be achieved through a single universal mechanism. Instead, we observe a gradual shift toward complementary defense paradigms that combine robust optimization, feature regularization, anomaly detection, model diversity, generative reconstruction, and certified guarantees (Table 5). This evolution is particularly visible in computer vision and cybersecurity, where adversarial failures may lead not only to prediction errors but also to operational, financial, or safety consequences.

A. Adversarial Training.

Adversarial training (AT) remains the most established empirical defense because it directly optimizes the model on adversarially perturbed samples. Across cybersecurity applications, multiple studies reported consistent gains when AT was integrated into intrusion detection and threat-monitoring pipelines [72,87,88]. Similar benefits were observed in communication systems, where AT improved robustness against signal-classification evasion under realistic wireless conditions [89]. In computer vision, newer variants such as asynchronous dual learners and Fast-M were introduced to reduce the computational burden of classical AT while maintaining competitive robustness [90,91]. Additional work also applied AT successfully to image quality assessment models exposed to multiple perturbation types [92]. More recent work extended AT to backdoor mitigation, where spatial adversarial training was shown to resist patch-based triggers [93], and to model compression, where on-manifold adversarial training preserved robustness in compressed networks suited to embedded deployment [94]. In intrusion detection, realistic constraint-aware adversarial training was applied to tabular network data to generate valid perturbations for both attack and defense [95], while adversarial retraining with feature fusion improved resilience in medical imaging [96]. The limitation of adversarial training is structural rather than incidental. Because the model is optimized on a specific family of perturbations, it tends to overfit to the threat seen during training and generalizes poorly to attacks it was not exposed to. The efficiency-oriented variants that reduce the cost of classical AT [90,91] do not remove this dependency since a faster training loop still learns from the same narrow perturbation set. This is compounded by a robustness–accuracy trade-off that is now well documented across the reviewed studies, where gains under attack are paid for by measurable drops on clean data. In sequential settings the problem deepens as defenses calibrated for single-step perturbations offer no guarantee once errors compound across time steps, which is why AT does not transfer cleanly from supervised classifiers to reinforcement learning agents (Section 4.1.2).

B. Regularization and Representation Smoothing.

A second family of defenses improves robustness indirectly by constraining the learned representation rather than explicitly generating adversarial examples. EMSF²Net, for example, used multi-stage feature fusion and regularization to strengthen CNN robustness under white-box attacks [44]. Other works introduced normalization with margin bal-

ancing [97], between-class mixing for feature regularization [98], and wavelet-domain smoothing combined with AT [99]. More recent comparative studies across CNNs and Vision Transformers showed that regularization-based defenses can sometimes rival AT depending on architecture and threat assumptions [100]. Gradient-regularized training was also shown to improve robustness against cropping-based attacks while preserving clean accuracy [101]. These approaches share a common ceiling. By smoothing the learned representation rather than confronting the attack directly, they raise the cost of finding an adversarial example without eliminating its existence, so a sufficiently strong or adaptive optimizer still succeeds. The comparative studies that report regularization rivaling AT [100] do so under fixed threat assumptions, and that parity tends to disappear once the attacker adapts to the specific smoothing being applied. Several of these methods also risk a subtler failure, where the regularizer flattens local gradients enough to obstruct gradient-based attacks during evaluation while leaving the model vulnerable to transfer- and gradient-free attacks. This form of apparent robustness through gradient obfuscation has repeatedly been shown to overestimate true robustness.

C. Detection and Monitoring Defenses.

Instead of modifying the classifier itself, another line of research attempts to detect adversarial behavior before harmful decisions are made. Local gradient analysis achieved over 90% detection accuracy on CIFAR-10 and MNIST without retraining the underlying model [102]. In cybersecurity, blockchain-assisted federated IDS frameworks combined with statistical anomaly detection reached approximately 96% accuracy in identifying adversarial traffic [103]. Symmetric GAN architectures were also proposed to detect adversarial DDoS attacks and outperformed several baseline ML methods [104]. Related work extended detection to control-flow-graph representations of malware, where an autoencoder-based detector combined with randomized feature extraction identified adversarial examples with high accuracy [105], and to image recognition, where multilayer filtering suppressed adversarial noise before classification [106]. Detection reframes the problem rather than solving it, and it inherits a cat-and-mouse dynamic that favors the attacker. A detector is itself a classifier, so an adversary with knowledge of it can craft inputs that fool the target model and evade the detector at the same time, which collapses the high accuracy these methods report under a static evaluation [102,103]. The reported figures are therefore best read as performance against non-adaptive attacks, and they say little about an adversary that optimizes against the full pipeline. This ceiling is the same one identified for XAI-based detection in Section 5, where explanation-based detectors are defeated by inputs that simultaneously satisfy the classifier and produce a normal-looking explanation.

D. Generative Purification and Reconstruction.

Generative models have increasingly been used to reconstruct clean inputs before final classification. In intrusion detection, GAN-based traffic purification reduced adversarial success by 99% while improving IDS accuracy by 37% [107]. Hybrid GAN + AT strategies also improved robustness by 20–30% across several datasets [108]. In face recognition, GAN-assisted defenses increased resilience against patch attacks by 18–25% [109]. Furthermore, training with GAN-generated adversarial examples yielded competitive robustness while reducing some of the optimization cost associated with standard AT [110]. In ransomware detection, GAN-based adversarial training of a sequential classifier improved robustness against Carlini–Wagner-style evasion while maintaining detection accuracy [111]. The strong recovery rates reported for generative purification [107–109] come with two structural costs. First, the purifier is an additional model in the pipeline, and it becomes a new attack surface that an adversary can target directly, so the defense can be bypassed by crafting inputs that survive reconstruction. Second, reconstruction is lossy, and pushing inputs back toward the training distribution introduces distortion on clean data that lowers

baseline accuracy, which is the mirror image of the robustness–accuracy trade-off seen in adversarial training. Purification also assumes that adversarial and clean inputs are separable in the generator’s latent space, an assumption that weakens as attacks are optimized to lie on the data manifold rather than off it.

E. Ensemble Learning.

Ensemble methods attempt to reduce vulnerability by combining multiple learners whose failure modes are less correlated. This strategy has shown encouraging results in malware detection, ransomware analysis, and network defense [79,112,113]. Particularly strong improvements were reported when ensemble learning was combined with AT, raising the F1-score of an intrusion detector from 0.29 to 0.91 under PGD attacks [114]. Subsequent studies reinforced this direction. Diversity-oriented ensembles that explicitly decorrelate member predictions improved robustness on large-scale benchmarks [115], and cost-efficient ensemble training reduced the overhead usually associated with the approach [116]. In network intrusion detection, multi-model and multi-view ensembles improved resilience to adversarial and distribution-shift conditions [117,118], while verification-based ensembles guarded against both adversarial examples and out-of-distribution inputs [119]. A general framework combining robustness assessment with defense also confirmed ensembling as a practical countermeasure for NIDS [120]. The premise of ensemble defense is that combining learners with uncorrelated failure modes reduces shared vulnerability, and this premise is exactly where the approach is fragile. Adversarial examples transfer well between models that solve the same task, so members trained on the same data and objective tend to share the very weaknesses the ensemble was meant to average out, and diversity that is not explicitly enforced against transferable perturbations provides little protection. The dramatic improvement reported when ensembles are paired with adversarial training [114] is telling since it suggests the robustness comes largely from the AT component rather than from diversity alone. Ensembles also multiply inference cost, which conflicts with the latency and throughput constraints of the operational settings where they would be deployed.

F. Distillation.

Distillation-based defenses seek to transfer knowledge from a teacher model into a smoother student model that is less sensitive to perturbations. Recent work combined self-distillation with adversarial learning for image super-resolution, improving both PSNR and perceptual robustness [121]. Similarly, preprocessing combined with distillation increased defense success from 41% to approximately 72.7% in image classification tasks [122]. Feature-squeezing mechanisms follow a related philosophy by suppressing irrelevant input information. Self-feature-squeezing layers, for instance, improved resistance to FGSM and JSMA with only minor clean-accuracy loss [123]. Distillation illustrates how quickly an empirical defense can be overturned. Defensive distillation was influential precisely because it appeared to harden models until it was shown that the smoothing it produces masks gradients rather than removing vulnerability, so attacks that do not rely on the target’s gradients defeat it. The reviewed variants [121,122] improve on the original but operate on the same principle of transferring a smoother decision surface, which limits how much genuine robustness they can add against a determined adaptive adversary. Feature-squeezing methods [123] face a related tension since aggressive squeezing discards information that both the attacker and the legitimate task depend on, forcing a trade-off between detection strength and clean accuracy.

G. Certified Robustness.

A more rigorous direction seeks formal guarantees rather than empirical robustness alone. In malware detection, randomized and de-randomized smoothing methods recently outperformed prior smoothing defenses while offering certifiable robustness guar-

antees [124]. Beyond malware, certified defenses based on randomized smoothing were applied to modulation classification in wireless systems [125] and extended to reinforcement learning, where certifiable bounds on state-action values provided guarantees under adversarial state uncertainty [126]. In computer vision, robustness on both in- and out-distribution inputs was shown to improve certified guarantees together with explanation quality [127], and ablation-based smoothing offered practical certification for deep malware detectors [128]. Certified methods provide what every empirical defense lacks, a provable guarantee, but that guarantee is narrow in ways that matter for deployment. The certified radius is typically small, so the guarantee holds only against perturbations well below the size a realistic attacker may use, and the bounds loosen quickly as input dimensionality grows. Certification also scales poorly since methods based on randomized smoothing require many forward passes per input, which is difficult to reconcile with real-time detection. The guarantee is defined for a specific perturbation model, usually an ℓ_p ball, and says nothing about semantic or functional perturbations of the kind that dominate cybersecurity, nor about common non-adversarial corruptions that have been shown to degrade certified models in practice. With only six studies in the reviewed corpus, certified robustness remains marginal relative to the operational demands placed on it.

Read together, these families reveal a recurring pattern rather than a set of isolated shortcomings. Each defense raises the cost of an attack within the threat model it was designed for, and each is undercut by an adversary that steps outside that model, whether by adapting to the defense, exploiting transfer, or targeting an auxiliary component the defense introduces. The empirical defenses trade clean accuracy for robustness against known attacks, while the one family offering formal guarantees pays for them in scope and scale. This is why no single mechanism has proven sufficient and why the most robust reported results consistently come from combinations, most often built around adversarial training. It also motivates the two directions examined in the remainder of this review, reinforcement learning as a way to make defenses adaptive rather than static and explainability as a way to audit why a defense succeeds or fails rather than only whether it does.

Table 5. Summary of defense studies grouped by defense family.

Ref.	Family	Mechanism/Approach	Domain
[87]	Adversarial Training	AT for intrusion detection pipelines	Cyber
[72]	Adversarial Training	AT for threat-monitoring	Cyber
[88]	Adversarial Training	AT-hardened IDS	Cyber
[89]	Adversarial Training	AT for wireless signal classification	Cyber
[90]	Adversarial Training	Asynchronous dual-learner AT	Vision
[91]	Adversarial Training	Fast-M efficient AT	Vision
[92]	Adversarial Training	AT for image quality assessment	Vision
[129]	Adversarial Training	Adversarial retraining of malware detectors	Cyber
[130]	Adversarial Training	Realistic PE-malware AT evaluation	Cyber
[131]	Adversarial Training	pix2pix-based AT for Android malware	Cyber
[95]	Adversarial Training	Constraint-aware AT for tabular NIDS	Cyber
[96]	Adversarial Training	AT + feature fusion (medical imaging)	Vision
[93]	Adversarial Training	Spatial AT against backdoor triggers	Vision
[94]	Adversarial Training	On-manifold AT for compressed models	Cyber
[44]	Regularization	Multi-stage feature fusion (EMSF ² Net)	Vision
[97]	Regularization	Normalization with margin balancing	Vision
[98]	Regularization	Between-class feature mixing	Vision
[99]	Regularization	Wavelet-domain smoothing + AT	Vision
[100]	Regularization	CNN vs. ViT regularization comparison	Vision
[101]	Regularization	Gradient-regularized training	Vision

Table 5. Cont.

Ref.	Family	Mechanism/Approach	Domain
[102]	Detection	Local gradient analysis detector	Vision
[103]	Detection	Blockchain-assisted federated IDS	Cyber
[104]	Detection	Symmetric GAN for DDoS detection	Cyber
[106]	Detection	Multilayer filtering defense	Vision
[105]	Detection	CFG autoencoder detector (Soteria)	Cyber
[107]	Generative Purification	GAN-based traffic purification	Cyber
[108]	Generative Purification	Hybrid GAN + AT	Vision
[109]	Generative Purification	GAN defense against patch attacks	Vision
[110]	Generative Purification	Training with GAN-generated examples	Vision
[111]	Generative Purification	GAN-based ransomware AT	Cyber
[79]	Ensemble	Static + dynamic ensemble (ransomware)	Cyber
[112]	Ensemble	Ensemble malware detection	Cyber
[113]	Ensemble	Ensemble network defense	Cyber
[114]	Ensemble	Ensemble + AT (PGD robustness)	Cyber
[117]	Ensemble	Multi-model NIDS ensemble	Cyber
[120]	Ensemble	Robustness assessment + ensemble (TIKI-TAKA)	Cyber
[115]	Ensemble	Diversity-oriented ensemble	Vision
[116]	Ensemble	Cost-efficient ensemble (FASTEN)	Vision
[119]	Ensemble	Verification-based ensemble (XEnsemble)	Vision
[118]	Ensemble	Multi-view fusion IDS	Cyber
[121]	Distillation	Self-distillation + adversarial learning	Vision
[122]	Distillation	Preprocessing + distillation	Vision
[123]	Distillation	Self-feature-squeezing layers	Vision
[124]	Certified Robustness	Randomized / de-randomized smoothing	Cyber
[128]	Certified Robustness	Ablation-based smoothing (malware)	Cyber
[125]	Certified Robustness	Filtered randomized smoothing	Cyber
[126]	Certified Robustness	Certifiable bounds for RL	Cyber
[127]	Certified Robustness	In/out-distribution robustness (RATIO)	Vision

4. Reinforcement Learning in Adversarial Machine Learning

4.1. Overview of Reinforcement Learning

4.1.1. RL Process

RL is a machine learning approach in which an agent learns optimal behavior through interaction with an environment, guided solely by scalar reward signals rather than labeled supervision [132,133]. Its formal foundation is the Markov Decision Process (MDP), defined by the tuple

$$(S, A, P, R, \gamma)$$

where:

- S : Set of environment states;
- A : Set of possible actions;
- $P(s'|s, a)$: Transition probability between states;
- $R(s, a)$: Reward function;
- γ : Discount factor that controls the importance of future rewards.

