Adversarial Machine Learning: Emerging Threats and Defense Mechanisms

Ankur Sharma
Bourns, Inc. / IT & Security
Western Governors University, Millcreek, USA

ABSTRACT
Artificial intelligence and machine learning are increasingly embedded in modern cybersecurity systems because of their capacity to automate threat detection, analysis, and anomaly identification. This shift toward intelligent defense has, however, introduced a new class of vulnerabilities, collectively termed adversarial machine learning (AML), in which attackers craft malicious inputs, poison training data, or mount inference-based attacks to subvert model behavior. This study reviews the emerging adversarial threats targeting cybersecurity applications and critically appraises the defense strategies proposed to strengthen model robustness and resilience. A structured literature review of 22 peer-reviewed studies published between 2022 and 2025 was conducted to identify dominant attack patterns, the most vulnerable application domains, and the limitations of current defenses. To ground the review in measurable evidence, a reproducible case study is additionally reported in which a neural intrusion-detection model is subjected to a Fast Gradient Sign Method (FGSM) evasion attack on the NSL-KDD dataset and then hardened through adversarial training. The undefended detector's accuracy collapses from 80.8% to near 0% as the perturbation budget grows, whereas the adversarial trained detector retains roughly 70-75% accuracy under the same attack, at the cost of a small reduction in clean accuracy. The findings confirm that adversarial attacks are becoming increasingly sophisticated, particularly against intrusion detection systems, autonomous systems, and deep-learning-based security tools, and that, although adversarial training, defensive distillation, and explainable AI are promising, open questions remain regarding their scalability, adaptability, and real-time applicability. The study underscores the need for multi-layered, adaptive security strategies to enhance the trustworthiness of AI-powered cybersecurity solutions.

General Terms
Security, Machine Learning, Experimentation, Algorithms, Reliability.

Keywords
Adversarial Machine Learning; Cybersecurity; Adversarial Attacks; Defense Mechanisms; Deep Learning Security; Poisoning Attacks; Evasion Attacks; Robust Artificial Intelligence.

1. INTRODUCTION
Artificial intelligence (AI) and machine learning (ML) have transformed modern cybersecurity by enabling intelligent threat detection, automated incident response, anomaly identification, and predictive security analytics. Because these techniques can analyze large and dynamic volumes of security data, they have been adopted in increasingly advanced systems such as intrusion detection systems, malware classifiers, autonomous infrastructures, biometric authentication, and network monitoring environments [1], [2]. As AI-driven defenses have become central to the digital security landscape, machine learning has emerged as a foundational element of intelligent cybersecurity architectures.

This reliance has, in turn, created a new attack surface. Adversarial machine learning refers to a class of attacks in which an adversary deliberately undermines a model by exploiting weaknesses in its training data, feature representations, parameters, or inference process, causing it to produce inaccurate or misleading outputs [3]. Unlike traditional cyberattacks that target software or network infrastructure, AML attacks target the learning and decision-making behavior of intelligent systems, creating complex security challenges for organizations that increasingly depend on automated defense technologies.

In recent years, adversarial attacks have grown more complex, adaptive, and difficult to detect across many cybersecurity domains. Evasion attacks, poisoning attacks, model inversion, membership inference, and backdoor manipulations have all demonstrated the capacity to compromise the integrity, confidentiality, and reliability of machine-learning systems [4], [5]. Evasion attacks perturb malicious inputs to mislead trained classifiers at inference time [6], whereas poisoning attacks insert or alter training samples to corrupt the learned model. In domains where misclassification carries severe operational and safety consequences, such as intrusion detection, autonomous-vehicle security, wireless communication, and deep-learning malware detection, these threats are especially dangerous [7], [8].

As AML threats have advanced, researchers and practitioners have raised significant concerns that existing defenses are insufficient. Several mitigation approaches have been proposed, including adversarial training, defensive distillation, robust optimization, explainable artificial intelligence, and input preprocessing; yet many remain vulnerable to adaptive and black-box attacks [9], [10]. The continually evolving nature of adversarial attacks further challenges the scalability, explainability, and real-time deployment of current defenses. Apruzzese et al. [11] show that realistic adversarial scenarios in network intrusion detection can expose substantial weaknesses in model robustness under operational conditions.

Beyond technical vulnerabilities, considerations of trust, reliability, governance, and economics are central to securing AI-driven infrastructure. Merkle et al. [12] argue that the economic and logistical costs of deploying adversarial defenses can materially affect the viability of machine learning in critical cybersecurity contexts. At the same time, the integration of AI into cloud computing, the Internet of Things (IoT), smart transportation, and autonomous systems continue to widen the available attack surface [2], [13].

As attacks become more sophisticated and pervasive, a comprehensive understanding of how adversarial threats emerge, and of the mechanisms available to counter them, is essential. Although individual dimensions of adversarial attacks and defenses have been studied, an integrated overview of the key challenges and opportunities for resilient AI security remains

needed. Accordingly, this study critically reviews the emerging challenges associated with adversarial machine learning and appraises contemporary defense strategies for their ability to improve the robustness and security of ML-based cybersecurity systems. The review identifies dominant attack patterns, evaluates the effectiveness of current mitigation techniques, and highlights priority directions for future research. To complement the qualitative synthesis with measurable evidence, the study also reports a reproducible experiment in which an evasion attack is mounted against a neural intrusion detector and then countered through adversarial training.

2. METHODOLOGY

2.1 Research Design

This study adopts a structured literature review (SLR) to critically examine the emerging threats posed by adversarial machine learning and the defense strategies used to counter them in cybersecurity environments. A structured approach was selected because AML research is evolving rapidly and a unified, cross-domain synthesis of the scholarly literature is needed. Structured reviews provide a broad analytical framework for synthesizing empirical and conceptual work from peer-reviewed sources and for evaluating current research trends, vulnerabilities, and mitigation strategies. The methodology was designed to ensure transparency, analytical rigor, and repeatability in identifying and assessing relevant publications, with a focus on adversarial attacks against ML-based cybersecurity systems and the defenses developed to make models robust, resilient, and reliable.

