Adversarial Machine Learning: Attack Methods and Defense Mechanisms

SurveyForge

Abstract— Adversarial Machine Learning (AML) explores the vulnerabilities in machine learning models arising from adversarial attacks and develops defense mechanisms to counteract them. This survey comprehensively covers attack methodologies such as gradient-based perturbations and physical-world evasions, alongside defenses like adversarial training and preprocessing techniques. It analyzes the intricate relationship between attack strategies and defense responses across diverse domains, including computer vision and natural language processing, highlighting the critical trade-offs between robustness and efficiency. Despite many advances, challenges persist in ensuring scalability, mitigating adaptive adversaries, and enhancing real-world applicability. The paper identifies future research directions, emphasizing adaptive defenses, interdisciplinary approaches, and robust evaluation frameworks to achieve resilient and trustworthy AI applications. This survey serves as a foundational reference for researchers and practitioners seeking to navigate the dynamic landscape of adversarial machine learning, directing attention to underexplored areas and novel collaborative potentials that harmonize theoretical advances with practice-oriented solutions.

Index Terms—adversarial attacks methods, robustness evaluation frameworks, defense mechanisms strategies

1 Introduction

Adversarial Machine Learning (AML), an innovative field at the intersection of artificial intelligence and cybersecurity, investigates the vulnerabilities of machine learning (ML) models to adversarial perturbations and seeks to develop mechanisms ensuring their robustness. Adversarial examples—carefully crafted perturbations imperceptible to humans but capable of causing ML models to fail—have raised significant concerns about the deployment of these systems in safety-critical applications, including autonomous driving, healthcare, and cybersecurity [1], [2]. This subsection provides an overview of AML, encompassing its core concepts, motivations, and the challenges that underscore the necessity of this domain.

Adversarial examples exploit the inherent weaknesses of ML models, particularly their sensitivity to small, distribution-shifting signals in high-dimensional feature spaces [3], [4]. These vulnerabilities stem from the complex decision boundaries of deep neural networks (DNNs), which are prone to over-reliance on non-robust features [3], [5]. Empirical studies have demonstrated the transferability of adversarial examples, showcasing their capability to mislead diverse ML architectures across both white-box and black-box attack paradigms [6], [7]. This robustness of attack strategies across varying systems underscores the urgency to develop resilient defenses.

The evolution of defense mechanisms has advanced from adversarial training, which enhances robustness by integrating adversarial examples into the training process, to formal verification methods that theoretically bound model susceptibility [8], [9]. However, defenses are often limited by trade-offs between clean accuracy and adversarial robustness, as well as computational costs [10], [11]. Furthermore, adaptive attacks, crafted to undermine these defenses, emphasize the escalating complexity of this arms race [12].

Emerging research focuses on the interdependence between attack strategies and defense paradigms, revealing a continuous co-evolution within the AML landscape. For instance, universal adversarial perturbations, which mislead models across diverse inputs, highlight the generalized vulnerabilities of ML models while informing robust defense strategies [13]. Similarly, domain-specific advancements, such as adversarial defenses in graph networks and multimodal systems, illuminate unique challenges and solutions across varied application contexts [14], [15].

The significance of AML spans multiple domains, including autonomous systems, natural language processing, and network intrusion detection, where maintaining security and reliability is paramount [16], [17]. This intersectional relevance necessitates unified evaluation frameworks, standardization of metrics, and cross-disciplinary collaboration to advance the field [18], [19]. Addressing open challenges such as scalability, adaptive threats, and ethical implications remains critical to ensuring the safe deployment of ML systems in real-world scenarios [5], [19].

In encapsulating the essence of adversarial ML through its conceptual underpinnings and practical implications, this survey highlights the need for robust, interdisciplinary solutions to preserve the integrity of ML systems against evolving adversarial threats.