The objective is to find an optimal policy $\pi^*(a|s)$ that maximizes the expected cumulative discounted reward:

$$J(\pi) = \mathbb{E}_{\pi} \left[\sum_{t=0}^{\infty} \gamma^t r_t \right].$$

The agent interacts continuously with its environment to learn optimal behavior. At each step, the agent observes the current state, which represents the situation of the environment, and selects an action according to its policy, a strategy that guides decision-making. The environment responds by providing a new state and a reward, a signal indicating the quality of the agent's action. Using this feedback, the agent updates its policy to maximize cumulative expected rewards over time. We illustrate the process in Figure 8.

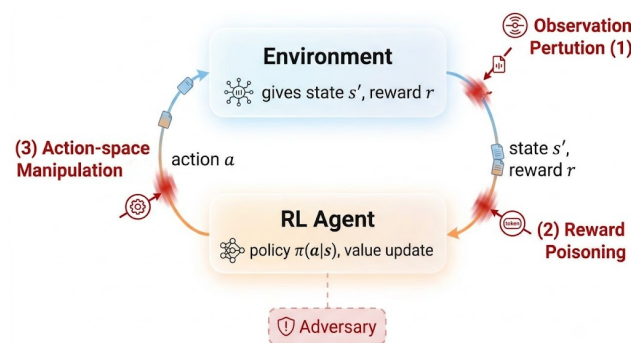


Figure 8. Reinforcement learning under adversarial conditions. The standard agent–environment loop (black) is exposed to three distinct attack surfaces: (1) observation perturbation corrupts the state signal, (2) reward poisoning manipulates the feedback signal, and (3) action-space manipulation alters the executed action.

In deep reinforcement learning (DRL), both the policy and the value function are parameterized by neural networks, enabling RL to scale to high-dimensional inputs such as raw image pixels and network traffic feature vectors [134]. Three algorithm families dominate the literature: value-based methods (such as DQN and its variants, which approximate $Q(s, a)$ and are particularly suited to discrete action spaces, such as network intrusion detection [135], PE binary modification [136], and sparse video-frame selection [137]). Policy-based methods, including PPO and TRPO, directly optimize $\pi(a | s)$ and effectively handle stochastic or continuous perturbation crafting [138,139], while actor–critic architectures, such as A3C, DDPG, and SAC, combine both approaches, reducing gradient variance at the cost of additional computational complexity [140,141]. Table 6 summarizes these three families and their relevance to AML.

Table 6. Families of RL algorithms and their relevance to AML.

Family	Algorithms	Mechanism	AML Relevance
Value-based	DQN, Double-DQN, Dueling-DQN	Learn $Q(s, a)$; policy derived greedily	Discrete attack spaces: PE evasion.
Policy-based	REINFORCE, PPO, TRPO	Directly optimize $\pi(a s)$; natural fit for continuous actions	Adversarial malware generation and continuous perturbation crafting in image classifiers.
Actor–Critic	A3C, DDPG, SAC, TD3	Separate actor and critic; lower variance than policy gradient alone	Attacks target actor decisions or critic estimates, while SAC/TD3 are studied for robustness

4.1.2. Vulnerabilities of RL Systems in Adversarial Settings

Three distinct attack surfaces have been identified against deployed RL agents, each corresponding to a different component of the MDP [142]. The most studied is observation perturbation, which Zhang et al. [143] formalized as the State-Adversarial MDP (SA-MDP). They demonstrated that bounded perturbations to input observations structurally analogous to FGSM on classifiers cause catastrophic policy failure in DRL agents trained on visual inputs, with attack effectiveness carrying over to perception systems in security-critical deployment scenarios. BARReL [144] confirmed this vulnerability in vision-based RL and showed that a bottleneck attention mechanism reduces its impact under bounded white-box perturbations, although not against adaptive adversaries. The second surface is reward poisoning, where Cai et al. [145] showed that an adversary exploiting RL exploration strategies can steer the agent toward a malicious policy with a surprisingly small fraction of corrupted reward signals, a disproportionate leverage that is absent in supervised learning. The sequential compounding of errors across time steps amplifies the effect of each corrupted signal. Han et al. [146] extended this finding to software-

defined network defense, showing that training-time poisoning remains effective even under partial observability of the environment. The third surface is action-space and environment manipulation. Elhami Fard and Selmic [147] demonstrated that time-delayed data transmission in multi-agent DRL systems allows an adversary to desynchronize agent observations and corrupt collaborative defense decisions, a vector that is realistic in distributed sensing architectures where network latency is unavoidable.

Across all three surfaces, a consistent finding is that standard AT, the dominant defense for supervised classifiers, does not transfer cleanly to RL [148] because perturbations compound over time steps, their downstream effects are disproportionate to their magnitude, and defenses calibrated for single-step attacks provide no guarantees in sequential settings.

4.2. Reinforcement Learning as an Attack Generation Tool

Framing adversarial example generation as a sequential decision problem is particularly effective in domains where perturbations must satisfy hard structural constraints that gradient-based methods cannot handle since many constraint-preserving operations are non-differentiable. This is most clearly illustrated in malware, where any modification to a PE binary must preserve both its maliciousness and executability. Song et al. [55] treated this as a multi-armed bandit problem (MAB-Malware), achieving evasion rates of 74–97% against ML-based detectors and 32–48% against commercial antivirus engines, often with a single-byte change. Fang et al. [71] used A3C (A3CMal) to produce targeted misclassification while preserving functionality, and Rathore et al. [75] demonstrated with PPO and TRPO agents that adversarial Android malware generated against one classifier transfers to independent classifiers at non-trivial rates, a practical concern for ensemble-based pipelines. Zhang et al. [149] extended the approach to GNN-based malware detectors that analyze binary call-graphs rather than raw bytes, formulating the attack as semantics-preserving instruction insertion and reporting evasion rates that substantially exceed three gradient-based baselines. The structured discrete nature of graph inputs is precisely where gradient-free RL exploration has the greatest advantage. Beyond evasion of a single classifier, Oh et al. [150] used DRL to simulate adaptive multi-step intrusion campaigns structured around the MITRE ATT&CK framework, modeling the sequential nature of real-world attacks rather than single-query evasion. Guan et al. [57] demonstrated the same RL evasion logic in web application security. Their SSQLi agent learns SQL injection mutation strategies and achieves bypass rates of 97.39% against ML- and DL-based web application firewalls without any gradient access. In computer vision, the constraint is geometric rather than structural. Song et al. [151] applied a self-attention RL policy (RLVS) to identify the minimal subset of video frames whose perturbation maximizes misclassification, achieving 83–92% fooling rates on UCF-101, HMDB-51, and Kinetics-400 while modifying only 3–4 frames per clip, a sparsity that gradient-based attacks cannot achieve on temporal inputs because frame-level importance is not directly accessible through differentiation. Yang et al. [152] used RL for adversarial patch placement (PatchAttack), framing the choice of patch location and texture as a policy optimization problem whose reward reflects misclassification under physical-world transformations. Recent work by [153] showed that randomized defenses, which add stochastic noise to model outputs to frustrate gradient estimation, are themselves exploitable by RL-based attackers because the RL agent learns from reward signals rather than gradients and is therefore indifferent to the noise. Several limitations are consistent across this body of work. Evaluations primarily target static analysis engines rather than dynamic detectors, action spaces require hand-engineering by domain experts, and Andrade et al. [154] found that ensembles of RL generators raise evasion rates substantially but at computational costs that have not been systematically addressed.

4.3. Reinforcement Learning as an Adaptive Defense Mechanism

The fundamental argument for RL-based defense is that conventional defenses are static: AT, input preprocessing, and certified robustness share one structural weakness, which is that, once deployed, they cannot change [28]. Every such technique produces a fixed policy, and an adaptive adversary that probes the defense boundary systematically will find its limits [29]. Wang et al. [155] showed that static defenses cannot adapt at test time, making them vulnerable to attackers that keep changing strategy. Tsingenopoulos et al. [156] confirmed this from a game-theoretic perspective. Self-adaptive adversaries that update their own policies through RL reliably defeat static defenses, including active detection mechanisms. RL addresses this directly by casting the defender as an agent whose reward is detection accuracy and whose environment implicitly contains the adversary, so the defense policy updates as the threat distribution shifts [157]. In practice, the transition from static to adaptive detection has been demonstrated across intrusion detection [158] and visual perception simultaneously, and the structural findings are consistent across both application areas. In network intrusion detection, Haydari et al. [159] showed that DRL-based detectors retain partial robustness under adversarial traffic manipulation when ensemble mechanisms are added, although the defenses were not designed for the adversarial RL case and offered no formal guarantees. Kim et al. [160] proposed a DRL-NIDS with a stacked LSTM that improved generalization to zero-day attack classes by training the agent against novel patterns not seen during initialization. Wang et al. [161] proposed IDNet, an instance-adaptive dynamic network that uses AT combined with curriculum learning to generate routing vectors that activate relevant detection paths according to traffic instance complexity, achieving competitive detection accuracy against known and unknown attacks while reducing inference cost relative to full-scale residual networks. Evaluation is conducted on multiple heterogeneous datasets and validated on resource-limited platforms, although the dynamic routing policy itself is not interpretable and the framework does not address adaptive adversaries that can probe the routing mechanism. Larriva-Novo et al. [162] integrated RL-based categorization with post hoc XAI, making the agent's detection decisions interpretable to human analysts, which is the only system in this survey that directly couples adaptive defense with principled explainability. Oh et al. [150] approached the problem from the attack side of the same co-evolutionary logic, generating MITRE ATT&CK-structured intrusion sequences with a DRL agent that adapts its trajectory to defender responses, producing campaigns that are harder to detect than fixed-sequence attacks and can stress-test adaptive defenders realistically. In visual perception, BARReL [144] combined AT with a bottleneck attention module that directs the agent's policy toward semantically stable image regions, improving robustness under both white-box observation perturbations and physical-world corruptions. The attention mechanism effectively narrows the exploitable input subspace without altering the policy architecture. REIN-EAD [163] took a structurally different approach by making defense active rather than reactive. Instead of classifying a fixed input under adversarial conditions, the RL policy controls sensor viewpoint, forcing the adversary to defeat both the classifier and the controller's choice of observation geometry.

Three findings cut across all the reviewed work (Table 7). RL-based defenses generalize well within their training distribution but poorly outside it. Robustness against unseen attack strategies depends entirely on how diverse the adversarial simulator was, a calibration problem none of the reviewed systems resolves. Every RL defense also inherits the vulnerability described in Section 4.2. The policy network is a neural network, and an attacker with query access can infer its properties and craft targeted counter-attacks. Finally, aside from Larriva-Novo et al. [162], none of the reviewed systems explains why

the agent made a given decision, a real barrier wherever auditability is required, and the gap that Section 5 addresses.

Table 7. Studies employing RL in AML.

Ref.	Year	Domain	How RL Is Applied	RL Algorithm	Dataset/Target	Key Result	Role
[71]	2021	Cybersecurity	A3CMal: targeted malware misclassification	A3C	Drebin	Misclassification while preserving functionality	Attack
[55]	2022	Cybersecurity	PE evasion as multi-armed bandit	MAB	EMBER, MalConv, AV	74–97% vs. ML and 32–48% vs. AV	Attack
[75]	2023	Cybersecurity	MalEAttack + MalVPatch defense	PPO/TRPO	Drebin, Android	96.87% fooling and defense reduces $\sim 3\times$	Att+Def
[149]	2022	Cybersecurity	Semantics-preserving GNN malware attack	RL+NOP	Windows malware GNN	Outperforms gradient baselines	Attack
[161]	2026	Cybersecurity	IDNet: dynamic routing via adversarial and curriculum learning based on traffic complexity	RL dynamic policy	Multiple IDS datasets	Competitive accuracy against known and unknown attacks	Defense
[150]	2024	Cybersecurity	Adaptive MITRE ATT&CK cyberattack simulation	DRL	Defense systems	Harder to detect than fixed attacks	Attack
[57]	2023	Cybersecurity	SQLi mutation learning to bypass WAFs	RL policy	ML/DL WAF	97.39% bypass success	Attack
[145]	2023	Cybersecurity	Reward poisoning against RL agents	RL auxiliary net (E-values)	RL cyber environments	Severe policy degradation	Attack
[151]	2025	Computer vision	Sparse video adversarial attack	DQN + attn.	UCF-101, HMDB-51	83–92% fooling rate	Attack
[152]	2020	Computer vision	PatchAttack: patch placement policy	RL policy	Image	Robust under physical transforms	Attack
[153]	2024	Computer vision	Adaptive RL breaks noise defenses	Adaptive RL	CIFAR-10	Randomized defenses fail	Attack
[162]	2024	Cybersecurity	IDS categorization + XAI	RL + XAI	IDS benchmarks	Better categorization + explainability	Defense
[143]	2020	Computer vision	SA-MDP observation attack model	SA-MDP	RL control + vision	Small perturbations break policies	Attack
[144]	2022	Computer vision	Attention bottleneck + AT	AT + attention	RL vision tasks	Improved robustness	Defense
[163]	2024	Computer vision	Active defense by sensor viewpoint control	RL + OAPA	3D face	Generalizes to unseen patches	Defense

5. Explainable Artificial Intelligence in Adversarial Machine Learning

5.1. Overview of Explainable Artificial Intelligence

XAI encompasses techniques that make the reasoning of ML models transparent and auditable, which is a practical necessity in intrusion detection, malware classification, and visual perception, where decisions carry operational and regulatory consequences [18,19]. The methods fall into four families. Intrinsic approaches use inherently interpretable architectures. Post hoc methods explain a trained black-box without modifying it, the canonical examples being LIME [164], SHAP [165], and Grad-CAM [166]. Model-specific methods are tied to a single architecture, and model-agnostic methods apply to any classifier regardless of internal structure.

In Figure 9, we present the process that begins when a machine learning model receives an input and produces a prediction. At this stage, the model simply delivers an outcome but without any explanation of how it reached that decision. Next, an XAI explainer such as LIME, SHAP, or Grad-CAM is applied to the model's output. The explainer analyzes the model's internal behavior or its reaction to different inputs and generates an interpretation that highlights the key elements that influenced the prediction. This explanation is then passed to a human expert, who examines it to understand why the model behaved the way it did. The expert evaluates whether the explanation is logical, whether the model can be trusted, and what decisions should follow. Each step naturally leads to the next: the model predicts, the explainer clarifies, and the human validates. In the context of this review, XAI matters for two structurally opposed reasons. It gives analysts an instrument for detecting adversarial behavior, and it is itself an attack surface that adversaries can subvert. That duality and the open problems it generates organize the subsections that follow.