2.2 Data Sources and Literature Selection

Relevant publications were identified using widely used scholarly databases and digital repositories in cybersecurity and artificial intelligence, namely IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, MDPI, and Google Scholar. These databases were chosen for their comprehensive coverage of the cybersecurity, machine-learning, and artificial-intelligence literature. To keep the review aligned with the most recent developments in adversarial techniques and defense architectures, the search emphasized studies published between 2022 and 2025, with a small number of earlier foundational works retained for conceptual continuity. Several keyword combinations were used during retrieval, including 'adversarial machine learning', 'cybersecurity', 'adversarial attacks', 'poisoning attacks', 'evasion attacks', 'intrusion detection systems', 'deep learning security', 'robust machine learning', 'AI security', and 'defense mechanisms'. Boolean operators (AND, OR, NOT) were applied to refine the results and remove irrelevant records.

2.3 Inclusion and Exclusion Criteria

Explicit inclusion and exclusion criteria were defined to maintain consistency and academic rigor. Studies were included when they were peer-reviewed; addressed adversarial threats, vulnerabilities, or defense methods; concerned machine-learning or deep-learning problems relevant to cybersecurity systems; and were written in English between 2022 and 2025. Studies were excluded when they treated AI solely as a tool rather than as the subject of the security problem, lacked sufficient methodological or analytical detail, were duplicate or non-peer-reviewed publications, addressed AI theory only peripherally relevant to cybersecurity, or reported outdated results no longer aligned with recent advances. This process kept the analyzed corpus within a well-defined scope closely aligned with the research objectives. The resulting study-selection workflow is summarized in Figure 1.

2.4 Data Collection Procedure

Data collection followed a staged process. A broad body of scholarly work related to adversarial machine learning and cybersecurity was first identified through database searches. Titles, abstracts, and keywords were then screened and assigned an initial relevance rating, and publications judged relevant were examined in full text with respect to their concepts, methodologies, and applicability to adversarial cybersecurity research. The shortlisted studies were organized into five thematic categories: adversarial attack techniques, targeted cybersecurity domains, machine-learning vulnerabilities, defensive architectures, and future security challenges. Particular attention was given to intrusion detection, malware classification, autonomous systems, wireless communication security, textual adversarial attacks, explainable artificial intelligence, and black-box attack scenarios, given their growing importance in contemporary cybersecurity architectures.

2.5 Analytical Framework

Thematic analysis was used to integrate and interpret the reviewed literature. Selected studies were compared and contrasted to identify recurring attack patterns, dominant defensive responses, implementation challenges, and emerging research directions. The framework emphasized the relationship between attack sophistication and defense robustness across cybersecurity applications. The limitations of current mitigation methods, including adversarial training, defensive distillation, feature squeezing, explainable artificial intelligence, and robust optimization, were also analyzed, and a comparative evaluation assessed the adaptability, scalability, computational overhead, and real-time usability of existing defense frameworks under changing adversarial conditions. Rather than emphasizing individual experimental results, the analysis prioritized synthesis and conceptual consistency across studies to surface broader insights, research gaps, and trends.

2.6 Reliability and Study Limitations

To strengthen reliability, preference was given to recent peer-reviewed studies, and multiple databases were consulted to reduce selection bias and broaden coverage. Structured screening criteria further supported methodological uniformity. Several limitations are nonetheless acknowledged. The field evolves quickly, so new attack and defense methods may emerge after the review window. The qualitative synthesis is also subject to heterogeneity across the primary studies, whose differing datasets, metrics, and experimental setups limit direct comparability. To partially address the absence of primary experimentation common to review studies, this work additionally includes a controlled, reproducible case study (Section 4) that empirically demonstrates an evasion attack and an adversarial-training defense on a standard intrusion-detection benchmark.

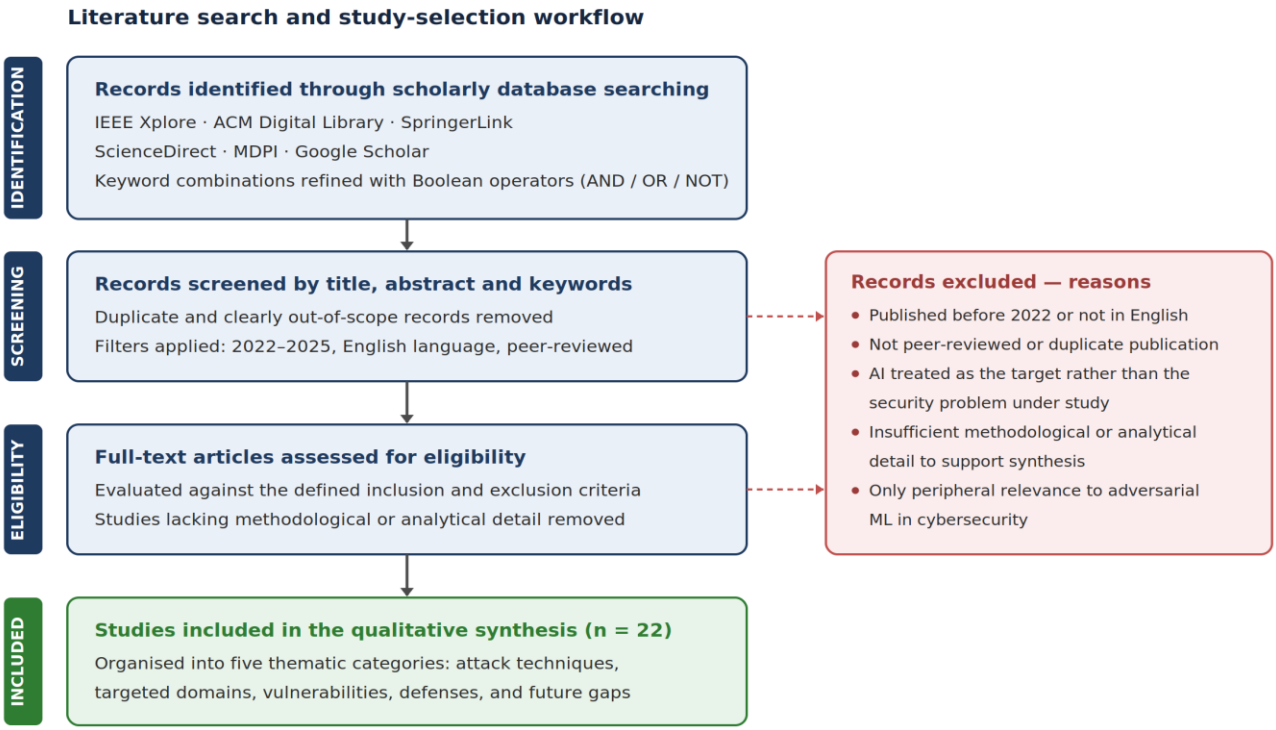

Figure 1. Literature search and study-selection workflow, showing the identification, screening, eligibility, and inclusion stages that yielded the final corpus of 22 studies.