2 Taxonomy and Techniques of Adversarial Attacks

2.1 Evasion Attacks

Evasion attacks, among the most prevalent forms of adversarial attacks, exploit the inference phase of machine learning (ML) systems by introducing carefully designed perturbations to input data. These perturbations are crafted to be minimally perceptible yet capable of causing model

misclassifications. Unlike data poisoning or backdoor attacks, evasion attacks do not modify the model or its training data, targeting instead the vulnerabilities in a model's decision boundaries and generalization behavior [20], [21].

A foundational approach to evasion attacks leverages gradient-based optimization techniques, such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), which compute the model's loss gradients with respect to inputs to craft adversarial perturbations that maximize misclassification likelihood under a specified perturbation norm. FGSM operates via a single-step update and is computationally efficient but can be suboptimal for robust defenses, whereas PGD utilizes iterative updates with carefully constrained step sizes, often achieving stronger adversarial examples under the same distortion thresholds [3], [10]. The perturbation is commonly measured using distance metrics like L_1, L_2, or L_∞ norms, where L_∞ ensures bounded maximal deviation per input pixel but may fail to capture aggregate distortions in other domains [2].

While gradient-based techniques excel in white-box scenarios, where the attacker has full model access, they falter in black-box settings. In such cases, attackers resort to transferability—the property that adversarial examples crafted on one model often succeed against structurally similar or identically trained models. Advanced black-box approaches, such as query-based methods or the use of surrogate models, refine this concept. Query-based attacks employ heuristics like Natural Evolution Strategies (NES) to approximate gradients without explicit model access, while surrogate model methods allow attackers to train local replicas to mimic the behavior of the target system [5], [7]. However, black-box attacks are limited by query efficiency and suffer reduced reliability against more diverse architectures.

Evasion attacks expand further into real-world applications, such as physical perturbations that deceive autonomous systems. Modifications to physical objects, like adversarial road signs, created under constrained conditions (e.g., environmental noise), have proven effective at misleading deep learning systems in safety-critical applications, including autonomous driving [1], [22]. However, crafting such attacks requires addressing complex challenges like robustness to transformations (e.g., lighting, perspective).

Despite the sophistication of multiple techniques, evasion attacks face challenges with advanced defenses, including adversarial training and input-processing mechanisms. Additionally, the interplay between transferability, adaptation, and norm constraints remains a fertile area for exploration. Future directions may focus on extending attacks to dynamic, multimodal systems and addressing limitations posed by novel defense strategies, such as certifiable robustness and context-aware detection algorithms [18], [23]. Such initiatives would deepen our understanding of the inherent vulnerabilities in ML systems and guide the development of more resilient architectures.

2.2 Data Poisoning and Backdoor Attacks

Data poisoning and backdoor attacks represent two potent mechanisms by which adversaries compromise the integrity of machine learning (ML) models during their training phase, contrasting sharply with evasion attacks that target the inference phase. These methods exploit the vulnerability of data-driven learning frameworks by introducing maliciously crafted data into the training set, embedding adversarial objectives that either degrade overall model performance or enable the model to exhibit specific malicious behaviors at deployment.

Data poisoning broadly aims to manipulate class boundaries or destabilize model performance, often targeting the robustness of the training process itself. For instance, poisoned samples that appear statistically normal on cursory analysis can covertly shift decision boundaries, leading to notable degradation in model accuracy [24]. Early poisoning techniques demonstrated how linear classifiers, such as support vector machines (SVMs), are susceptible to targeted attacks, where adversaries use gradient-based optimization to amplify the impact of poisoning [25]. Poisoning deep learning models is more challenging due to the high dimensionality and inherent non-linearity of feature spaces. Advanced approaches like "Back-gradient Optimization" leverage automatic differentiation to craft poisoned data, enabling adversaries to undermine multi-class systems systematically [26]. Additionally, generative adversarial frameworks have emerged to craft poisoned data capable of circumventing statistical anomaly detection, forcing models to overfit misleading patterns [27]. While stealthy and impactful, these poisoning methods pose long-term threats to the reliability of ML systems, particularly in sensitive or large-scale deployments.