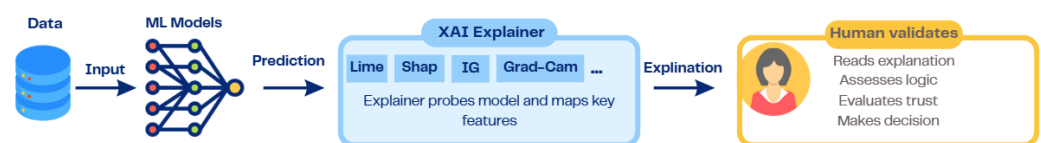


Figure 9. XAI process flow.

5.2. Explainable Artificial Intelligence as an Attack Surface

The assumption underlying most XAI deployments that explanations faithfully reflect the model's true reasoning does not hold in adversarial settings, and the evidence for this spans gradient-based attacks, scaffolding attacks, backdoors, and structural fragility. Shi et al. [167] proposed Ada-FGSM, an adaptive iterative attack that adjusts step size dynamically according to gradient information, demonstrating that the gradient landscape of a neural network is directly exploitable to increase attack effectiveness. The study is limited to image classifiers and does not address how the same gradient information could be used to manipulate or stabilize XAI attributions in security-critical deployment scenarios. Slack et al. [21] showed that a biased classifier wrapped in a shadow model can fool LIME and SHAP entirely. Perturbation queries are handled by the benign proxy, real inputs reach the biased model, and auditors see innocuous explanations. This exposes the central distinction between faithfulness, an explanation that reflects the model's actual computation, and plausibility, one that merely looks credible. The study was not evaluated against adaptive adversaries that know the scaffolding structure. Baniecki and Biecek [168] formalized the full attack surface with a unified taxonomy covering adversarial examples with misleading attributions, data poisoning invisible to post hoc explainers, model manipulation that shifts explanations without changing predictions, and backdoor attacks that produce both a target misclassification and a pre-specified attribution map. The survey covers attacks more thoroughly than defenses, and evaluation protocols for adversarial XAI robustness remain unstandardized. Noppel et al. [169] demonstrated explanation-aware backdoors that redirect Grad-CAM and gradient saliency maps to irrelevant regions when a trigger fires, making the attack invisible to analysts who rely on visual attribution. The attack requires model access at training time and has not been demonstrated against certified explanation defenses. Ghorbani et al. [20] showed that saliency maps, DeepLIFT, and integrated gradients produce dramatically different outputs for perceptually identical inputs with the same predicted label, a fragility that is structural rather than incidental. The study evaluates fragility but does not propose certified countermeasures, leaving the practical defense gap unaddressed. Henna and Rathnayake [170] demonstrated from the complementary direction that gradient-derived XAI signals, specifically Gradient-XAI,

Gradient-I-XAI, and I-Gradient-XAI, can be repurposed offensively: features identified as most influential by the explanation layer are precisely the targets that maximize adversarial misclassification under FGM–GAN hybrid perturbations, confirming that higher-fidelity explanations paradoxically increase an adversary’s attack surface. The finding is limited to DNS tabular data and has not been validated against defenses that deliberately obscure attribution information. Repetto et al. [171] provided the most systematic 2026 empirical evidence for structural XAI fragility, evaluating ten post hoc methods, including SmoothGrad, Grad-CAM++, AttnLRP, and LibraGrad, across eight CNN and transformer architectures on three medical imaging datasets under both adversarial and natural perturbations. Even the most robust methods exhibited substantial attribution instability under imperceptible adversarial noise, with no single method maintaining consistency across all the evaluation metrics, reinforcing that the fragility documented by Ghorbani et al. [20] extends to the latest generation of XAI techniques and is not an artifact of older methods.

5.3. XAI Under Operational Constraints in Cyber–Physical Systems

The deployment context matters as much as the method. Industrial control systems (ICSs), supervisory control and data acquisition (SCADA) environments, the Industrial Internet of Things (IIoT), and cyber–physical systems (CPSs) impose constraints that most reviewed XAI studies, developed on static image or tabular benchmarks, do not account for. Three constraints are particularly limiting.

Explanations in these settings must be produced in real time and under resource budgets. A perturbation-based method such as LIME or SHAP requires many model queries per input, which is difficult to reconcile with the latency requirements of an industrial controller or an in-vehicle detector. Recent work on IoT intrusion detection quantifies this tension, benchmarking SHAP and LIME under adversarial conditions and showing that explanation fidelity and stability degrade precisely when attacks are present, which is when a security operator most needs a reliable explanation [172]. Related evaluations on consumer IoT confirm that XAI can support adversarial detection in these environments but only when the explanation method is chosen with the operational profile in mind [173].

Explanations must also remain faithful under domain-specific perturbations. Adversarial examples in CPSs are constrained by physics and protocol semantics, so a valid perturbation must preserve the functional behavior of a control signal or a network flow rather than simply stay within an ℓ_p ball. An explanation method validated only against pixel-space attacks gives little assurance in this regime. The reviewed evidence on robustness-aware XAI in remote sensing and industrial monitoring [174,175] points toward explanation methods that are evaluated against realistic domain-constrained attacks, which we regard as a necessary direction for security-critical deployment.

Finally, explanations carry auditability and accountability obligations that are now partly legal. In sectors governed by safety regulation, and under frameworks such as the EU AI Act, a security decision that affects physical infrastructure may need to be accompanied by a per-decision explanation that is suitable for audit. None of the reviewed XAI-for-security systems was designed against this requirement, which we identify in Section 6 as one of the concrete gaps separating research prototypes from deployable systems.

5.4. Emerging Directions in Explainable AI for Adversarial Settings

The XAI methods discussed so far are predominantly post hoc attribution techniques applied to a single trained model. Three emerging directions extend explainability in ways that are directly relevant to adversarial robustness, and the reviewed literature has only begun to address them.

The first is causal explainability. Attribution methods such as SHAP and Grad-CAM report correlations between input features and outputs, which makes them fragile precisely because an adversary can shift these correlations without changing the underlying decision logic. Causal approaches aim to explain a prediction in terms of the features that genuinely drive it rather than those that merely co-vary with it, which in principle yields explanations that are harder to manipulate. This robustness is not automatic, and the reviewed evidence shows that adversarial perturbation can still distort explanation maps even after robust training [176], but causal framing offers a route toward attributions whose stability can be reasoned about rather than only measured empirically.

The second is explainability for structured and graph-based models. Most security-relevant data, including malware call graphs, network communication patterns, and control-flow representations, is naturally graph structured. Explaining decisions over such data requires attribution over nodes and edges rather than over pixels, which introduces attack surfaces that pixel-space methods do not face since an adversary can perturb graph topology as well as node features. This direction remains underexplored in the adversarial setting, and it represents one of the clearest gaps between the current XAI practice and the data structures that dominate cybersecurity.

The third is explainability for large language models. As LLMs are integrated into security workflows for tasks such as log analysis and threat-report generation, the question of why a model produced a given output becomes both harder and more consequential. Explanation techniques for these models are still immature, and their vulnerability to adversarial prompting and manipulated rationales is an open problem that the current corpus does not yet cover. We identify this as a priority direction rather than a settled area, and we return to it among the open challenges in Section 6. Table 8 summarizes the studies on XAI in AML reviewed in this section.

Table 8. Studies on XAI in AML.

Ref.	Year	Contribution	XAI Method	Dataset	Key Result	Limitation
[167]	2020	Adaptive FGSM with dynamic step size	Ada-FGSM	ImageNet, CIFAR	Higher ASR using gradients	No XAI attribution analysis
[21]	2020	Scaffolding hides biased classifier from audits	LIME, SHAP	COMPAS	Plausible but unfaithful explanations	No adaptive defense study
[169]	2023	Explanation-aware backdoors mislead Grad-CAM	Grad-CAM	CIFAR-10, malware	Controls prediction and explanation jointly	Requires training-time access
[168]	2024	Taxonomy of attacks against XAI	SHAP, LIME, IG	Multiple	Highlights faithfulness/stability gap	Focuses more on attacks than defenses
[170]	2026	FGM-GAN hybrid adversarial learning for DNS classification with multi-level XAI	Gradient-XAI, Gradient-I-XAI, I-Gradient-XAI	CIC-BELL-DNS-2021	FGM-GAN outperforms single-method baselines in transferability. I-Gradient-XAI highest explanation stability	Single DNS dataset
[177]	2025	Six attacks against XAI in cybersecurity	SHAP, LIME, IG	Phishing, IDS, fraud	Audit layer bypassed in all cases	No certified defense proposed
[178]	2022	AT + XAI for DL-NIDS	SHAP, Grad-CAM	CICIDS	Robustness and explainability coexist	No adaptive XAI-aware attacks
[174]	2022	LLM + SVDD + XAI for manipulation detection	LLM, SVDD, XAI	DNS, platooning	XAI reveals hidden anomalies	Limited dataset diversity

Table 8. Cont.

Ref.	Year	Contribution	XAI Method	Dataset	Key Result	Limitation
[179]	2024	Fuzzy XAI for VANET adversarial detection	SHAP, fuzzy DM	VANET traffic	High explainability + robustness	No second-order attack evaluation
[180]	2025	Detects adversarial word substitutions	Integrated Gradients	Text classifiers	Attribution deviation flags attacks	Fails if attribution preserved
[181]	2026	XAI-enhanced adversarial IDS for CIIoT/IIoT edge	SHAP (DeepExplainer)	Edge-IIoTset, CIC-IoT2023	99%+ detection under Auto-PGD, CW, DeepFool. 32.1 ms latency. SHAP fidelity 96.4%	SHAP deferred to post-inference. No test against attribution-preserving evasions
[162]	2024	RL-based IDS with interpretable decisions	Post hoc XAI	IDS benchmarks	Adaptive defense + explainability	Limited step-level policy insight
[182]	2024	XAI-enabled Digital Twin monitoring	XAI-enabled DT	Smart healthcare	Detects temporal manipulation patterns	Twin model itself unprotected
[183]	2024	Attribution fingerprinting for IoT IDS	SHAP DeepExplainer	IoT benchmark	Better than AT baseline	Single dataset only
[173]	2024	Real-time SHAP anomaly detector	SHAP	CIoT	96% adversarial detection	Assumes white-box SHAP access
[184]	2025	Compare XAI robustness for NIDS	LIME, SHAP, IG, DeepLIFT	UNSW-NB15	DeepLIFT most faithful	No certified robustness
[171]	2026	XAI robustness evaluation under adversarial and natural perturbations across 10 methods, 8 architectures, 3 medical datasets	SmoothGrad, Grad-CAM++, AttnLRP, LibraGrad, IG, DeepLIFT	DermaMNIST, PneumoniaMNIST, OctMNIST	SmoothGrad most robust for CNNs, AttnLRP for transformers	Medical imaging only. No certified defense proposed

6. Discussion and Open Challenges

6.1. Why RL and XAI Must Be Studied Together in AML

None of the three fields alone gives a complete picture of secure AI. AML treats attacks and defenses as fixed confrontations between static agents. RL produces adaptive policies but ones that are opaque to their designers. XAI makes decisions inspectable but is itself exploitable and cannot certify the faithfulness of its own outputs. Their intersection is not a theoretical convenience; it is where the real problems live. An adaptive RL defense that cannot explain its decisions cannot be deployed in any jurisdiction that has adopted the EU AI Act or GDPR, where auditability is mandatory. A static XAI detector that does not update its expected attribution distribution as the attacker adapts will be defeated once the adversary characterizes it. An RL policy whose explanation is available to the adversary creates a direct roadmap for crafting perturbations that exploit its most sensitive states. Static evaluation, fixed attack, and fixed defense tell almost nothing about real-world security because the operational threat model is always an informed adaptive adversary. Only a framework that treats attack, defense, and explanation as co-evolving processes produces results that hold under deployment conditions.

6.2. Limitations of the Reviewed Work

Five structural limitations recur across the reviewed studies. Benchmark narrowness: cybersecurity evaluations concentrate on CICIDS-2017/2018, UNSW-NB15, NSL-KDD, and EMBER and computer vision evaluations on CIFAR-10 and ImageNet datasets that are well understood by researchers and therefore not constraining for adversaries. Non-adaptive adversary assumption: defenses are routinely tested against fixed attack methods; the adaptive adversary that knows the defense of the realistic threat model is almost never eval-

uated. Simulator calibration in RL defense: the quality of an RL-based defense is bounded by the diversity of its adversarial simulator, yet no reviewed system provides a principled method for calibrating it. XAI faithfulness under attack: no reviewed XAI detection system certifies that its attribution properties are maintained under adversarial perturbation despite documented instability. Computational feasibility: Shapley values at line rate in a high-throughput NIDS remain infeasible without quality-degrading approximations. The overhead of the most faithful alternative, DeepLIFT, has not been characterized at operational throughput.

6.3. Open Challenges and Future Directions

Six problems remain unresolved across the reviewed literature. Each is stated with a concrete forward direction rather than a generic call for more research.

- **Certified robustness and real-time XAI in one pipeline:** Certified defenses guarantee stable classifier behavior under bounded perturbations, and XAI-based detectors generate interpretable anomaly signals, yet the two have never appeared in the same pipeline. The gap is not merely architectural but reflects a deeper unresolved question about whether formal guarantees and human-readable explanations can be reconciled at operational throughput. A concrete path forward is to extend certified explanation robustness using interval bound propagation for SHAP or DeepLIFT, then integrate the certified explainer directly into an operational NIDS, treating the attribution certificate as a first-class output alongside the prediction.
- **Attribution stability as a defense layer:** Explanation stability under adversarial perturbation is routinely assumed rather than verified. No reviewed system checks whether a small change to the input would cause a large change to the attribution before acting on that attribution, which means the detection logic may be trusting explanations that are in practice arbitrary. A local Lipschitz estimate of the explainer computed per input requires only one additional forward pass and would flag unstable attributions as potentially adversarial without adding meaningful latency.
- **Second-order attacks on XAI-based detection:** Detection pipelines that use attribution vectors as anomaly signals create a second-order attack surface. An adversary that knows the mechanism can craft inputs that simultaneously fool the classifier and produce a normal-looking attribution, bypassing both layers. The natural countermeasure is a multi-explainer consistency metric requiring the adversary to evade SHAP, DeepLIFT, and Grad-CAM simultaneously, a substantially harder task that raises the cost of a successful attack without requiring a separate detector.
- **Adaptive adversary evaluation as standard protocol:** Static evaluation against fixed attack strategies remains the dominant benchmarking practice even though the threat models it produces have known exploits. Adopting co-evolutionary RL evaluation, where attacker and defender agents are trained jointly, the standard robustness benchmark would replace point-in-time accuracy scores with equilibrium behavior under mutual adaptation, which is a far more honest measure of how a deployed system would actually perform.
- **Interpretable RL defense for regulated deployment:** RL-based defense systems learn policies that work but almost never explain why any given action was taken. In regulated environments, that is a deployment barrier, not a minor limitation. Applying SHAP-based policy attribution to explain each agent action from its state features at every time step is computationally tractable when supported by approximation or recurrent-state caching, and it would close the only gap that currently prevents adaptive RL defense from satisfying auditability requirements in practice.