3. RESULTS AND ANALYSIS

3.1 Overview of the Adversarial Machine Learning Threat Landscape

The reviewed literature indicates that adversarial machine learning is among the most pressing threats to contemporary cybersecurity infrastructures that rely on advanced ML techniques. Across the analyzed studies, adversarial threats are pervasive and capable of compromising the reliability, integrity, confidentiality, and operational effectiveness of machine-learning systems deployed in security-sensitive environments. The growing adoption of deep-learning architectures has widened the attack surface available to adversaries and enabled increasingly advanced model-manipulation techniques [3], [5]. The temporal and thematic distribution of the reviewed corpus is summarized in Figure 2, and the individual studies are catalogued in Table 1.

A recurring theme is the migration of adversarial attacks from theory to practice. Much early AML research was conducted in controlled environments in which attacks were simulated to characterize their behavior [14], [15]; more recent work, however, documents adversarial attacks within operational settings such as wireless networks, autonomous systems, malware classifiers, intrusion detection systems, and cloud-based security platforms [11], [16]. This transition indicates that adversarial machine learning has matured from a research curiosity into an applied threat with direct implications for digital-security resilience.

The literature further notes that adversarial attacks are becoming more adaptive and automated. Several studies report that attackers employ iterative optimization to probe ML models and generate highly evasive yet functional adversarial samples [4], [17]. Such samples substantially complicate conventional detection, because the appearance or behavior of malicious inputs closely resembles that of legitimate data.

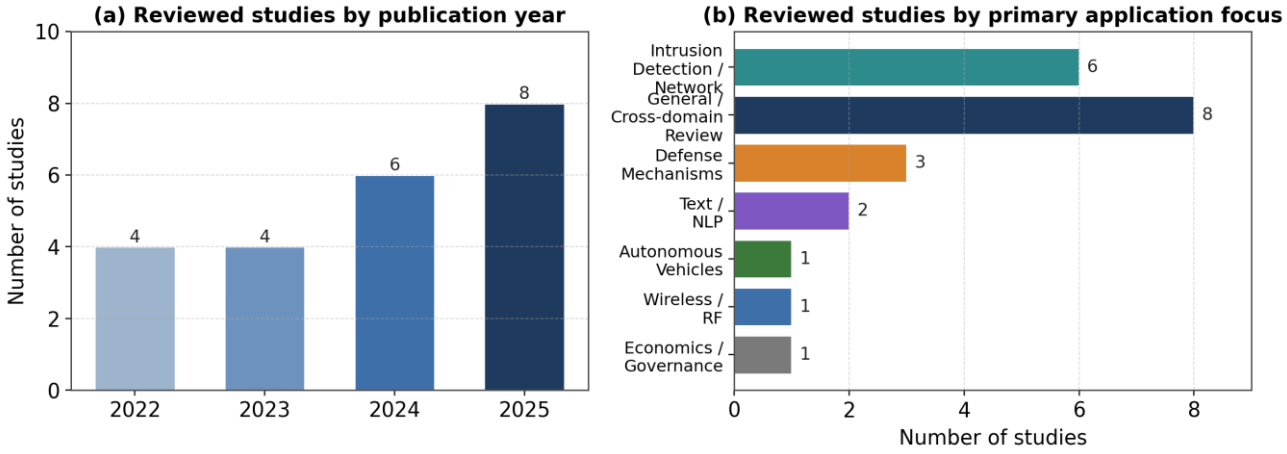

Figure 2. Bibliometric synthesis of the reviewed corpus: (a) studies by publication year; (b) studies by primary application focus.

Table 1. Summary of the 22 reviewed studies, with primary focus, attack types considered, and main contribution.

Ref.	Study (Year)	Primary Focus / Domain	Attacks Considered	Main Contribution
[1]	Ejeofobiri et al. (2024)	AI in cybersecurity (review)	Threat detection	Survey of AI threat detection, response, and prevention
[2]	Ali et al. (2025)	AI-cybersecurity fusion (review)	Cross-domain	Trends, advanced techniques, and policy implications
[3]	Malik et al. (2024)	AML attacks and controls (SLR)	Evasion, poisoning	Robustness methods and adversarial-training analysis
[4]	Zhou et al. (2023)	Deep-learning attacks and defenses	Evasion, inversion	Cybersecurity-oriented attack/defense taxonomy
[5]	Pelekis et al. (2025)	AML methods and sectors	Inference, inversion	Methods, tools, and critical-sector mapping
[6]	Alotaibi & Rassam (2023)	IDS evasion (survey)	Evasion	Attack strategies and defenses for IDS
[7]	Girdhar et al. (2023)	Autonomous vehicles (SLR)	Evasion (perception)	Attack and defense models for AV security
[8]	Adesina et al. (2023)	Wireless / RF ML (review)	Evasion (RF)	Review of RF adversarial attacks and defenses
[9]	Vaccari et al. (2022)	XAI-based defense	Evasion	Explainable, reliable detection of manipulation
[10]	Barik, Misra & Lopez-Baldominos (2025)	Black-box defense	Black-box evasion	Empirical black-box defense analysis
[11]	Apruzzese et al. (2022)	Realistic NIDS attacks	Evasion, poisoning	Realistic adversarial threat modeling for NIDS
[12]	Merkle et al. (2024)	Economics of AML	Cost analysis	Economic framing of adversarial defense
[13]	Kiranbabu et al. (2025)	AI-driven security challenges	Multiple	Challenge analysis for AI-driven security
[14]	McCarthy et al. (2022)	Functionality-preserving AML (survey)	Evasion	Survey of robust classification for IDS
[15]	Ododo & Sadiq (2025)	AML perspective	Multiple	ML-perspective overview of attacks
[16]	Ennaji et al. (2025)	NIDS adversarial challenges	Evasion, poisoning	Research insights and future prospects
[17]	Barik & Misra (2024)	Empirical defense analysis	Evasion	Defense-evaluation approach in cybersecurity
[18]	Khan & Ghafoor (2024)	Network-security AML	Poisoning, evasion	Challenges and solutions for network security
[19]	Ke et al. (2025)	AML attacks and defenses	Backdoor, evasion	Overview of attacks and corresponding defenses
[20]	Paya et al. (2024)	Apollon defense (IDS)	Evasion	Robust defense system for intrusion detection
[21]	Jiang et al. (2025)	Cybersecurity NER	Textual	Textual adversarial attacks on entity recognition
[22]	Alsmadi et al. (2022)	Text-processing AML (survey)	Textual	Literature survey of text-based attacks