Backdoor attacks, unlike general poisoning attacks, embed dormant vulnerabilities into models during training, activating them only when triggered by specific input patterns. These triggers, often imperceptible or camouflaged, are designed to hijack predictions without raising suspicion. For example, pixel-level perturbations can be carefully embedded to blend seamlessly with benign training data, remaining undetectable during validation but inducing malicious behavior when encountered during inference [28]. Visual tasks, such as facial recognition, are particularly prone to these attacks, as adversaries can exploit natural artifacts like lighting or image texture to conceal their triggers [29]. More advanced strategies integrate clean-label attacks, where both the label and feature space align with legitimate data distributions, further complicating detection efforts by relying on generative poisoning methods [24]. This level of sophistication highlights the insidious nature of backdoor attacks as they blur the line between authentic and malicious training data.

Defenses against data poisoning and backdoor attacks remain an evolving area of research and often involve a combination of preprocessing, robust training, and anomaly detection techniques. Statistical methods that leverage robust estimators to filter anomalous data have shown effectiveness against simpler poisoning strategies. However, these methods often assume independence across data points, limiting their applicability against attacks designed with coordinated corruption [30]. When applied to high-dimensional models, such defenses can also incur high computational costs and may degrade accuracy on clean data. Backdoor-specific defenses focus on detecting or miti-

gating trigger-based manipulations, typically utilizing fine-grained data transformations, such as input compression, or inspecting feature space for latent perturbation patterns. For example, auditing feature maps in hidden layers has been shown to isolate trigger-related artifacts effectively [27].

Emerging research highlights the potential of hybrid frameworks that combine pre-training data validation with anomaly detection, sometimes leveraging generative adversarial networks to uncover poisoned data prior to model deployment [27]. Additionally, adversarial training paradigms are beginning to account for poisoning scenarios during synthesis, providing models with broader resilience toward perturbed training data. Still, significant challenges persist in defending against poisoning and backdoor attacks, particularly in decentralized systems such as federated learning. These environments exacerbate vulnerabilities by limiting centralized inspection capabilities for training data and amplifying adversarial opportunities while prioritizing user privacy [31].

In conclusion, the multifaceted vulnerabilities exposed by data poisoning and backdoor attacks underscore the need for comprehensive safeguards spanning anomaly detection, robust optimization, and distributed security paradigms. Unlike evasion attacks that exploit model fragilities post-deployment, these training-based threats destabilize the very foundation of ML systems, emphasizing the critical role of resilient training practices in securing the future of machine learning.

2.3 Attack Models Based on Adversarial Knowledge

Adversarial attacks on machine learning models are often characterized by the extent of the attacker's access to the target model's parameters, architecture, and input-output mappings. These attacks are broadly categorized into white-box, black-box, gray-box, and adaptive attacks, each reflecting a distinct level of adversarial knowledge and capability. Understanding their interplay reveals critical insights into the vulnerabilities and design of robust defensive systems.

White-box attacks assume complete access to the target model, encompassing architecture, gradients, hyperparameters, and training data distribution. This comprehensive knowledge enables highly optimized attacks, often leveraging gradient-based methods such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). Attackers can efficiently manipulate model predictions by calculating precise adversarial perturbations, as demonstrated in studies on evasion attacks [32]. Despite their potency, the practicality of white-box attacks is limited in real-world scenarios where such complete model access is rare.

Conversely, black-box attacks operate without direct access to the model internals, relying solely on query-response feedback or transferability principles. Attacks in this category exploit surrogate models trained to mimic the behavior of the target system. For example, substitute models have been effectively utilized to craft transferable adversarial examples capable of deceiving commercial systems hosted by Amazon and Google [33], [34]. A major limitation of black-box attacks, however, arises from the dependency on the fidelity of surrogate models and the number of queries required, often triggering detection mechanisms in practical deployments.