- Federated AML and XAI: Recent work has begun to connect pairs of these properties in federated settings. Rezaei et al. [185] introduced Adapt-LFA, an adaptive gradient-guided label flipping attack against federated learning-based intrusion detection in IoT, showing that a poisoning adversary can degrade CNN and RNN detectors trained on CSE-CICIDS2018 and CICIoV2024 while remaining difficult to detect, which demonstrates that the federated aggregation process itself constitutes an attack surface. On the defensive side, Taheri et al. [186] integrated SHAP into a federated intrusion detection framework for connected vehicles, using attribution analysis both to interpret model decisions and help to identify anomalous updates originating from malicious clients. Each of these studies addresses two of the three properties that real-world deployment increasingly requires at once. Among the studies reviewed, however, none jointly addresses federated learning, adversarial robustness, and explainability. Evaluating federated SHAP approximations under adversarial poisoning conditions of the kind demonstrated by Adapt-LFA therefore represents a concrete and practically urgent research direction, particularly given the explainability obligations now mandated under the EU AI Act.

7. Conclusions

We examined AML, RL, and XAI not as parallel research threads but as an interdependent problem space, and the central finding is that each field, studied in isolation, produces defenses and evaluation results that break down the moment the other two are brought into the picture. Static defenses, which dominate the AML literature, fail against adaptive adversaries. RL-based defenses cannot be deployed in regulated environments without principled explanations of their decisions. XAI tools used for detection can themselves be subverted, with explanations made to look faithful while hiding a biased or backdoored model, and attribution methods are unstable under exactly the adversarial conditions they are meant to flag. The result is a circular dependency that is only visible when all three fields are studied together. The concrete findings support a gap between reported performance and operational security that is probably larger than the field currently acknowledges. RL-based attack frameworks consistently achieve evasion rates of 75–97% against ML-based detectors across both cybersecurity and computer vision benchmarks. XAI-based detection systems achieve promising accuracy under stationary conditions, but none has been evaluated against an adversary that knows the detection mechanism and adapts to it, and recent work shows that such adversaries can craft inputs that defeat both the classifier and the attribution check simultaneously. One tension runs through every section of this review without resolution: more robust models are harder to interpret, and more interpretable models are less robust. AT, certified defenses, and RL robustness mechanisms impose conflicting demands on the internal geometry of a model. The field does not yet have a principled framework for navigating this trade-off, and the papers that claim to address both robustness and interpretability simultaneously have not been evaluated under conditions that would reveal whether the claims hold. Three gaps remain. No existing system combines certified adversarial robustness with real-time XAI-based detection in a single evaluated pipeline. No RL-based defense, with one exception, provides a per-decision explanation that would satisfy a regulatory audit. We found that no study evaluates any reviewed system against an adaptive adversary that knows both the classifier and the explanation mechanism. These are not niche technical shortcomings. They are the difference between a research prototype and a deployable security system, and closing them is the work that remains.

Author Contributions: Conceptualization, H.O., F.K. and A.A.; methodology, H.O.; investigation, H.O.; writing—original draft preparation, H.O.; writing—review and editing, H.O., F.K. and A.A.; supervision, F.K. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The authors thank the reviewers for their constructive feedback, which improved the clarity and scope of this manuscript. No generative AI tools were used in the preparation of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AML	Adversarial Machine Learning
RL	Reinforcement Learning
XAI	Explainable Artificial Intelligence
DRL	Deep Reinforcement Learning
MDP	Markov Decision Process
IDS	Intrusion Detection System
NIDS	Network Intrusion Detection System
FGSM	Fast Gradient Sign Method
PGD	Projected Gradient Descent
BIM	Basic Iterative Method
C&W	Carlini and Wagner
JSMA	Jacobian-Based Saliency Map Attack
UAP	Universal Adversarial Perturbation
GAN	Generative Adversarial Network
AT	Adversarial Training
DQN	Deep Q-Network
PPO	Proximal Policy Optimization
TRPO	Trust Region Policy Optimization
A3C	Asynchronous Advantage Actor–Critic
DDPG	Deep Deterministic Policy Gradient
SAC	Soft Actor–Critic
LIME	Local Interpretable Model-Agnostic Explanations
SHAP	SHapley Additive exPlanations
PE	Portable Executable
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SA-MDP	State-Adversarial Markov Decision Process
CNN	Convolutional Neural Network
ASR	Attack Success Rate
TPR	True Positive Rate
WAF	Web Application Firewall
PSNR	Peak Signal-to-Noise Ratio
IoT	Internet of Things
VANET	Vehicular Ad Hoc Network
GDPR	General Data Protection Regulation

Appendix A. Evidence Matrix

Table A1. Evidence matrix mapping each of the 180 included empirical studies to its domain, threat model, attack/defense role, RL and XAI involvement, and the research question(s) it addresses.

Ref.	Domain	Threat Model	Role	RL	XAI	RQ
[42]	Vision	White-box	Attack	–	–	RQ1
[43]	Vision	White-box	Attack	–	–	RQ1
[44]	Vision	White-box	Defense	–	–	RQ1
[45]	Cyber	White-box	Both	–	–	RQ1
[46]	Cyber	White-box	Both	–	–	RQ1
[47]	Vision	Black-box	Attack	–	–	RQ1
[48]	Vision	Black-box	Attack	–	–	RQ1
[49]	Vision	Black-box	Attack	–	–	RQ1
[50]	Vision	Black-box	Attack	–	–	RQ1
[51]	Vision	Black-box	Attack	–	–	RQ1
[52]	Vision	Black-box	Attack	–	–	RQ1
[53]	Cyber	Black-box	Attack	–	–	RQ1
[54]	Vision	Black-box	Attack	–	–	RQ1
[55]	Cyber	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[56]	Cyber	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[57]	Cyber	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[58]	Cyber	Black-box	Attack	–	–	RQ1
[59]	Cyber	Black-box	Attack	–	–	RQ1
[60]	Cyber	Black-box	Attack	–	–	RQ1
[61]	Cyber	Black-box	Defense	–	–	RQ1
[62]	Vision	Black-box	Attack	–	–	RQ1
[63]	Vision	Black-box	Attack	–	–	RQ1
[64]	Vision	Black-box	Attack	–	–	RQ1
[65]	Vision	Black-box	Attack	–	–	RQ1
[66]	Cyber	Black-box	Attack	–	–	RQ1
[67]	Vision	White-box	Attack	–	–	RQ1
[68]	Vision	Black-box	Attack	–	–	RQ1
[69]	Vision	Black-box	Attack	–	–	RQ1
[70]	Vision	Black-box	Attack	–	–	RQ1
[74]	Vision	Training-time	Attack	–	–	RQ1
[75]	Cyber	Sequential (MDP)	Both	✓	–	RQ1/RQ2
[76]	Cyber	Sequential (MDP)	Both	✓	–	RQ1/RQ2
[77]	Cyber	Black-box	Defense	–	–	RQ1
[78]	Cyber	Mixed	Defense	–	–	RQ1
[79]	Cyber	Mixed	Defense	–	–	RQ1
[80]	Vision	White-box	Attack	–	–	RQ1
[81]	Cyber	White-box	Attack	–	–	RQ1
[82]	Vision	White-box	Attack	–	–	RQ1
[83]	Vision	White-box	Attack	–	–	RQ1
[73]	Vision	White-box	Attack	–	–	RQ1
[84]	Vision	Black-box	Defense	–	–	RQ1
[86]	Cyber	White-box	Attack	–	–	RQ1
[87]	Cyber	White-box	Both	–	–	RQ1
[72]	Cyber	White-box	Defense	–	–	RQ1
[88]	Cyber	White-box	Defense	–	–	RQ1
[89]	Cyber	White-box	Defense	–	–	RQ1
[90]	Vision	White-box	Defense	–	–	RQ1
[91]	Vision	White-box	Defense	–	–	RQ1
[92]	Vision	White-box	Defense	–	–	RQ1
[97]	Vision	White-box	Defense	–	–	RQ1
[98]	Vision	White-box	Defense	–	–	RQ1
[99]	Vision	White-box	Defense	–	–	RQ1
[100]	Vision	White-box	Defense	–	–	RQ1

Table A1. Cont.

Ref.	Domain	Threat Model	Role	RL	XAI	RQ
[101]	Vision	White-box	Defense	–	–	RQ1
[102]	Vision	Black-box	Defense	–	–	RQ1
[103]	Cyber	Black-box	Defense	–	–	RQ1
[104]	Cyber	Black-box	Defense	–	–	RQ1
[107]	Cyber	Black-box	Defense	–	–	RQ1
[108]	Vision	Black-box	Defense	–	–	RQ1
[109]	Vision	Black-box	Defense	–	–	RQ1
[110]	Vision	Black-box	Defense	–	–	RQ1
[112]	Cyber	Mixed	Defense	–	–	RQ1
[113]	Cyber	Mixed	Defense	–	–	RQ1
[114]	Cyber	Mixed	Defense	–	–	RQ1
[121]	Vision	White-box	Defense	–	–	RQ1
[122]	Vision	White-box	Defense	–	–	RQ1
[123]	Vision	White-box	Defense	–	–	RQ1
[124]	Cyber	Provable	Defense	–	–	RQ1
[187]	Vision	White-box	Defense	–	–	RQ1
[135]	Cyber	Sequential (MDP)	Defense	✓	–	RQ1/RQ2
[143]	Vision	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[144]	Vision	White-box	Defense	✓	–	RQ1/RQ2
[145]	Cyber	Training-time	Attack	✓	–	RQ1/RQ2
[146]	Cyber	Training-time	Attack	✓	–	RQ1/RQ2
[147]	Cyber	Varies	Attack	✓	–	RQ1/RQ2
[71]	Cyber	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[149]	Cyber	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[150]	Cyber	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[151]	Vision	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[152]	Vision	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[153]	Vision	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[154]	Cyber	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[155]	Vision	Varies	Defense	–	–	RQ1
[156]	Cyber	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[157]	Cyber	Sequential (MDP)	Defense	✓	–	RQ1/RQ2
[158]	Cyber	Sequential (MDP)	Defense	✓	–	RQ1/RQ2
[159]	Cyber	Sequential (MDP)	Both	✓	–	RQ1/RQ2
[160]	Cyber	Sequential (MDP)	Both	✓	–	RQ1/RQ2
[161]	Cyber	Sequential (MDP)	Defense	✓	–	RQ1/RQ2
[162]	Cyber	Sequential (MDP)	Defense	✓	✓	RQ1/RQ2/RQ3
[163]	Vision	Sequential (MDP)	Defense	✓	–	RQ1/RQ2
[167]	Vision	White-box	Attack	–	✓	RQ1/RQ3
[169]	Vision	Training-time	Attack	–	✓	RQ1/RQ3
[20]	Vision	White/Black-box	Attack	–	✓	RQ1/RQ3
[170]	Cyber	Black-box	Attack	–	✓	RQ1/RQ3
[171]	Vision	Post hoc	Evaluation	–	✓	RQ3/RQ4
[177]	Cyber	White/Black-box	Attack	–	✓	RQ1/RQ3
[178]	Cyber	White-box	Defense	–	✓	RQ1/RQ3
[174]	Cyber	Post hoc	Defense	–	✓	RQ1/RQ3
[179]	Cyber	Post hoc	Defense	–	✓	RQ1/RQ3
[181]	Cyber	Post hoc	Defense	–	✓	RQ1/RQ3
[182]	Cyber	White/Black-box	Attack	–	✓	RQ1/RQ3
[183]	Cyber	Post hoc	Defense	–	✓	RQ1/RQ3
[173]	Cyber	Post hoc	Defense	–	✓	RQ1/RQ3
[184]	Cyber	Post hoc	Evaluation	–	✓	RQ3/RQ4
[188]	Cyber	Sequential (MDP)	Both	✓	–	RQ1/RQ2
[189]	Cyber	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[190]	Cyber	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[191]	Cyber	Sequential (MDP)	Both	✓	–	RQ1/RQ2
[192]	Cyber	Sequential (MDP)	Attack	✓	–	RQ1/RQ2

Table A1. Cont.