3.2 Classification of Adversarial Attacks

The reviewed studies identify several attack categories targeting ML-based cybersecurity systems. Although specific techniques vary across operational settings, the literature consistently distinguishes evasion, poisoning, model inversion, membership inference, and backdoor attacks, together with the orthogonal distinction between black-box and white-box threat models. The principal categories are summarized in Table 2, and a consolidated taxonomy is depicted in Figure 3.

3.2.1 Evasion Attacks

Evasion attacks are the most extensively studied category. They modify inputs to mislead trained models at inference time while leaving the model itself unchanged. The reviewed studies show that evasion is especially effective against intrusion detection and malware classification systems, where a small change to an input can drastically alter the classification outcome without affecting the malicious functionality of the input [6].

Multiple studies report that deep-learning classifiers are highly susceptible to gradient-based evasion, particularly in black-box settings where the adversary has limited knowledge of model internals [10]. Adversarial examples produced through optimization can evade anomaly detectors because they exhibit only minor statistical deviations from normal traffic.

3.2.2 Poisoning Attacks

Poisoning attacks were consistently identified as among the most damaging threats during the training phase. The adversary manipulates the training set by injecting false or malicious examples that distort the model's learned behavior. The literature reports that poisoning can substantially degrade detection accuracy, raise false-negative rates, and reduce overall system reliability [3].

Poisoning is particularly harmful in distributed and continually learning systems that ingest data from automated or external sources. Studies of network intrusion detection show that even modest poisoning can shift decision boundaries and impair a model's ability to generalize under operational conditions [11], [18].

3.2.3 Model Inversion and Membership Inference Attacks

The reviewed studies highlight growing attention to privacy-oriented attacks such as model inversion and membership inference. Model inversion seeks to reconstruct sensitive data from model outputs, whereas membership inference determines whether a particular sample was part of the training set [5].

Such attacks pose significant confidentiality risks for biometric authentication, healthcare, and identity-centric security systems. Several works note that high-capacity deep-learning models can inadvertently leak private information through prediction confidence scores and inference behavior, raising fundamental privacy concerns [4].

3.2.4 Backdoor and Trojan Attacks

Backdoor, or Trojan, attacks were singled out as especially stealthy because they allow adversaries to embed malicious behavior in a model during training. A compromised model behaves normally under ordinary conditions but produces attacker-chosen outputs when a specific trigger is present [19].

The literature indicates that backdoor attacks are difficult to detect, since performance on clean data may be indistinguishable from that of a benign model. This characteristic heightens the risk associated with model outsourcing, third-party pretrained models, and collaborative learning.

Table 2. Classification of adversarial machine learning attacks in cybersecurity systems.

Attack Type	Attack Phase	Primary Objective	Targeted Systems	Potential Impact
Evasion	Inference / testing	Deceive trained models by modifying malicious inputs without altering attack functionality	IDS, malware classifiers, spam filters	Reduced accuracy, more false negatives, security bypass
Poisoning	Training	Manipulate training data and corrupt model learning behavior	Network-security models, adaptive and federated learning	Model degradation, biased predictions, compromised reliability
Model Inversion	Inference	Reconstruct sensitive training data from model outputs	Biometric and healthcare security applications	Privacy leakage, unauthorized data reconstruction
Membership Inference	Inference	Determine whether specific samples were in the training set	Cloud-based ML, identity-management systems	Exposure of confidential user and training-data information
Backdoor / Trojan	Training and deployment	Implant hidden behavior triggered under specific conditions	Pretrained, outsourced, and autonomous ML systems	Covert manipulation, unauthorized model behavior
Black-box	Inference	Exploit vulnerabilities without knowledge of model architecture	Commercial and cloud-hosted ML applications	Difficult detection, transferable exploitation
White-box	Training and inference	Exploit full access to parameters and architecture	Research environments, exposed AI infrastructures	Highly optimized perturbations, severe robustness loss