Gray-box attacks fall between these paradigms, assuming partial knowledge of the system. Attackers might, for instance, have access to the model architecture but not the training data or exact parameter values. In such cases, hybrid methodologies are employed, combining heuristic approximations of gradients with data augmentation to enhance attack efficacy. Studies have demonstrated the feasibility of such strategies in diverse domains, including poisoning attacks on node embeddings [35].

Adaptive attacks represent a dynamically adversarial approach wherein attackers modify their strategies based on the defense mechanisms in place. For example, defenses such as gradient obfuscation or adversarial training can be circumvented by attackers employing strategies specifically formulated to undermine these techniques, such as designing adaptive backdoors undetectable via statistical or spectral inspections [36], [37]. These attacks remain a persistent challenge due to their evolving nature and the arms race they perpetuate against defensive strategies.

Each adversarial knowledge model reveals distinct trade-offs between attack success rate, resource requirements, and practical viability. White-box attacks, while theoretically powerful, are constrained by their unrealistic assumptions, whereas black-box attacks face challenges related to query efficiency and surrogate modeling. Gray-box and adaptive frameworks, bridging varying levels of knowledge and adaptability, represent a middle ground where real-world relevance and theoretical robustness intersect.

Future research must explore domain-specific implications and scalable defenses against knowledge-dependent attacks. For instance, combining techniques from differential privacy with secure model deployment practices may obfuscate critical model details, deterring white-box and gray-box exploits [38]. Similarly, reducing over-reliance on vulnerable surrogate models in black-box settings remains an open area of investigation. A robust understanding of the adversary's knowledge and its constraints is essential for building resilient machine learning ecosystems capable of withstanding increasingly sophisticated threats.

2.4 Universal and Advanced Attack Modalities

Advanced adversarial attack methodologies transcend the boundaries of conventional perturbation-based strategies, uncovering universal and innovative mechanisms that exploit latent vulnerabilities in machine learning (ML) systems. These approaches often operate in task-agnostic settings, highlighting structural deficiencies in deep learning models and broadening the scope of adversarial research.

Universal adversarial perturbations (UAPs) signify a key evolution in adversarial techniques by crafting a single perturbation that remains effective across multiple input samples within a task. Unlike instance-specific perturbations, UAPs exploit shared decision boundaries inherent to models, showcasing their transferability across various architectures and datasets. These perturbations are derived by solving optimization problems over representative datasets to maximize misclassification rates, while adhering to constraints on the ℓ_p-norm and visual imperceptibility. Studies [13] [34] underscore how UAPs generalize effectively, presenting significant challenges to robustness and

detection frameworks. Yet, their susceptibility to targeted defenses, particularly methods like gradient regularization and adversarial data augmentation, curtails their universal applicability [39].

Semantic and context-based adversarial attacks pivot from low-level perturbations to the manipulation of higher-level features, particularly in multimodal models or task-specific feature extraction stages. By targeting models' underlying semantic representations, these methods induce errors that propagate across modalities, such as in vision-language frameworks. For instance, altering semantic embeddings can result in cross-modal misclassifications [15]. However, the often discrete and structured characteristics of multimodal inputs, such as the syntactic constraints in natural language or the topology in graph-based ML systems, complicate the generation of adversarial examples that preserve perceptual or functional integrity [16].

Distributional and spatio-temporal attacks target systems that rely on dynamic data patterns, such as autonomous vehicles or video recognition models, by manipulating either input distributions or temporal features. Instead of perturbing individual inputs, distributionally adversarial attacks recalibrate entire data distributions to amplify risk under spatio-temporal dependencies [40]. Similarly, trajectory perturbations in autonomous vehicle systems illustrate the challenges posed to safety-critical tasks like motion prediction and object tracking, where subtle disruptions in temporal behavior can cascade into severe operational failures [41].

Adaptive physical attacks aim to safeguard adversarial efficacy under real-world conditions, addressing variabilities such as environmental noise, occlusions, and object deformations. These attacks generate perturbations that remain effective across diverse physical scenarios, often using generative adversarial methods to dynamically adapt their payloads for persistent stealth and impact. For example, dynamic trigger patterns embedded in backdoor attacks [37] showcase real-time adaptation, enabling adversarial actions to remain potent under fluctuating settings. Despite their sophistication, such attacks face challenges in maintaining performance when subjected to rapidly changing environmental dynamics.