Ref.	Domain	Threat Model	Role	RL	XAI	RQ
[193]	Cyber	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[194]	Cyber	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[195]	Cyber	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[196]	Cyber	Sequential (MDP)	Both	✓	–	RQ1/RQ2
[129]	Vision	White-box	Defense	–	–	RQ1
[197]	Cyber	Sequential (MDP)	Defense	✓	–	RQ1/RQ2
[130]	Cyber	White-box	Both	–	✓	RQ1/RQ3
[111]	Cyber	Black-box	Defense	–	–	RQ1
[131]	Cyber	White-box	Defense	–	–	RQ1
[95]	Cyber	White-box	Both	–	–	RQ1
[198]	Vision	White-box	Attack	–	–	RQ1
[199]	Vision	Post hoc	Evaluation	–	✓	RQ3/RQ4
[176]	Vision	Post hoc	Both	–	✓	RQ1/RQ3
[173]	Cyber	Post hoc	Defense	–	✓	RQ1/RQ3
[200]	Cyber	Black-box	Attack	–	–	RQ1
[201]	Cyber	White-box	Attack	–	–	RQ1
[202]	Vision	Post hoc	Defense	–	✓	RQ1/RQ3
[203]	Cyber	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[204]	Vision	Post hoc	Defense	–	✓	RQ1/RQ3
[205]	Vision	Post hoc	Defense	–	✓	RQ1/RQ3
[206]	Vision	White/Black-box	Evaluation	–	✓	RQ3/RQ4
[207]	Vision	White/Black-box	Evaluation	–	✓	RQ3/RQ4
[208]	Vision	White-box	Attack	–	–	RQ1
[209]	Vision	White-box	Attack	–	–	RQ1
[210]	Vision	Black-box	Attack	–	–	RQ1
[211]	Vision	White-box	Evaluation	–	–	RQ4
[106]	Vision	Black-box	Defense	–	–	RQ1
[96]	Vision	White-box	Both	–	–	RQ1
[212]	Vision	Provable	Evaluation	–	–	RQ4
[128]	Cyber	Provable	Defense	–	–	RQ1
[125]	Cyber	Provable	Defense	–	–	RQ1
[213]	Vision	White-box	Evaluation	–	–	RQ4
[117]	Cyber	Mixed	Defense	–	–	RQ1
[126]	Vision	Provable	Defense	✓	–	RQ1/RQ2
[214]	Vision	White/Black-box	Evaluation	–	✓	RQ3/RQ4
[215]	Vision	White/Black-box	Attack	–	✓	RQ1/RQ3
[216]	Vision	Post hoc	Defense	–	✓	RQ1/RQ3
[93]	Vision	White-box	Defense	–	–	RQ1
[217]	Vision	White-box	Defense	–	–	RQ1
[218]	Cyber	Sequential (MDP)	Defense	✓	–	RQ1/RQ2
[120]	Cyber	Mixed	Both	–	–	RQ1
[115]	Vision	Mixed	Defense	–	–	RQ1
[116]	Vision	Mixed	Defense	–	–	RQ1
[119]	Vision	Mixed	Defense	–	–	RQ1
[118]	Cyber	Mixed	Defense	–	–	RQ1
[172]	Cyber	Post hoc	Evaluation	–	✓	RQ3/RQ4
[174]	Cyber	Post hoc	Defense	–	✓	RQ1/RQ3
[175]	Vision	Post hoc	Defense	–	✓	RQ1/RQ3
[105]	Cyber	Black-box	Defense	–	–	RQ1
[219]	Cyber	Training-time	Attack	–	–	RQ1
[220]	Cyber	Sequential (MDP)	Both	✓	–	RQ1/RQ2
[221]	Vision	Sequential (MDP)	Defense	✓	–	RQ1/RQ2
[222]	Vision	Sequential (MDP)	Defense	✓	–	RQ1/RQ2
[223]	Cyber	Sequential (MDP)	Defense	✓	–	RQ1/RQ2
[224]	Cyber	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[225]	Vision	Sequential (MDP)	Defense	✓	–	RQ1/RQ2

Table A1. Cont.

Ref.	Domain	Threat Model	Role	RL	XAI	RQ
[157]	Cyber	Sequential (MDP)	Defense	✓	–	RQ1/RQ2
[226]	Cyber	Sequential (MDP)	Defense	✓	–	RQ1/RQ2
[227]	Cyber	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[151]	Vision	Sequential (MDP)	Attack	✓	–	RQ1/RQ2
[228]	Cyber	Sequential (MDP)	Defense	✓	–	RQ1/RQ2
[229]	Cyber	Sequential (MDP)	Defense	✓	–	RQ1/RQ2
[230]	Cyber	Sequential (MDP)	Defense	✓	–	RQ1/RQ2
[231]	Cyber	Sequential (MDP)	Defense	✓	–	RQ1/RQ2
[232]	Vision	Mixed	Evaluation	–	–	RQ4
[127]	Vision	Provable	Defense	–	–	RQ1
[233]	Cyber	White-box	Evaluation	–	–	RQ4
[234]	Vision	White-box	Attack	–	–	RQ1
[94]	Cyber	White-box	Defense	–	–	RQ1

✓ = attribute applies to the corresponding study. – = attribute not applicable or not reported.

References

- Yuan, X.; He, P.; Zhu, Q.; Li, X. Adversarial examples: Attacks and defenses for deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 2805–2824. [\[CrossRef\]](#)
- Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2015**, arXiv:1412.6572.
- Akhtar, N.; Mian, A. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1978–2001.
- Stanford Institute for Human-Centered Artificial Intelligence. *The 2026 AI Index Report*; Stanford University: Stanford, CA, USA, 2026.
- European Union Agency for Cybersecurity (ENISA). *ENISA Threat Landscape 2025*; European Union Agency for Cybersecurity: Heraklion, Greece, 2025.
- Biggio, B.; Roli, F. Wild patterns: Ten years after the rise of adversarial machine learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 3125–3140.
- Qiu, X.; He, X.; Liu, X. A survey of adversarial attacks on deep learning. *IEEE Access* **2019**, *7*, 117897–117918.
- Xu, K.; Qi, X.; Liu, L. Adversarial attacks and defenses in malware detection systems: A survey. *IEEE Access* **2020**, *8*, 199093–199112.
- Kumar, R.; Singh, A. Advances in adversarial machine learning for cybersecurity: Review and perspectives. *Comput. Secur.* **2023**, *120*, 103463.
- Papernot, N.; McDaniel, P.; Goodfellow, I. Distillation as a defense to adversarial perturbations against deep neural networks. *arXiv* **2016**, arXiv:1511.04508.
- Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv* **2018**, arXiv:1706.06083.
- Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; pp. 39–57. [\[CrossRef\]](#)
- Xu, K.; Liu, P. Adversarial defenses in deep learning: A survey. *Comput. Secur.* **2021**, *102*, 102367.
- Lin, Y.; Chen, Y. Tactics of adversarial attacks against reinforcement learning agents. *arXiv* **2017**, arXiv:1702.02284.
- Huang, L.; Wu, J.; Liu, W. Adversarial attacks and defenses in reinforcement learning: A survey. *Neural Netw.* **2022**, *150*, 345–362.
- Ilahi, S.; Mahmood, A. Reinforcement learning for security: A survey. *Expert Syst. Appl.* **2021**, *169*, 115054.
- Nguyen, T.; Tran, T. Adversarial reinforcement learning: Taxonomies, challenges, and future directions. *Artif. Intell.* **2024**, *329*, 104006.
- Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [\[CrossRef\]](#)
- Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bannetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [\[CrossRef\]](#)
- Ghorbani, A.; Abid, A.; Zou, J. Interpretation of neural networks is fragile. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 3681–3688.
- Slack, D.; Hilgard, S.; Jia, J.; Singh, S.; Lakkaraju, H. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*; Association for Computing Machinery: New York, NY, USA, 2020.

22. Boopathy, A.; Liu, S.; Zhang, G.; Liu, C.; Chen, P.-Y.; Chang, S.; Daniel, L. Proper network interpretability helps adversarial robustness in classification. In *Proceedings of the 37th International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2020; Volume 119, pp. 1014–1023.
23. Noack, A.; Ahern, I.; Dou, D.; Li, B. An empirical study on the relation between network interpretability and adversarial robustness. *SN Comput. Sci.* **2021**, *2*, 32. [\[CrossRef\]](#)
24. Mohammed, A. Artificial intelligence-powered cyber attacks: Adversarial machine learning. *Authorea Preprints* **2025**. [\[CrossRef\]](#)
25. Ling, X.; Wu, L.; Zhang, J.; Qu, Z. Adversarial attacks against Windows PE malware detection: A survey of the state-of-the-art. *Comput. Secur.* **2023**, *128*, 103134. [\[CrossRef\]](#)
26. McCarthy, A.; Ghadafi, E.; Andriotis, P.; Legg, P. Functionality-preserving adversarial machine learning for robust classification in cybersecurity and intrusion detection domains: A survey. *J. Cybersecur. Priv.* **2022**, *2*, 154–190. [\[CrossRef\]](#)
27. Zhou, S.; Liu, C.; Ye, D.; Zhu, T.; Zhou, W.; Yu, P.S. Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity. *ACM Comput. Surv.* **2023**, *55*, 1–39.
28. Macas, M.; Wu, C.; Fuertes, W. Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems. *Expert Syst. Appl.* **2024**, *238*, 122223. [\[CrossRef\]](#)
29. Alotaibi, A.; Rassam, M.A. Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense. *Future Internet* **2023**, *15*, 62. [\[CrossRef\]](#)
30. Rosenberg, I.; Shabtai, A.; Elovici, Y.; Rokach, L. Adversarial machine learning attacks and defense methods in the cyber security domain. *arXiv* **2020**, arXiv:2007.02407.
31. Martins, N.; Cruz, J.M.; Cruz, T.; Abreu, P.H. Adversarial machine learning applied to intrusion and malware scenarios: A systematic review. *IEEE Access* **2020**, *8*, 35403–35419. [\[CrossRef\]](#)
32. Vitorino, J.; Praça, I.; Maia, E. SoK: Realistic adversarial attacks and defenses for intelligent network intrusion detection. *Comput. Secur.* **2023**, *134*, 103433. [\[CrossRef\]](#)
33. Ke, H.; Xu, J.; Wang, Y.; Chen, H.; Shen, Z. Adversarial machine learning in cybersecurity: Attacks and defenses. *Int. J. Manag. Sci. Res.* **2025**, *8*, 26–33. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Khaleel, Y.L.; Habeeb, M.A.; Albahri, A.S.; Al-Quraishi, T.; Albahri, O.S.; Alamoodi, A.H. Network and cybersecurity applications of defense in adversarial attacks: A state-of-the-art using machine learning and deep learning methods. *J. Intell. Syst.* **2024**, *33*, 20240153. [\[CrossRef\]](#)
35. Rane, N.L.; Mallick, S.K.; Rane, J. *Artificial Intelligence and Machine Learning for Enhancing Resilience: Concepts, Applications, and Future Directions*; Deep Science Publishing: London, UK, 2025.
36. Liang, H.; He, E.; Zhao, Y.; Jia, Z.; Li, H. Adversarial attack and defense: A survey. *Electronics* **2022**, *11*, 1283. [\[CrossRef\]](#)
37. Stanly, H.; Shalinie, M.S.; Paul, R. A review of generative and non-generative adversarial attack on context-rich images. *Eng. Appl. Artif. Intell.* **2023**, *124*, 106595. [\[CrossRef\]](#)
38. Li, Y.; Xie, B.; Guo, S.; Yang, Y.; Xiao, B. A survey of robustness and safety of 2D and 3D deep learning models against adversarial attacks. *ACM Comput. Surv.* **2024**, *56*, 1–37. [\[CrossRef\]](#)
39. Akhtar, N.; Mian, A.; Kardan, N.; Shah, M. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access* **2021**, *9*, 155161–155196. [\[CrossRef\]](#)
40. Costa, J.C.; Roxo, T.; Proenca, H.; Inacio, P.R.M. How deep learning sees the world: A survey on adversarial attacks & defenses. *IEEE Access* **2024**, *12*, 61113–61136. [\[CrossRef\]](#)
41. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [\[CrossRef\]](#) [\[PubMed\]](#)
42. Aggarwal, S.; Mittal, A.; Aggarwal, S.; Singh, A.K. Dynamic programming-based white box adversarial attack for deep neural networks. *AI* **2024**, *5*, 1216–1234. [\[CrossRef\]](#)
43. Chen, T.; Ling, J.; Sun, Y. White-box content camouflage attacks against deep learning. *Comput. Secur.* **2022**, *117*, 102676. [\[CrossRef\]](#)
44. Zhang, J.; Maeda, K.; Ogawa, T.; Haseyama, M. Regularization meets enhanced multi-stage fusion features: Making CNN more robust against white-box adversarial attacks. *Sensors* **2022**, *22*, 5431. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Roshan, K.; Zafar, A.; Haque, S.B.U. Untargeted white-box adversarial attack with heuristic defence methods in real-time deep learning based network intrusion detection system. *Comput. Commun.* **2024**, *218*, 97–113. [\[CrossRef\]](#)
46. Ouazza, H.; Khennou, F.; Abdellaoui, A. Adversarial retraining and white-box attacks for robust malware detection. In *Proceedings of the 2025 13th International Symposium on Digital Forensics and Security (ISDFS)*, Boston, MA, USA, 24–25 April 2025; pp. 1–6.
47. Gao, J.; Zheng, K.; Wang, X.; Wu, C.; Wu, B. EFSAttack: Edge noise-constrained black-box attack using artificial fish swarm algorithm. *Electronics* **2024**, *13*, 2446. [\[CrossRef\]](#)
48. Cao, H.; Si, C.; Sun, Q.; Liu, Y.; Li, S.; Gope, P. ABCAttack: A gradient-free optimization black-box attack for fooling deep image classifiers. *Entropy* **2022**, *24*, 412. [\[CrossRef\]](#) [\[PubMed\]](#)