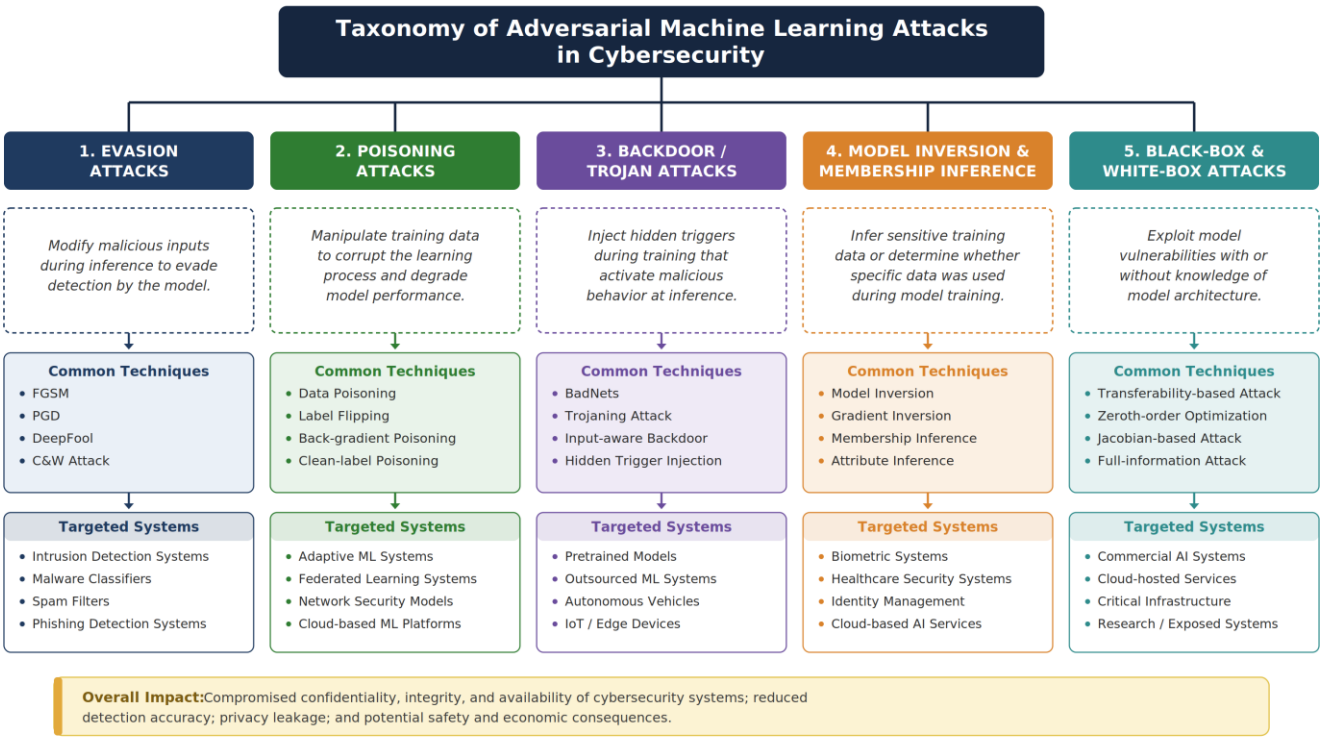

Figure 3. Taxonomy of adversarial machine learning attacks across contemporary cybersecurity environments.

3.3 Impact of Adversarial Attacks on Cybersecurity Applications

Adversarial attacks affect cybersecurity domains to differing degrees. Intrusion detection systems were among the most vulnerable applications identified in the literature: adversarial perturbations consistently degrade detection performance and substantially increase misclassification rates in network-security settings [16], [20].

Autonomous systems and intelligent transportation infrastructures were also found to be highly exposed. Girdhar et al. [7] show that adversarial attacks can deceive autonomous-vehicle perception systems, disrupting object recognition and decision-making and thereby introducing serious safety risks.

Wireless communication systems are similarly susceptible to adversarial interference against RF-based machine-learning models. Adesina et al. [8] report that carefully crafted perturbations in wireless environments can markedly impair communication reliability and signal-classification accuracy.

Textual adversarial attacks are increasingly prevalent against natural-language-processing systems used for phishing detection, threat-intelligence extraction, and cybersecurity entity recognition. Jiang et al. [21] demonstrate that subtle textual modifications can leave semantic readability largely intact while substantially changing model interpretation, and broader surveys document the breadth of text-based attack strategies [22].

3.4 Evaluation of Defense Mechanisms

The reviewed studies show that many defense techniques can improve the robustness of machine-learning systems against adversarial threats, but the literature is clear that no single strategy offers complete protection across all conditions. A comparative summary of major defenses and their operational characteristics is presented in Table 3, and a conceptual layered-defense architecture is illustrated in Figure 4.

3.4.1 Adversarial Training

Adversarial training is among the most widely used defenses. By exposing models to adversarial samples during training, it improves robustness against subsequent attacks, and several studies report increased resistance to known attack patterns [3]. However, the literature notes that adversarial training can be computationally expensive and offers limited protection against novel, previously unseen attacks.

3.4.2 Defensive Distillation and Robust Optimization

Defensive distillation and robust optimization were examined as means of reducing model sensitivity to adversarial perturbations. The reviewed studies suggest they can increase stability against controlled attacks but often underperform when confronted by adaptive adversaries that dynamically adjust their strategies [17].

3.4.3 Explainable Artificial Intelligence-Based Defenses

Several recent studies emphasize the growing relevance of explainable artificial intelligence (XAI) within adversarial-defense frameworks. Interpretability techniques can improve model transparency [9] and help analysts detect anomalous decision-making indicative of adversarial manipulation. Explainable architectures are expected to play an important role in future detection systems, particularly where critical infrastructure demands interpretability and trust.

3.4.4 Hybrid and Layered Defense Architectures

The literature increasingly favors hybrid, multi-layered defenses that combine preprocessing, anomaly detection, adversarial training, and explainability. Studies of layered architectures report stronger protection against both black-box and white-box attacks than reactive defenses alone [20]. Nonetheless, several works caution that scalability, computational complexity, and deployment overhead remain challenges when such defenses are applied in real-time cybersecurity environments.

Table 3. Comparative analysis of defense mechanisms against adversarial attacks.

Defense Mechanism	Core Principle	Strengths	Limitations	Practical Effectiveness
Adversarial Training	Incorporates adversarial samples during training to improve robustness	Improves resistance to known attacks; enhances resilience	Computationally expensive; limited generalization to unseen attacks	Moderately effective against white-box and gradient-based attacks
Defensive Distillation	Smooths prediction boundaries to reduce model sensitivity	Enhances stability under certain conditions	Vulnerable to adaptive and stronger attacks	Effective in controlled settings but limited in dynamic ones
Feature Squeezing	Reduces input complexity to limit perturbations	Simple to implement; reduces attack surface	May reduce accuracy; weak against advanced attacks	Suitable as a supplementary defense
Input Preprocessing	Filters or transforms inputs before evaluation	Detects and removes suspicious perturbations	Attackers may design preprocessing-aware attacks	Useful for lightweight, real-time protection
Robust Optimization	Trains under worst-case adversarial assumptions	Improves overall robustness and stability	High computational and training complexity	Strong theoretical robustness but limited scalability
Explainable AI Defense	Uses interpretability to flag abnormal behavior	Enhances transparency and analyst trust	Cannot, alone, fully prevent attacks	Valuable for investigation and anomaly interpretation
Hybrid / Layered Defense	Combines multiple strategies into one framework	Broad protection against diverse attacks	Increased complexity and resource demands	Among the most promising for practical deployment
Anomaly Detection	Identifies deviations from normal patterns	Effective against previously unseen attacks	High false-positive rates in dynamic settings	Effective when integrated with adaptive monitoring