Recent innovations in adversarial research emphasize the application of meta-learning algorithms and generative approaches for advancing the adaptability and generalization of attacks. Leveraging meta-transferability across models enables efficient generation of adversarial samples with reduced query complexity, thereby enhancing the practicality of black-box attack techniques [42]. However, the computational overhead associated with these methods continues to pose significant scalability challenges.

The evolution of advanced adversarial methodologies unveils critical insights into the inherent vulnerabilities of ML systems. Bridging the gaps between adversarial robustness, cross-domain generalization, and real-world applicability requires interdisciplinary research focused on robust optimization, multimodal resilience, and comprehensive validations. Such efforts are pivotal for ensuring the trustworthiness and reliability of machine learning systems amid increasingly sophisticated adversarial landscapes.

2.5 Cross-domain and Task-specific Attacks

Adversarial attacks have demonstrated diverse manifestations across application domains, each defined by unique modalities, objectives, and underlying technical challenges. This subsection delves into task-specific adversarial attack methodologies, highlighting domain-specific vulnerabilities in vision, natural language processing (NLP), cybersecurity, and multimodal systems, with a focus on comparative analysis, emerging trends, and challenges.

In computer vision, adversarial attacks exploit the inherent susceptibility of visual models to small, imperceptible perturbations in image inputs. For example, gradient-based attacks targeting image classifiers achieve high misclassification rates by maximizing the perturbations' impact within l_p-bounded norms [1], [21]. Additionally, in physical contexts such as autonomous driving, adversarial patches and real-world perturbations like Robust Physical Perturbations (RP2) deceive models under varying environmental conditions, highlighting the difficulty of maintaining robustness in dynamic real-world scenarios [22], [43], [44]. These attacks expose critical safety concerns in vision-dependent systems, necessitating sophisticated defense strategies for robust deployment.

Conversely, in NLP systems, the crafting of adversarial text inputs involves preserving semantic coherence while misleading the model. Substitution-based attacks, leveraging word embeddings or imperceptible encoding changes, showcase significant efficacy while maintaining input readability [45], [46]. Attacks on large language models (LLMs) further reflect the dual vulnerability of these systems to textual perturbations and "jailbreak" attacks that bypass safety constraints [47], [48]. However, the constraints of linguistic structures and the adherence to grammar add substantial complexity to adversarial text generation, marking this domain distinct from vision-centric adversarial research.

In cybersecurity, adversarial attacks target intrusion detection systems (IDSs) or other machine learning-based cybersecurity tools. These attacks craft data perturbations that mimic real-world patterns, allowing malicious entities to bypass detection. For instance, adversarial botnets highlight the feasibility of realistic adversarial samples capable of avoiding IDS mechanisms [31], [49]. Furthermore, domain-specific constraints in cybersecurity, such as maintaining logical coherence in tabular data, shape the design of attacks like the Adaptive Perturbation Pattern Method (A2PM) [50].

Emerging threats to multimodal systems reveal a merging of domain-specific challenges. Vision-language pre-trained models (VLPs) combine vulnerabilities across both text and vision domains. Attacks leveraging adversarial perturbations in cross-modal embeddings create new dimensions of risk. Techniques like collaborative multimodal attacks demonstrate how perturbations in one modality, such as adversarial images, can propagate and disrupt aligned outputs [51], [52]. These challenges underscore the complexity of protecting systems that integrate heterogeneous inputs.

The convergence of task-specific and cross-domain adversarial research points to critical avenues for future work. Broader exploration of adaptive, multimodal, and task-

specific attack-defense dynamics is necessary to bridge the gap between research methodologies and real-world applications. Furthermore, standardization in benchmarking across domains can facilitate the development of resilient, transferable defenses capable of supporting trustworthy machine learning systems.