49. Lapid, R.; Haramaty, Z.; Sipper, M. An evolutionary, gradient-free, query-efficient, black-box algorithm for generating adversarial instances in deep convolutional neural networks. *Algorithms* **2022**, *15*, 407. [\[CrossRef\]](#)
50. Yu, M.; Sun, S. FE-DaST: Fast and effective data-free substitute training for black-box adversarial attacks. *Comput. Secur.* **2022**, *113*, 102555. [\[CrossRef\]](#)
51. Chen, Z.; Ding, J.; Wu, F.; Zhang, C.; Sun, Y.; Sun, J.; Liu, S.; Ji, Y. An optimized black-box adversarial simulator attack based on meta-learning. *Entropy* **2022**, *24*, 1377. [\[CrossRef\]](#) [\[PubMed\]](#)
52. Park, H.; Ryu, G.; Choi, D. Partial retraining substitute model for query-limited black-box attacks. *Appl. Sci.* **2020**, *10*, 7168. [\[CrossRef\]](#)
53. Zhao, S.; Li, J.; Wang, J.; Zhang, Z.; Zhu, L.; Zhang, Y. attackGAN: Adversarial attack against black-box IDS using generative adversarial networks. *Procedia Comput. Sci.* **2021**, *187*, 128–133. [\[CrossRef\]](#)
54. Wang, F.; Ma, Z.; Zhang, X.; Li, Q.; Wang, C. DDSG-GAN: Generative adversarial network with dual discriminators and single generator for black-box attacks. *Mathematics* **2023**, *11*, 1016. [\[CrossRef\]](#)
55. Song, W.; Li, X.; Afroz, S.; Garg, D.; Kuznetsov, D.; Yin, H. MAB-Malware: A reinforcement learning framework for blackbox generation of adversarial malware. In Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security (Asia CCS '22), Nagasaki, Japan, 30 May–3 June 2022; pp. 990–1003.
56. Wang, Q.; Yang, H.; Wu, G.; Choo, K.K.R.; Zhang, Z.; Miao, G.; Ren, Y. Black-box adversarial attacks on XSS attack detection model. *Comput. Secur.* **2022**, *113*, 102554. [\[CrossRef\]](#)
57. Guan, Y.; He, J.; Li, T.; Zhao, H.; Ma, B. SSQLi: A black-box adversarial attack method for SQL injection based on reinforcement learning. *Future Internet* **2023**, *15*, 133. [\[CrossRef\]](#)
58. Bundt, J.; Davinroy, M.; Agadakos, I.; Oprea, A.; Robertson, W. Black-box attacks against neural binary function detection. In Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses (RAID '23), Hong Kong, China, 16–18 October 2023; pp. 1–16.
59. Bostani, H.; Moonsamy, V. EvadeDroid: A practical evasion attack on machine learning for black-box Android malware detection. *Comput. Secur.* **2024**, *139*, 103676. [\[CrossRef\]](#)
60. Wu, S.; Xue, J.; Wang, Y.; Kong, Z. Black-box evasion attack method based on confidence score of benign samples. *Electronics* **2023**, *12*, 2346. [\[CrossRef\]](#)
61. Kong, Z.; Xue, J.; Liu, Z.; Wang, Y.; Han, W. MalDBA: Detection for query-based malware black-box adversarial attacks. *Electronics* **2023**, *12*, 1751. [\[CrossRef\]](#)
62. Mahmood, K.; Gurevin, D.; van Dijk, M.; Nguyen, P.H. Beware the black-box: On the robustness of recent defenses to adversarial examples. *Entropy* **2021**, *23*, 1359. [\[CrossRef\]](#) [\[PubMed\]](#)
63. Yang, Y.; Liu, Z.; Lei, Z.; Wu, S.; Chen, Y. Black-box boundary attack based on gradient optimization. *Electronics* **2024**, *13*, 1009. [\[CrossRef\]](#)
64. Xiang, F.; Xu, J.; Zhang, W.; Wang, W. A distributed biased boundary attack method in black-box attack. *Appl. Sci.* **2021**, *11*, 10479. [\[CrossRef\]](#)
65. Pomponi, J.; D'Antoni, D.; Alessandro, N.; Scardapane, S. Rearranging pixels is a powerful black-box attack for RGB and infrared deep learning models. *IEEE Access* **2023**, *11*, 11298–11306. [\[CrossRef\]](#)
66. Roshan, K.; Zafar, A. Black-box adversarial transferability: An empirical study in cybersecurity perspective. *Comput. Secur.* **2024**, *141*, 103853. [\[CrossRef\]](#)
67. Jo, J.; Kim, J.; Suh, Y.-J. Exploring public data vulnerabilities in semi-supervised learning models through gray-box adversarial attack. *Electronics* **2024**, *13*, 940. [\[CrossRef\]](#)
68. Zhang, Y.; Shin, S.-Y.; Tan, X.; Xiong, B. A self-adaptive approximated-gradient-simulation method for black-box adversarial sample generation. *Appl. Sci.* **2023**, *13*, 1298. [\[CrossRef\]](#)
69. Chitic, R.; Topal, A.O.; Leprevost, F. Empirical perturbation analysis of two adversarial attacks: Black box versus white box. *Appl. Sci.* **2022**, *12*, 7339. [\[CrossRef\]](#)
70. Bai, Y.; Wang, Y.; Zeng, Y.; Jiang, Y.; Xia, S.-T. Query efficient black-box adversarial attack on deep neural networks. *Pattern Recognit.* **2023**, *133*, 109037. [\[CrossRef\]](#)
71. Fang, Z.; Wang, J.; Geng, J.; Zhou, Y.; Kan, X. A3CMal: Generating adversarial samples to force targeted misclassification by reinforcement learning. *Appl. Soft Comput.* **2021**, *109*, 107505. [\[CrossRef\]](#)
72. Haroon, M.S.; Ali, H.M. Adversarial training against adversarial attacks for machine learning-based intrusion detection systems. *Comput. Mater. Contin.* **2022**, *73*, 3513–3527. [\[CrossRef\]](#)
73. Wang, H.; Cai, D.; Wang, L.; Xiong, Z. Crafting targeted universal adversarial perturbations: Considering images as noise. *IEEE Access* **2023**, *11*, 131651–131660. [\[CrossRef\]](#)
74. Lee, C.; Lee, S. Vulnerability of clean-label poisoning attack for object detection in maritime autonomous surface ships. *J. Mar. Sci. Eng.* **2023**, *11*, 1179. [\[CrossRef\]](#)

75. Rathore, H.; Nandanwar, A.; Sahay, S.K.; Sewak, M. Adversarial superiority in android malware detection: Lessons from reinforcement learning based evasion attacks and defenses. *Forensic Sci. Int. Digit. Investig.* **2023**, *44*, 301511. [[CrossRef](#)]
76. Rathore, H.; Sasan, A.; Sahay, S.K.; Sewak, M. Defending malware detection models against evasion based adversarial attacks. *Pattern Recognit. Lett.* **2022**, *164*, 119–125. [[CrossRef](#)]
77. Pawlicki, M.; Chorás, M.; Kozik, R. Defending network intrusion detection systems against adversarial evasion attacks. *Future Gener. Comput. Syst.* **2020**, *110*, 148–154. [[CrossRef](#)]
78. Pillai, M.J.; Remya, S.; Devika, V.; Ramasubbareddy, S.; Cho, Y. Evasion attacks and defense mechanisms for machine learning-based web phishing classifiers. *IEEE Access* **2024**, *12*, 19375–19387. [[CrossRef](#)]
79. Ahmed, U.; Lin, J.C.-W.; Srivastava, G. Mitigating adversarial evasion attacks of ransomware using ensemble learning. *Comput. Electr. Eng.* **2022**, *100*, 107903. [[CrossRef](#)]
80. Hirano, H.; Takemoto, K. Simple iterative method for generating targeted universal adversarial perturbations. *Algorithms* **2020**, *13*, 268. [[CrossRef](#)]
81. Kucuk, Y.; Yan, G. Deceiving portable executable malware classifiers into targeted misclassification with practical adversarial examples. In Proceedings of the 10th ACM Conference on Data and Application Security and Privacy (CODASPY '20), New Orleans, LA, USA, 16–18 March 2020; pp. 341–352.
82. Zhang, M.; Chen, Y.; Li, H.; Qian, C.; Kuang, X. Dynamic loss yielding more transferable targeted adversarial examples. *Neurocomputing* **2024**, *590*, 127754. [[CrossRef](#)]
83. Zheng, S.; Han, D.; Lu, C.; Hou, C.; Han, Y.; Hao, X.; Zhang, C. Transferable targeted adversarial attack on synthetic aperture radar (SAR) image recognition. *Remote Sens.* **2025**, *17*, 146. [[CrossRef](#)]
84. Qu, Y.; Huang, S.; Chen, X.; Wang, X.; Yao, Y. Detection of backdoor attacks using targeted universal adversarial perturbations for deep neural networks. *J. Syst. Softw.* **2024**, *207*, 111859. [[CrossRef](#)]
85. Buckman, J.; Roy, A.; Raffel, C.; Goodfellow, I. Thermometer encoding: One hot way to resist adversarial examples. In Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
86. Ding, R.; Sun, L.; Zang, W.; Dai, L.; Ding, Z.; Xu, B. Towards universal and transferable adversarial attacks against network traffic classification. *Comput. Netw.* **2024**, *254*, 110790. [[CrossRef](#)]
87. Mohammadian, H.; Ghorbani, A.A.; Lashkari, A.H. A gradient-based approach for adversarial attack on deep learning-based network intrusion detection systems. *Appl. Soft Comput.* **2023**, *137*, 110173. [[CrossRef](#)]
88. Rashid, M.M.; Kamruzzaman, J.; Hassan, M.M.; Imam, T.; Wibowo, S.; Gordon, S.; Fortino, G. Adversarial training for deep learning-based cyberattack detection in IoT-based smart city applications. *Comput. Secur.* **2022**, *120*, 102783. [[CrossRef](#)]
89. McClintick, K.W.; Harer, J.; Flowers, B.; Headley, W.C.; Wyglinski, A.M. Countering physical eavesdropper evasion with adversarial training. *IEEE Open J. Commun. Soc.* **2022**, *3*, 1820–1833. [[CrossRef](#)]
90. He, Z.; Liu, W.; Huang, Z.; Chen, Y.; Zhang, S. A3GT: An adaptive asynchronous generalized adversarial training method. *Electronics* **2024**, *13*, 4052. [[CrossRef](#)]
91. Ma, Y.; An, D.; Gu, Z.; Lin, J.; Liu, W. Fast-M adversarial training algorithm for deep neural networks. *Appl. Sci.* **2024**, *14*, 4607. [[CrossRef](#)]
92. Chistyakova, A.; Antsiferova, A.; Khrebtov, M.; Lavrushkin, S.; Arkhipenko, K.; Vatolin, D.; Turdakov, D. Increasing the robustness of image quality assessment models through adversarial training. *Technologies* **2024**, *12*, 220. [[CrossRef](#)]
93. Gao, Y.; Wu, D.; Zhang, J.; Gan, G.; Xia, S.-T.; Niu, G.; Sugiyama, M. On the effectiveness of adversarial training against backdoor attacks. In Proceedings of the 10th International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022.
94. Kwon, J.; Lee, S. Improving the robustness of model compression by on-manifold adversarial training. *Future Internet* **2021**, *13*, 300. [[CrossRef](#)]
95. Vitorino, J.; Oliveira, N.; Praça, I. Adaptive perturbation patterns: Realistic adversarial learning for robust intrusion detection. *Future Internet* **2022**, *14*, 108. [[CrossRef](#)]
96. Lal, S.; Rehman, S.U.; Shah, J.H.; Meraj, T.; Rauf, H.T.; Damaševičius, R.; Mohammed, M.A.; Abdulkareem, K.H. Adversarial attack and defence through adversarial training and feature fusion for diabetic retinopathy recognition. *Sensors* **2021**, *21*, 3922. [[CrossRef](#)] [[PubMed](#)]
97. Cheng, Z.; Zhu, F.; Zhang, X.-Y.; Liu, C.-L. Adversarial training with distribution normalization and margin balance. *Pattern Recognit.* **2023**, *136*, 109182. [[CrossRef](#)]
98. Wang, D.; Jin, W.; Wu, Y. Between-class adversarial training for improving adversarial robustness of image classification. *Sensors* **2023**, *23*, 3252. [[CrossRef](#)] [[PubMed](#)]
99. Yan, J.; Yin, H.; Zhao, Z.; Ge, W.; Zhang, H.; Rigoll, G. Wavelet regularization benefits adversarial training. *Inf. Sci.* **2023**, *649*, 119650. [[CrossRef](#)]
100. Dingeto, H.; Kim, J. Comparative study of adversarial defenses: Adversarial training and regularization in vision transformers and CNNs. *Electronics* **2024**, *13*, 2534. [[CrossRef](#)]

101. Yoshida, M.; Namura, H.; Okuda, M. Adversarial examples for image cropping: Gradient-based and Bayesian-optimized approaches for effective adversarial attack. *IEEE Access* **2024**, *12*, 86541–86552. [\[CrossRef\]](#)
102. Chen, R.; Jin, H.; Chen, J.; Zheng, H.; Zheng, S.; Yang, X.; Yang, X. AdvCheck: Characterizing adversarial examples via local gradient checking. *Comput. Secur.* **2024**, *136*, 103540. [\[CrossRef\]](#)
103. Aliyu, I.; Van Engelenburg, S.; Mu’azu, M.B.; Kim, J.; Lim, C.G. Statistical detection of adversarial examples in blockchain-based federated forest in-vehicle network intrusion detection systems. *IEEE Access* **2022**, *10*, 109366–109384. [\[CrossRef\]](#)
104. Shieh, C.-S.; Nguyen, T.-T.; Lin, W.-W.; Lai, W.K.; Horng, M.-F.; Miu, D. Detection of adversarial DDoS attacks using symmetric defense generative adversarial networks. *Electronics* **2022**, *11*, 1977. [\[CrossRef\]](#)
105. Alasmary, H.; Abusnaina, A.; Jang, R.; Abuhamad, M.; Anwar, A.; Nyang, D. Soteria: Detecting adversarial examples in control flow graph-based malware classifiers. In Proceedings of the 2020 IEEE 40th International Conference on Distributed Computing Systems, Singapore, 29 November–1 December 2020.
106. Wang, M.; Liu, Z. Defense against adversarial attacks in image recognition based on multilayer filters. *Appl. Sci.* **2024**, *14*, 8119. [\[CrossRef\]](#)
107. Costa, J.; Apolinário, F.; Ribeiro, C. ARGAN-IDS: Adversarial resistant intrusion detection systems using generative adversarial networks. In Proceedings of the 19th International Conference on Availability, Reliability and Security (ARES ’24), Vienna, Austria, 30 July–2 August 2024.
108. Sajeeda, A.; Hossain, B.M.M. Exploring generative adversarial networks and adversarial training. *Int. J. Cogn. Comput. Eng.* **2022**, *3*, 78–89. [\[CrossRef\]](#)
109. Hwang, R.-H.; Lin, J.-Y.; Hsieh, S.-Y.; Lin, H.-Y.; Lin, C.-L. Adversarial patch attacks on deep-learning-based face recognition systems using generative adversarial networks. *Sensors* **2023**, *23*, 853. [\[CrossRef\]](#) [\[PubMed\]](#)
110. Zhao, W.; Mahmoud, Q.H.; Alwidian, S. Evaluation of GAN-based model for adversarial training. *Sensors* **2023**, *23*, 2697. [\[CrossRef\]](#) [\[PubMed\]](#)
111. Wang, P.; Lin, H.-C.; Chen, J.-H.; Lin, W.-H.; Li, H.-C. Improving cyber defense against ransomware: A generative adversarial networks-based adversarial training approach for long short-term memory network classifier. *Electronics* **2025**, *14*, 810. [\[CrossRef\]](#)
112. Chen, J.; Yuan, C.; Li, J.; Tian, D.; Ma, R.; Jia, X. ELAMD: An ensemble learning framework for adversarial malware defense. *J. Inf. Secur. Appl.* **2023**, *75*, 103508. [\[CrossRef\]](#)
113. Baldini, G. Mitigation of adversarial attacks in 5G networks with a robust intrusion detection system based on extremely randomized trees and infinite feature selection. *Electronics* **2024**, *13*, 2405. [\[CrossRef\]](#)
114. Lin, Y.D.; Pratama, J.H.; Sudyana, D.; Lai, Y.C.; Hwang, R.H.; Lin, P.C.; Lin, H.-Y.; Lee, W.-B.; Chiang, C.-K. ELAT: Ensemble learning with adversarial training in defending against evaded intrusions. *J. Inf. Secur. Appl.* **2022**, *71*, 103348. [\[CrossRef\]](#)
115. Chen, X.; Huang, W.; Peng, Z.; Guo, W.; Zhang, F. Diversity supporting robustness: Enhancing adversarial robustness via differentiated ensemble predictions. *Comput. Secur.* **2024**, *142*, 103861. [\[CrossRef\]](#)
116. Huang, L.; Huang, Q.; Qiu, P.; Wei, S.; Gao, C. FASTEN: Fast ensemble learning for improved adversarial robustness. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 2565–2580. [\[CrossRef\]](#)
117. Le, T.-T.-H.; Cho, J.; Shin, D.; Kim, H. Toward adversarial robustness network intrusion detection based on multi-model ensemble approach. *Sensors* **2026**, *26*, 2478. [\[CrossRef\]](#) [\[PubMed\]](#)
118. Li, M.; Qiao, Y.; Lee, B. Adversarial robustness evaluation for multi-view deep learning cybersecurity anomaly detection. *Future Internet* **2025**, *17*, 459. [\[CrossRef\]](#)
119. Wei, W.; Liu, L. Robust deep learning ensemble against deception. *IEEE Trans. Dependable Secur. Comput.* **2020**, *18*, 1513–1527. [\[CrossRef\]](#)
120. Zhang, C.; Costa-Pérez, X.; Patras, P. Adversarial attacks against deep learning-based network intrusion detection systems and defense mechanisms. *IEEE/ACM Trans. Netw.* **2022**, *30*, 1294–1311. [\[CrossRef\]](#)
121. Park, H. Semantic super-resolution via self-distillation and adversarial learning. *IEEE Access* **2024**, *12*, 2361–2370. [\[CrossRef\]](#)
122. Hong, I.; Lee, S. Exploring synergy of denoising and distillation: Novel method for efficient adversarial defense. *Appl. Sci.* **2024**, *14*, 10872. [\[CrossRef\]](#)
123. Menet, F.; Berthier, P.; Gagnon, M.; Fernandez, J.M. Spartan networks: Self-feature-squeezing neural networks for increased robustness in adversarial settings. *Comput. Secur.* **2020**, *88*, 101537. [\[CrossRef\]](#)
124. Gibert, D.; Zizzo, G.; Le, Q.; Planes, J. Adversarial robustness of deep learning-based malware detectors via (de)randomized smoothing. *IEEE Access* **2024**, *12*, 61152–61162. [\[CrossRef\]](#)
125. Zhang, W.; Zhong, M.; Tandon, R.; Krunz, M. Filtered randomized smoothing: A new defense for robust modulation classification. In *MILCOM 2024—2024 IEEE Military Communications Conference (MILCOM)*; IEEE: New York, NY, USA, 2024.
126. Everett, M.; Lütjens, B.; How, J.P. Certifiable robustness to adversarial state uncertainty in deep reinforcement learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 4184–4198. [\[CrossRef\]](#) [\[PubMed\]](#)
127. Augustin, M.; Meinke, A.; Hein, M. Adversarial robustness on in- and out-distribution improves explainability. In *European Conference on Computer Vision (ECCV)*; Springer International Publishing: Cham, Switzerland, 2020; pp. 228–245.