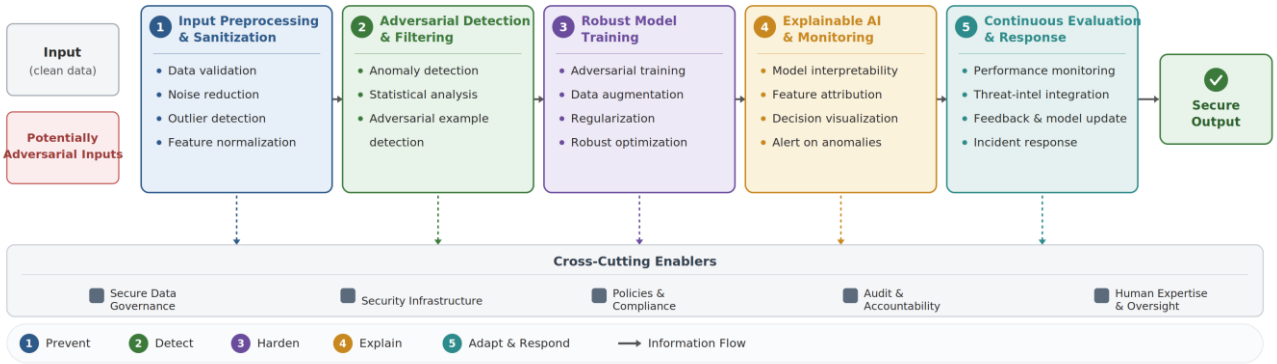

Figure 4. Layered defense framework for adversarial machine learning security.

3.5 Emerging Research Trends and Identified Gaps

Several trends emerge from the reviewed literature. Research is shifting toward adaptive defenses that can adjust dynamically to evolving attack strategies, and increasing attention is directed at securing federated learning and distributed AI infrastructures within cybersecurity ecosystems [2].

The analysis also reveals persistent research gaps, including the lack of standardized evaluation metrics, limited use of realistic adversarial benchmarks, insufficient cross-domain transferability analysis, and unresolved questions around scalable defense deployment. Many studies still rely on controlled laboratory settings that do not fully reflect operational conditions. These gaps motivate the controlled experiment reported next, which quantifies the impact of an evasion attack and the protection offered by adversarial training on a widely used intrusion-detection benchmark.

4. EXPERIMENTAL CASE STUDY

To ground the preceding synthesis in measurable evidence, this section reports a controlled and fully reproducible experiment that instantiates two of the review's central themes: the effectiveness of evasion attacks against undefended intrusion detectors and the protective value of adversarial training. The experiment was

implemented from first principles and executed on real data; all reported figures are produced directly by the experimental pipeline rather than drawn from secondary sources.

4.1 Objective

The experiment evaluates how an intrusion-detection classifier behaves under a Fast Gradient Sign Method (FGSM) evasion attack of increasing strength, and whether adversarial training restores robustness. FGSM was selected because it is a canonical, well-understood gradient-based attack [23], and adversarial training is adopted as the corresponding defense following the robust-optimization formulation of Madry et al. [24].

4.2 Dataset and Preprocessing

The experiment uses the NSL-KDD benchmark [25], a refined version of the KDD'99 intrusion-detection dataset that removes redundant records and mitigates known bias. The standard training and test partitions were used without modification. Categorical attributes (protocol, service, and flag) were one-hot encoded, producing a 122-dimensional feature vector, and all features were scaled to a common range. Each record was mapped to a binary label distinguishing normal traffic from attacks. The composition of the dataset is reported in Table 4.

Table 4. Composition of the NSL-KDD dataset used in the experiment.

Subset	Normal	Attack	Total
Training (KDDTrain+)	67,343	58,630	125,973
Test (KDDTest+)	9,711	12,833	22,544

4.3 Model and Attack Configuration

A feed-forward neural network with two hidden layers (128 and 64 units, ReLU activations) was trained as the intrusion detector using the Adam optimizer and a cross-entropy objective. The FGSM attack was applied at inference time across a range of perturbation budgets (epsilon from 0 to 0.30); perturbations were restricted to continuous features to preserve the semantic validity of categorical fields. For the defense, a second model of identical architecture was trained with adversarial training, in which FGSM-perturbed examples were incorporated into each training batch. Both models were evaluated on identical clean and adversarial test sets to ensure a fair comparison.

4.4 Results

On clean test data, the baseline detector achieved 80.8% accuracy, 96.6% precision, 68.6% recall, and an F1-score of 80.2%. The adversarially trained detector achieved 77.9% accuracy, 92.4% precision, 66.6% recall, and an F1-score of 77.4%, reflecting the small clean-accuracy cost that typically accompanies robustness. These clean-performance figures are summarized in Table 5.

Table 5. Clean-data performance of the baseline and adversarially trained detectors.

Metric (%)	Baseline	Adv.-trained
Accuracy	80.8	77.9
Precision	96.6	92.4
Recall	68.6	66.6
F1-score	80.2	77.4

Under attack, the two models diverge sharply. As shown in Table 6 and Figure 5, the baseline detector's accuracy falls from 80.8% to 33.2% at epsilon = 0.05, to 24.6% at epsilon = 0.10, and to 7.3% at epsilon = 0.20, approaching 0% at epsilon = 0.30. The adversarially trained detector, by contrast, retains 75.1%, 74.3%, and 73.3% accuracy at the same budgets and still achieves 70.3% accuracy at epsilon = 0.30. Adversarial training therefore converts an almost complete loss of detection capability into a graceful and bounded degradation.

Table 6. Detection accuracy and F1-score under FGSM evasion at increasing perturbation budgets.