128. Gibert, D.; Zizzo, G.; Le, Q. Towards a practical defense against adversarial attacks on deep learning-based malware detectors via randomized smoothing. In Proceedings of the ESORICS 2023 International Workshops, The Hague, The Netherlands, 25–29 September 2023.
129. Patil, S.; Varadarajan, V.; Walimbe, D.; Gulechha, S.; Shenoy, S.; Raina, A.; Kotecha, K. Improving the robustness of AI-based malware detection using adversarial machine learning. *Algorithms* **2021**, *14*, 297. [[CrossRef](#)]
130. Imran, M.; Appice, A.; Malerba, D. Evaluating realistic adversarial attacks against machine learning models for Windows PE malware detection. *Future Internet* **2024**, *16*, 168. [[CrossRef](#)]
131. Wang, C.; Zhang, L.; Zhao, K.; Ding, X.; Wang, X. AdvAndMal: Adversarial training for Android malware detection and family classification. *Symmetry* **2021**, *13*, 1081. [[CrossRef](#)]
132. Adawadkar, A.M.K.; Kulkarni, N. Cyber-security and reinforcement learning—A brief survey. *Eng. Appl. Artif. Intell.* **2022**, *114*, 105116. [[CrossRef](#)]
133. Gholizadeh, N.; Kazemi, N.; Musilek, P. A comparative study of reinforcement learning algorithms for distribution network reconfiguration with deep Q-learning-based action sampling. *IEEE Access* **2023**, *11*, 13714–13723. [[CrossRef](#)]
134. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-level control through deep reinforcement learning. *Nature* **2015**, *518*, 529–533. [[CrossRef](#)] [[PubMed](#)]
135. Lopez-Martin, M.; Carro, B.; Sanchez-Esguevillas, A. Deep Q-learning based reinforcement learning approach for network intrusion detection. *Computers* **2022**, *11*, 41. [[CrossRef](#)]
136. Tong, Y.; Liang, H.; Ma, H.; Zhang, S.; Yang, X. A survey on reinforcement learning-driven adversarial sample generation for PE malware. *Electronics* **2025**, *14*, 2422. [[CrossRef](#)]
137. Mnih, V.; Badia, A.P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; Kavukcuoglu, K. Asynchronous methods for deep reinforcement learning. In Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML), New York, NY, USA, 19–24 June 2016; pp. 1928–1937.
138. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal policy optimization algorithms. *arXiv* **2017**, arXiv:1707.06347.
139. Schulman, J.; Levine, S.; Abbeel, P.; Jordan, M.I.; Moritz, P. Trust region policy optimization. In Proceedings of the 32nd International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015; pp. 1889–1897.
140. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. In Proceedings of the 4th International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016.
141. Haarnoja, T.; Zhou, A.; Abbeel, P.; Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Proceedings of the 35th International Conference on Machine Learning (ICML), Stockholm, Sweden, 10–15 July 2018; pp. 1861–1870.
142. Wu, Y.; Wang, Y.; Ding, P.; Wang, H.; Zhu, B.; Liu, C. Enhancing security in deep reinforcement learning: A comprehensive survey on adversarial attacks and defenses. *Neurocomputing* **2026**, 133503. [[CrossRef](#)]
143. Zhang, H.; Chen, H.; Xiao, C.; Li, B.; Liu, M.; Boning, D.; Hsieh, C.-J. Robust deep reinforcement learning against adversarial perturbations on state observations. In *Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; pp. 8829–8841.
144. Bykovets, E.; Metz, Y.; El-Assady, M.; Keim, D.A.; Buhmann, J.M. BARreL: Bottleneck attention for adversarial robustness in vision-based reinforcement learning. In Proceedings of the ECCV 2022 Workshop on Adversarial Robustness in the Real World (AROW), Virtual, 23 October 2022.
145. Cai, K.; Zhu, X.; Hu, Z. Reward poisoning attacks in deep reinforcement learning based on exploration strategies. *Neurocomputing* **2023**, *553*, 126578. [[CrossRef](#)]
146. Han, Y.; Hubczenko, D.; Montague, P.; De Vel, O.; Abraham, T.; Rubinstein, B.I.P.; Leckie, C.; Alpcan, T.; Erfani, S. Adversarial reinforcement learning under partial observability in autonomous computer network defence. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8.
147. Elhami Fard, N.; Selmic, R.R. Adversarial attacks on heterogeneous multi-agent deep reinforcement learning system with time-delayed data transmission. *J. Sens. Actuator Netw.* **2022**, *11*, 45. [[CrossRef](#)]
148. Mohan, A.; Schön, T. Toward robust agents: A survey of adversarial attacks and defenses in deep reinforcement learning. *IEEE Access* **2026**, *14*, 14481–14497. [[CrossRef](#)]
149. Zhang, L.; Liu, P.; Choi, Y.-H.; Chen, P. Semantics-preserving reinforcement learning attack against graph neural networks for malware detection. *IEEE Trans. Dependable Secur. Comput.* **2023**, *20*, 1390–1402. [[CrossRef](#)]
150. Oh, S.H.; Kim, J.; Park, J. Dynamic cyberattack simulation: Integrating improved deep reinforcement learning with the MITRE-ATT&CK framework. *Electronics* **2024**, *13*, 2831. [[CrossRef](#)]
151. Song, J.; Yu, D.; Teng, H.; Chen, Y. RLVS: A reinforcement learning-based sparse adversarial attack method for black-box video recognition. *Electronics* **2025**, *14*, 245. [[CrossRef](#)]

152. Yang, C.; Kortylewski, A.; Xie, C.; Cao, Y.; Yuille, A. PatchAttack: A black-box texture-based attack with reinforcement learning. In Proceedings of the 2020 European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 681–698.
153. Bakos, S.; Madani, P.; Davoudi, H. Noise as a double-edged sword: Reinforcement learning exploits randomized defenses in neural networks. In Proceedings of the 2024 IEEE 7th Int. Conf. Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 24–27 May 2024.
154. Louthanová, P.; Kozák, M.; Jureček, M.; Stamp, M. A comparison of adversarial learning techniques for malware detection. *J. Comput. Virol. Hacking Tech.* **2024**, *20*, 531–549.
155. Wang, D.; Ju, A.; Shelhamer, E.; Wagner, D.; Darrell, T. Fighting gradients with gradients: Dynamic defenses against adversarial attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2021.
156. Tsingenopoulos, I.; Rimmer, V.; Preuveneers, D.; Pierazzi, F.; Cavallaro, L.; Joosen, W. Adversarial Markov games: On adaptive decision-based attacks and defenses. In Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, 19–23 May 2024.
157. Yao, Q.; Wang, Y.; Xiong, X.; Wang, P.; Li, Y. Adversarial decision-making for moving target defense: A multi-agent Markov game and reinforcement learning approach. *Entropy* **2023**, *25*, 605. [[CrossRef](#)] [[PubMed](#)]
158. Hammar, K.; Stadler, R. Intrusion prevention through optimal stopping. *IEEE Trans. Netw. Serv. Manag.* **2022**, *19*, 2333–2348. [[CrossRef](#)]
159. Haydari, A.; Zhang, M.; Chuah, C.-N. Adversarial attacks and defense in deep reinforcement learning (DRL)-based traffic signal controllers. *IEEE Open J. Intell. Transp. Syst.* **2021**, *2*, 402–416. [[CrossRef](#)]
160. Kim, M.-S. Deep reinforcement learning-based adversarial attack and defense in industrial control systems. *Mathematics* **2024**, *12*, 3900. [[CrossRef](#)]
161. Wang, T.; Liu, Q.; Shen, H.; Bai, G. IDNet: Instance-adaptive dynamic network with adversarial training for intrusion detection. *Comput. Netw.* **2026**, 112336.
162. Larriva-Novo, X.; Pérez Miguel, L.; Villagra, V.A.; Álvarez-Campana, M.; Sanchez-Zas, C.; Jover, Ó. Post-hoc categorization based on explainable AI and reinforcement learning for improved intrusion detection. *Appl. Sci.* **2024**, *14*, 11511. [[CrossRef](#)]
163. Yang, X.; Wu, L.; Wang, L.; Ying, C.; Su, H.; Zhu, J. Reinforced embodied active defense: Exploiting adaptive interaction for robust visual perception in adversarial 3D environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 9078–9094. [[CrossRef](#)]
164. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why should I trust you?”: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
165. Lundberg, S.M.; Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2017.
166. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
167. Shi, Y.; Han, Y.; Zhang, Q.; Kuang, X. Adaptive iterative attack towards explainable adversarial robustness. *Pattern Recognit.* **2020**, *105*, 107309. [[CrossRef](#)]
168. Baniecki, H.; Biecek, P. Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Inf. Fusion* **2024**, *107*, 102303. [[CrossRef](#)]
169. Noppel, M.; Peter, L.; Wressnegger, C. Disguising attacks with explanation-aware backdoors. In Proceedings of the 44th IEEE Symposium on Security and Privacy (S&P), San Francisco, CA, USA, 21–25 May 2023; pp. 664–681.
170. Henna, S.; Rathnayake, U. Fast-gradient-guided generative adversarial learning for explainable cyber threat intelligence. *Appl. Soft Comput.* **2026**, *194*, 114911. [[CrossRef](#)]
171. Repetto, S.; Maljkovic, I.; Lotto, M.; Cinà, A.E.; Vascon, S.; Roli, F. Evaluating the robustness of explainable AI in medical image recognition under natural and adversarial data corruption. *Mach. Learn.* **2026**, *115*, 4. [[CrossRef](#)]
172. Munilla, J.; Khammas, R.M. Evaluation of explainable artificial intelligence in IoT intrusion detection systems under DeepFool adversarial conditions. *Sensors* **2026**, *26*, 2924. [[CrossRef](#)] [[PubMed](#)]
173. Ojo, S.; Krichen, M.; Alamro, M.A.; Mihoub, A. TXAI-ADV: Trustworthy XAI for defending AI models against adversarial attacks in realistic CIoT. *Electronics* **2024**, *13*, 1769. [[CrossRef](#)]
174. Vaccari, I.; Carlevaro, A.; Narteni, S.; Cambiaso, E.; Mongelli, M. eXplainable and reliable against adversarial machine learning in data analytics. *IEEE Access* **2022**, *10*, 83949–83970. [[CrossRef](#)]
175. Tasneem, S.; Islam, K.A. Improve adversarial robustness of AI models in remote sensing via data-augmentation and explainable-AI methods. *Remote Sens.* **2024**, *16*, 3210. [[CrossRef](#)]
176. Agrawal, N.; Pendharkar, I.; Shroff, J.; Raghuvanshi, J.; Neogi, A.; Patil, S.; Walambe, R.; Kotecha, K. A-XAI: Adversarial machine learning for trustable explainability. *AI Ethics* **2024**, *4*, 1143–1174. [[CrossRef](#)]

177. Chaudhry, U.; Khan, M.A.; Hayat, M.F.; Razzak, I. Adversarial attacks on explainable AI methods in cybersecurity. *arXiv* **2025**, arXiv:2510.03623.
178. Sauka, K.; Shin, G.-Y.; Kim, D.-W.; Han, M.-M. Adversarial robust and explainable network intrusion detection systems based on deep learning. *Appl. Sci.* **2022**, *12*, 6451. [\[CrossRef\]](#)
179. Albahri, A.S.; Hamid, R.A.; Abdulnabi, A.R.; Albahri, O.S.; Alamoodi, A.H.; Deveci, M.; Pedrycz, W.; Alzubaidi, L.; Santamaría, J.; Gu, Y. Fuzzy decision-making framework for explainable golden multi-machine learning models for real-time adversarial attack detection in vehicular ad-hoc networks. *Inf. Fusion* **2024**, *105*, 102208. [\[CrossRef\]](#)
180. Moraliyage, H.; Kulawardana, G.; De Silva, D.; Issadeen, Z.; Manic, M.; Katsura, S. Explainable artificial intelligence with integrated gradients for the detection of adversarial attacks on text classifiers. *Appl. Syst. Innov.* **2025**, *8*, 17. [\[CrossRef\]](#)
181. Saheed, Y.K.; Chukwuere, J.E. XAI-enhanced adversarial resilient deep learning framework for transparent and secure edge deployment in consumer Internet of Things/industrial Internet of Things environments. *Comput. Electr. Eng.* **2026**, *133*, 111080. [\[CrossRef\]](#)
182. Imam, N.H. Adversarial examples on XAI-enabled DT for smart healthcare systems. *Sensors* **2024**, *24*, 6891. [\[CrossRef\]](#) [\[PubMed\]](#)
183. Albu-Salih, A.; Sallabi, F.; Qayyum, A.; Qadir, J. Enhancing adversarial robustness of IoT intrusion detection systems via SHAP-based fingerprinting. *arXiv* **2024**, arXiv:2511.06197.
184. Gaspar, D.; Pereira, T.; Ferreira, A.; Correia, M. Explainable AI for intrusion detection systems: LIME and SHAP applicability on multi-layer perceptron. *IEEE Access* **2024**, *12*, 28672–28690. [\[CrossRef\]](#)
185. Rezaei, H.; Taheri, R.; Jordanov, I.; Shiaeles, S. Adapt-LFA: Adaptive Gradient-Guided Label Flipping Attack Against Federated Learning-Based Intrusion Detection in IoT. In Proceedings of the IEEE Int. Conf. Cyber Security and Resilience (CSR), Chania, Greece, 4–6 August 2025; pp. 407–412.
186. Taheri, R.; Jafari, R.; Arabikhan, F.; Gegov, A.; Ichtev, A. Explainable AI for Federated Learning-Based Intrusion Detection Systems in Connected Vehicles. *Electronics* **2025**, *14*, 4508. [\[CrossRef\]](#)
187. Nikfam, F.; Marchisio, A.; Martina, M.; Shafique, M. AccelAT: A framework for accelerating the adversarial training of deep neural networks through accuracy gradient. *IEEE Access* **2022**, *10*, 108997–109007. [\[CrossRef\]](#)
188. Rathore, H.; Sahay, S.K.; Nikam, P.; Sewak, M. Robust Android malware detection system against adversarial attacks using Q-learning. *Inf. Syst. Front.* **2021**, *23*, 867–882. [\[CrossRef\]](#)
189. Labaca-Castro, R.; Franz, S.; Rodosek, G.D. AIMED-RL: Exploring adversarial malware examples with reinforcement learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; Springer International Publishing: Cham, Switzerland, 2021; pp. 37–52.
190. Kozák, M.; Jureček, M.; Stamp, M.; Di Troia, F. Creating valid adversarial examples of malware. *J. Comput. Virol. Hacking Tech.* **2024**, *20*, 607–621. [\[CrossRef\]](#)
191. Fang, Y.; Zeng, Y.; Li, B.; Liu, L.; Zhang, L. DeepDetectNet vs. RLAttackNet: An adversarial method to improve deep learning-based static malware detection model. *PLoS ONE* **2020**, *15*, e0231626. [\[CrossRef\]](#) [\[PubMed\]](#)
192. Zhong, F.; Hu, P.; Zhang, G.; Li, H.; Cheng, X. Reinforcement learning based adversarial malware example generation against black-box detectors. *Comput. Secur.* **2022**, *121*, 102869. [\[CrossRef\]](#)
193. Zhan, D.; Zhang, Y.; Zhu, L.; Chen, J.; Xia, S.; Guo, S.; Pan, Z. Enhancing reinforcement learning based adversarial malware generation to evade static detection. *Alex. Eng. J.* **2024**, *98*, 32–43. [\[CrossRef\]](#)
194. Phan, T.D.; Luong, T.D.; An, N.H.Q.; Huu, Q.N.; Nghi, H.K.; Pham, V.-H. Leveraging reinforcement learning and generative adversarial networks to craft mutants of Windows malware against black-box malware detectors. In Proceedings of the 11th International Symposium on Information and Communication Technology (SoICT), Hanoi, Vietnam, 1–3 December 2022; pp. 31–38.
195. Rigaki, M.; Garcia, S. The power of MEME: Adversarial malware creation with model-based reinforcement learning. In *European Symposium on Research in Computer Security*; Springer Nature: Cham, Switzerland, 2024; pp. 44–64.
196. Randhawa, R.H.; Aslam, N.; Alauthman, M.; Khalid, M.; Rafiq, H. Deep reinforcement learning based evasion generative adversarial network for botnet detection. *Future Gener. Comput. Syst.* **2023**, *150*, 294–302. [\[CrossRef\]](#)
197. Paya, A.; Arroni, S.; García-Díaz, V.; Gómez, A. Apollon: A robust defense system against adversarial machine learning attacks in intrusion detection systems. *Comput. Secur.* **2023**, *136*, 103546. [\[CrossRef\]](#)
198. Lee, C.; Lee, S. Evaluating the vulnerability of YOLOv5 to adversarial attacks for enhanced cybersecurity in MASS. *J. Mar. Sci. Eng.* **2023**, *11*, 947. [\[CrossRef\]](#)
199. Wang, Z.; Liu, Y. Cloud-based XAI services for assessing open repository models under adversarial attacks. In Proceedings of the International Conference on Services Computing (SCC), Bangkok, Thailand, 16–19 November 2024.
200. Qiu, H.; Dong, T.; Zhang, T.; Lu, J.; Memmi, G.; Qiu, M. Adversarial attacks against network intrusion detection in IoT systems. *IEEE Internet Things J.* **2021**, *8*, 10327–10335. [\[CrossRef\]](#)

201. Aiken, J.; Scott-Hayward, S. Investigating adversarial attacks against network intrusion detection systems in SDNs. In Proceedings of the IEEE Conference on Network Functions Virtualization and Software Defined Networking (NFV-SDN), Madrid, Spain, 9–12 November 2020.
202. Fidel, G.; Bitton, R.; Shabtai, A. When explainability meets adversarial learning: Detecting adversarial examples using SHAP signatures. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Virtual, 19–24 July 2020.
203. Oh, S.H.; Kim, J.; Nah, J.H.; Park, J. Employing deep reinforcement learning to cyber-attack simulation for enhancing cybersecurity. *Electronics* **2024**, *13*, 555. [\[CrossRef\]](#)
204. Kao, C.-Y.; Chen, J.; Markert, K.; Böttinger, K. Rectifying adversarial inputs using XAI techniques. In Proceedings of the 31st European Signal Processing Conference (EUSIPCO), Helsinki, Finland, 4–8 September 2023.
205. Tan, W.; Renkhoff, J.; Velasquez, A.; Wang, Z.; Li, L.; Wang, J.; Niu, S.; Yang, F.; Liu, Y.; Song, H. NoiseCAM: Explainable AI for the boundary between noise and adversarial attacks. In Proceedings of the 2023 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Incheon, Republic of Korea, 13–17 August 2023.
206. Galli, A.; Marrone, S.; Moscato, V.; Sansone, C. Reliability of eXplainable artificial intelligence in adversarial perturbation scenarios. In *Pattern Recognition. ICPR International Workshops and Challenges*; Springer: Cham, Switzerland, 2021; pp. 243–256.
207. Gowrisankar, B.; Thing, V.L.L. An adversarial attack approach for eXplainable AI evaluation on deepfake detection models. *Comput. Secur.* **2024**, *139*, 103684. [\[CrossRef\]](#)
208. Chen, J.; Chen, H.; Chen, K.; Zhang, Y.; Zou, Z.; Shi, Z. Diffusion models for imperceptible and transferable adversarial attack. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *47*, 961–977. [\[CrossRef\]](#) [\[PubMed\]](#)
209. Zhong, Y.; Deng, W. Towards transferable adversarial attack against deep face recognition. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 1452–1466. [\[CrossRef\]](#)
210. Khedr, Y.M.; Liu, X.; Lu, H.; He, K. Transferable adversarial attacks against face recognition using surrogate model fine-tuning. *Appl. Soft Comput.* **2025**, *172*, 112983. [\[CrossRef\]](#)
211. Smagulova, K.; Bacha, L.; Fouda, M.E.; Kanj, R.; Eltawil, A. Robustness and transferability of adversarial attacks on different image classification neural networks. *Electronics* **2024**, *13*, 592. [\[CrossRef\]](#)
212. Zeng, J.; Xu, J.; Zheng, X.; Huang, X. Certified robustness to text adversarial attacks by randomized [MASK]. *Comput. Linguist.* **2023**, *49*, 395–427. [\[CrossRef\]](#)
213. Ibrahim, A.D.M.; Shang, Z.; Hong, J.-E. How resilient are Kolmogorov–Arnold networks in classification tasks? A robustness investigation. *Appl. Sci.* **2024**, *14*, 10173. [\[CrossRef\]](#)
214. Zhang, J.; Chao, H.; Dasegowda, G.; Wang, G.; Kalra, M.K.; Yan, P. Overlooked trustworthiness of saliency maps. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer Nature: Cham, Switzerland, 2022; pp. 451–461.
215. Dai, Z.; Liu, S.; Li, Q.; Tang, K. Saliency attack: Towards imperceptible black-box adversarial attack. *ACM Trans. Intell. Syst. Technol.* **2022**, *14*, 1–20. [\[CrossRef\]](#)
216. Wang, S.; Gong, Y. Adversarial example detection based on saliency map features. *Appl. Intell.* **2022**, *52*, 6262–6275. [\[CrossRef\]](#)
217. Lu, R.; Yu, P.; Huang, Z.; Xia, Z.; Jiang, X. BAM: Backdoor defense based on adversarial mitigation. *Pattern Recognit.* **2025**, *162*, 112662. [\[CrossRef\]](#)
218. Eghtesad, T.; Vorobeychik, Y.; Laszka, A. Adversarial deep reinforcement learning based adaptive moving target defense. In Proceedings of the 11th Conference on Decision and Game Theory for Security (GameSec 2020), College Park, MD, USA, 26–30 October 2020; pp. 58–79.
219. Babaagba, K.O.; Tan, Z. Mitigating metamorphic malware through adversarial learning techniques. *Network* **2026**, *6*, 22. [\[CrossRef\]](#)
220. He, Z.; Homayoun, H.; Sayadi, H. Beyond conventional defenses: Proactive and adversarial-resilient hardware malware detection using deep reinforcement learning. In Proceedings of the 61st ACM/IEEE Design Automation Conference (DAC), San Francisco, CA, USA, 23–27 June 2024.
221. Hu, Y.; Sun, S. RL-VAEGAN: Adversarial defense for reinforcement learning agents via style transfer. *Knowl.-Based Syst.* **2021**, *221*, 106967. [\[CrossRef\]](#)
222. García, J.; Sagredo, I. Instance-based defense against adversarial attacks in deep reinforcement learning. *Eng. Appl. Artif. Intell.* **2021**, *107*, 104514. [\[CrossRef\]](#)
223. Tan, K.L.; Esfandiari, Y.; Lee, X.Y.; Aakanksha; Sarkar, S. Robustifying reinforcement learning agents via action space adversarial training. In Proceedings of the 2020 American Control Conference (ACC), Denver, CO, USA, 1–3 July 2020.
224. Zhan, D.; Liu, X.; Bai, W.; Li, W.; Guo, S.; Pan, Z. GAME-RL: Generating adversarial malware examples against API call based detection via reinforcement learning. *IEEE Trans. Dependable Secur. Comput.* **2025**, *22*, 5431–5447. [\[CrossRef\]](#)
225. Erdem, M.; Üre, N.K. Learning to balance mixed adversarial attacks for robust reinforcement learning. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 108. [\[CrossRef\]](#)
226. Benaddi, H.; Jouhari, M.; Ibrahim, K.; Ben Othman, J.; Amhoud, E.M. Anomaly detection in Industrial IoT using distributional reinforcement learning and generative adversarial networks. *Sensors* **2022**, *22*, 8085. [\[CrossRef\]](#) [\[PubMed\]](#)

227. Zhou, Z.; Liu, G.; Guo, W.; Zhou, M. Adversarial attacks on multiagent deep reinforcement learning models in continuous action space. *IEEE Trans. Syst. Man Cybern. Syst.* **2024**, *54*, 7633–7646. [[CrossRef](#)]
228. Banik, S.; Ramachandran, T.; Bhattacharya, A.; Bopardikar, S.D. Automated adversary-in-the-loop cyber-physical defense planning. *ACM Trans. Cyber-Phys. Syst.* **2023**, *7*, 1–25. [[CrossRef](#)]
229. Alnfai, M.M. AI-powered cyber resilience: A reinforcement learning approach for automated threat hunting in 5G networks. *EURASIP J. Wirel. Commun. Netw.* **2025**, *2025*, 68. [[CrossRef](#)]
230. Farooq, M.O. Adaptive cyber defense through hybrid learning: From specialization to generalization. *Future Internet* **2025**, *17*, 464. [[CrossRef](#)]
231. Makarem, M.A.; Razaz, M.A.; Saleh, Z. Co-adaptive attacker–defender learning over attack graphs: A stochastic game approach to dynamic network defense. *Future Internet* **2026**, *18*, 239. [[CrossRef](#)]
232. Rahman, M.; Roy, P.; Frizell, S.S.; Qian, L. Evaluating pretrained deep learning models for image classification against individual and ensemble adversarial attacks. *IEEE Access* **2025**, *13*, 35230–35242. [[CrossRef](#)]
233. Wang, M.; Yang, N.; Gunasinghe, D.H.; Weng, N. On the robustness of ML-based network intrusion detection systems: An adversarial and distribution shift perspective. *Computers* **2023**, *12*, 209. [[CrossRef](#)]
234. Abraha, T.M.; Graham-Knight, J.B.; Lasserre, P.; Najjaran, H.; Lucet, Y. ConvNet-generated adversarial perturbations for evaluating 3D object detection robustness. *Sensors* **2025**, *25*, 6026. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.