FGSM budget (epsilon)	Undefended accuracy (%)	Undefended F1 (%)	Adv.-trained accuracy (%)	Adv.-trained F1 (%)
0.00	80.8	80.2	77.9	77.4
0.01	71.9	72.2	77.3	76.7
0.02	59.3	63.2	76.6	75.9
0.05	33.2	47.3	75.1	74.0
0.10	24.6	39.5	74.3	72.9
0.15	21.6	35.5	73.7	72.1
0.20	7.3	13.5	73.3	71.4
0.30	0.0	0.0	70.3	67.2

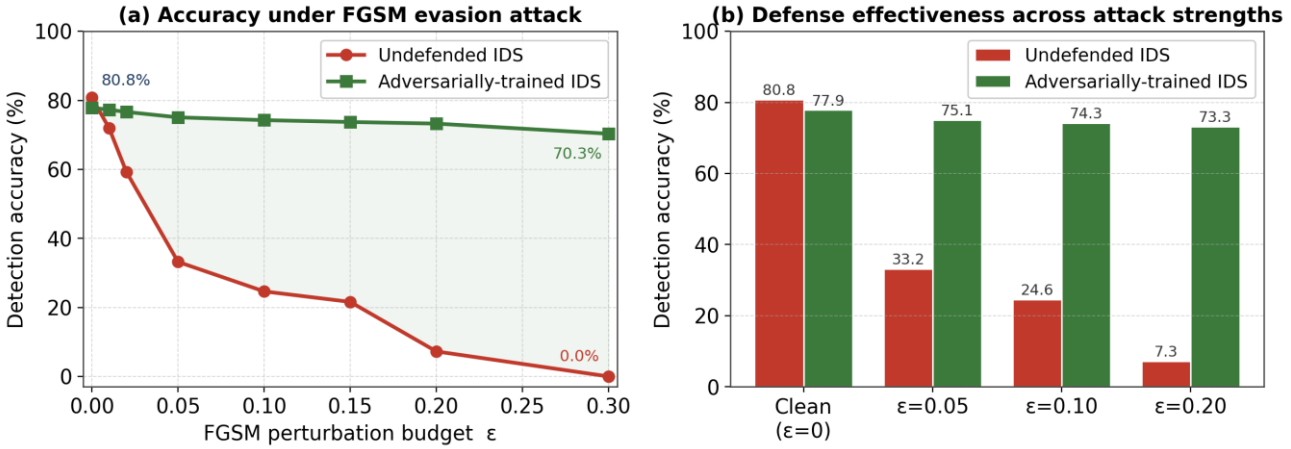

Figure 5. Experimental results on NSL-KDD: (a) detection accuracy as a function of the FGSM perturbation budget; (b) defense effectiveness at selected attack strengths.

4.5 Discussion of Experimental Findings

The experiment provides concrete support for two claims that recur throughout the reviewed literature. First, undefended deep-learning detectors are highly fragile: an imperceptible, low-budget perturbation is sufficient to collapse detection accuracy, confirming that high clean-data accuracy is not a reliable indicator of operational robustness. Second, adversarial training is an effective and practical countermeasure that preserves most of the detection capability across the full range of attack strengths.

The results also reproduce the well-documented robustness-accuracy trade-off: the hardened model sacrifices a few percentage points of clean accuracy in exchange for a large gain in adversarial robustness. For security-critical deployments, where the cost of a missed detection is high, this trade-off is generally favorable. At the same time, the experiment is deliberately scoped to a single attack and dataset; consistent with the gaps identified in Section 3.5, broader evaluation across multiple attacks, datasets, and adaptive adversaries remains an important direction for future work.

5. DISCUSSION

Taken together, the review and the experiment indicate that adversarial machine learning is a multidimensional problem spanning not only the weaknesses of individual models but also data integrity, reliability, trustworthiness, and governance of AI-enabled systems. The literature consistently describes a paradox: integrating machine learning into cybersecurity improves automated defense while simultaneously introducing new, exploitable vulnerabilities. This duality reflects the adversarial dynamic in which more capable automated systems are met by correspondingly more capable attacks.

A dominant theme is the increasing operational realism of attacks. Whereas early research was largely confined to laboratory conditions, recent studies, and the experiment reported here, demonstrate that adversarial attacks can effectively target real cybersecurity systems, particularly intrusion detection systems, malware classifiers, autonomous infrastructures, and wireless platforms [11], [16]. Adversarial machine learning should therefore be regarded not as a speculative concern but as a maturing threat capable of affecting intelligent security infrastructure at scale.

The findings also expose the structural tension that adversarial attacks exploit: the gap between predictive accuracy and robustness. High-performing deep-learning models generally excel under normal conditions yet remain acutely sensitive to small, carefully designed perturbations that are often imperceptible to humans [4]. The experimental collapse of the undefended detector concretely illustrates this fragility and reinforces the conclusion that robustness must be treated as a first-class objective alongside accuracy.

Poisoning attacks present a complementary, longer-term risk by corrupting the training process itself. In continually learning environments that adapt to new threats, poisoning can gradually erode model integrity [3], a concern that is especially acute in cloud-based and distributed settings dependent on automated data collection. These observations underscore the importance of data governance, provenance, and training-pipeline verification in addition to model-level defenses.

The growing overlap between adversarial machine learning and privacy is a further concern. Membership inference and model inversion threaten sensitive information embedded in trained models [5], with significant implications for biometric security, healthcare, financial systems, and identity management. The reviewed literature indicates that high-capacity models can leak training characteristics through their probabilistic outputs, enabling privacy breaches and unauthorized data reconstruction.

Defensive progress notwithstanding, current mechanisms face technical and operational limits. Adversarial training, though effective, can be attack-specific and computationally costly [17], and it may not generalize to adaptive adversaries, an asymmetry the experiment also reflects, since the defense was strongest against the attack family on which it was trained. Explainable AI offers valuable support for detection, accountability, and governance [9] but is insufficient on its own and must be combined with prevention, detection, response, and monitoring. The diversity of attack surfaces, spanning text, wireless, autonomous, and IoT systems [8], [21], further complicates standardization and argues for context-aware, adaptive defenses. Finally, the economic burden of sustaining robustness [12] and the absence of consistent evaluation methodologies across studies remain significant obstacles to universally accepted robustness benchmarks.

6. CONCLUSION AND FUTURE WORK

Adversarial machine learning is among the most significant challenges to the reliability, security, and trustworthiness of AI-powered cybersecurity systems. This study combined a structured review of 22 recent peer-reviewed works with a reproducible experiment to characterize the adversarial threat landscape and the defenses developed to counter it. The review shows that evasion, poisoning, backdoor, and inference-based attacks are growing in sophistication and increasingly endanger critical

applications such as intrusion detection, autonomous platforms, and intelligent network infrastructures, while the experiment demonstrates concretely that an undefended detector can be driven to near-zero accuracy by a simple evasion attack and that adversarial training restores robust, bounded performance.

The evidence indicates that resilient machine learning cannot rely on isolated defensive techniques. Future cybersecurity frameworks should integrate adaptive defense architectures, explainable AI, robust training, secure data governance, and continuous adversarial testing into coherent, multi-layered ecosystems rather than treating these components in isolation. Priority directions include standardized robustness benchmarks, evaluation against adaptive and cross-domain attacks, and scalable, real-time defenses suitable for operational deployment. As adversarial techniques continue to evolve, building robust, transparent, and trustworthy AI will be essential to securing digital ecosystems against an expanding threat landscape.

7. REFERENCES

[1] Ejeofobiri, C. K., Fadare, A. A., Fagbo, O. O., Ejiofor, V. O., and Fabusoro, A. T. 2024. The role of artificial intelligence in enhancing cybersecurity: A comprehensive review of threat detection, response, and prevention techniques. International Journal of Science and Research Archive, 13(2), 310-316.

[2] Ali, S., Wang, J., and Leung, V. C. M. 2025. AI-driven fusion with cybersecurity: Exploring current trends, advanced techniques, future directions, and policy implications for evolving paradigms - A comprehensive review. Information Fusion, 118.

[3] Malik, J., Muthalagu, R., and Pawar, P. M. 2024. A systematic review of adversarial machine learning attacks, defensive controls, and technologies. IEEE Access, 12, 99382-99421.

[4] Zhou, S., Liu, C., Ye, D., Zhu, T., Zhou, W., and Yu, P. S. 2023. Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity. ACM Computing Surveys, 55(8).

[5] Pelekis, S., Koutroubas, T., Blika, A., Berdelis, A., Karakolis, E., Ntanos, C., and Askounis, D. 2025. Adversarial machine learning: A review of methods, tools, and critical industry sectors. Artificial Intelligence Review, 58(8).

[6] Alotaibi, A., and Rassam, M. A. 2023. Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense. Future Internet, 15(2).

[7] Girdhar, M., Hong, J., and Moore, J. 2023. Cybersecurity of autonomous vehicles: A systematic literature review of adversarial attacks and defense models. IEEE Open Journal of Vehicular Technology, 4, 417-437.

[8] Adesina, D., Hsieh, C. C., Sagduyu, Y. E., and Qian, L. 2023. Adversarial machine learning in wireless communications using RF data: A review. IEEE Communications Surveys and Tutorials, 25(1), 77-100.

[9] Vaccari, I., Carlevaro, A., Narteni, S., Cambiaso, E., and Mongelli, M. 2022. explainable and reliable against adversarial machine learning in data analytics. IEEE Access, 10, 83949-83970.

[10] Barik, K., Misra, S., and Lopez-Baldominos, I. 2025. Black-box adversarial attack defense approach: An empirical analysis from cybersecurity perspective. Results in Engineering, 26.

[11] Apruzzese, G., Andreolini, M., Ferretti, L., Marchetti, M., and Colajanni, M. 2022. Modeling realistic adversarial attacks against network intrusion detection systems. Digital Threats: Research and Practice, 3(3).

[12] Merkle, F., Samsinger, M., Schoettle, P., and Pevny, T. 2024. On the economics of adversarial machine learning. IEEE Transactions on Information Forensics and Security, 19, 4670-4685.

[13] Kiranbabu, M. N. V., Jeraldine Viji, A., Chandanan, A. K., Birchha, V., Pandey, T. K., and Sar, S. K. 2025. The challenge of adversarial attacks on AI-driven cybersecurity systems. Journal of Cybersecurity and Information Management, 15(1), 288-297.

[14] McCarthy, A., Ghadafi, E., Andriotis, P., and Legg, P. 2022. Functionality-preserving adversarial machine learning for robust classification in cybersecurity and intrusion detection domains: A survey. Journal of Cybersecurity and Privacy, 2(1), 154-190.

[15] Ododo, F. R., and Sadiq, R. R. 2025. Adversarial attacks in cybersecurity: A machine learning perspective. Journal of Science Innovation and Technology Research.

[16] Ennaji, S., De Gaspari, F., Hitaj, D., Kbidi, A., and Mancini, L. V. 2025. Adversarial challenges in network intrusion detection systems: Research insights and future prospects. IEEE Access, 13, 148613-148645.

[17] Barik, K., and Misra, S. 2024. Adversarial attack defense analysis: An empirical approach in cybersecurity perspective. Software Impacts, 21.

[18] Khan, M., and Ghafoor, L. 2024. Adversarial machine learning in the context of network security: Challenges and solutions. Journal of Computational Intelligence and Robotics, 4(1), 51-63.

[19] Ke, H., Xu, J., Wang, Y., Chen, H., and Shen, Z. 2025. Adversarial machine learning in cybersecurity: Attacks and defenses. International Journal of Management Science Research, 8(2), 26-33.

[20] Paya, A., Arroni, S., Garcia-Diaz, V., and Gomez, A. 2024. Apollon: A robust defense system against adversarial machine learning attacks in intrusion detection systems. Computers and Security, 136.

[21] Jiang, T., Liu, Y., and Cui, X. 2025. Textual adversarial attacks in cybersecurity named entity recognition. Computers and Security, 150.

[22] Alsmadi, I., Aljaafari, N., Nazzal, M., Alhamed, S., Sawalmeh, A. H., Vizcarra, C. P., and Al-Humam, A. 2022. Adversarial machine learning in text processing: A literature survey. IEEE Access, 10, 17043-17077.

[23] Goodfellow, I. J., Shlens, J., and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In Proceedings of the International Conference on Learning Representations (ICLR). arXiv:1412.6572.

[24] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. 2018. Towards deep learning models resistant to adversarial attacks. In Proceedings of the International Conference on Learning Representations (ICLR). arXiv:1706.06083.

[25] Tavallaee, M., Bagheri, E., Lu, W., and Ghorbani, A. A. 2009. A detailed analysis of the KDD CUP 99 data set. In Proceedings of the IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 1-6.