3 Defense Mechanisms in Adversarial Machine Learning

3.1 Adversarial Training and Data Augmentation

Adversarial training represents a cornerstone in enhancing the robustness of machine learning models against adversarial attacks by incorporating adversarial examples into the training process. Originating from early findings on the linear susceptibilities of neural networks to perturbations [3], adversarial training has evolved into a versatile and widely adopted defense mechanism. Fundamentally, it can be formulated as a min-max optimization problem that seeks to minimize the model's loss over the worst-case adversarial examples within a bounded perturbation space ϵ, such that $\min_\theta \max_{\delta \in S} \mathcal{L}(f_\theta(x+\delta), y)$, where f_θ denotes the model parameterized by θ, $\mathcal{L}$ the loss, and S the perturbation set. This approach effectively builds intrinsic robustness by exposing the model to adversarial scenarios throughout training [6].

Standard adversarial training commonly involves generating adversarial examples via methods like the Projected Gradient Descent (PGD) attack, which iteratively aligns perturbations with gradient directions to maximize the loss within ϵ constraints [2], [6]. PGD-based adversarial training has been shown to substantially enhance robustness against first-order attacks, earning it recognition as a baseline in adversarial defenses. However, this comes with trade-offs: high computational costs due to repeated gradient computations during adversarial example generation, and challenges in maintaining clean data performance, often referred to as the robustness-accuracy trade-off [2], [8].

Beyond standard methods, advanced techniques address limitations in both efficiency and generalization. Curriculum adversarial training progressively increases perturbation strength during training to balance robustness and accuracy [10]. Label smoothing integrates uncertainty into target labels to reduce overconfidence on perturbed inputs, while distributional adversarial training frames perturbations as adversarial shifts in data distributions, improving robustness in diverse threat models [8], [23]. Despite their improvements, these approaches must contend with adaptive attacks that specifically exploit their defensive structures [12].

Adversarial data augmentation complements adversarial training by enriching training sets with diverse adversarial samples, often crafted to reflect distinct attack modalities. By increasing exposure to varied perturbations, this approach strengthens the generalization of defenses [6]. Nonetheless, concerns about computational demands and dataset imbalance persist, emphasizing the need for automated augmentation frameworks that are both efficient and scalable [53].

Challenges in adversarial training include the tendency to overfit specific attack types, limited transferability across threat models, and scalability to large-scale datasets. Future research should explore hybrid strategies that combine adversarial training with preprocessing defenses and formal verification methods [9], [54]. Additionally, emerging work on utilizing adversarial training in non-traditional domains (e.g., graphs, speech, and multimodal systems) promises to broaden its applicability while addressing domain-specific vulnerabilities [8], [14].

3.2 Input Preprocessing and Transformation-based Defenses

Input preprocessing and transformation-based defenses serve as an essential layer in adversarial machine learning frameworks, aiming to mitigate adversarial perturbations by modifying input data prior to model inference. Functionally, these mechanisms transform the input space to diminish the impact of malicious perturbations while preserving the semantic integrity of the original data. Grounded in methods of denoising, encoding, and feature-space transformations, this subsection examines the primary categories of such defenses, evaluating their strengths, limitations, and the persistent challenges surrounding their deployment.

A prominent class of these defenses comprises denoising and filtering techniques, which function in either spatial or frequency domains to smooth or remove high-frequency perturbations often introduced by adversarial attacks. Gaussian filtering and wavelet-based denoising, for example, have shown efficacy in suppressing perturbations that disproportionately target high-frequency components [17], [43]. These methods operate on the assumption that adversarial perturbations primarily occupy such frequency bands. However, they face significant challenges in disentangling adversarial signals from overlapping legitimate high-frequency features in clean data, and their application can sometimes alter the fidelity of original inputs, leading to degradation in model accuracy.

Encoding and discretization transformations offer another avenue of preprocessing-based defenses by projecting inputs into alternate representations to disrupt adversarial patterns. JPEG compression, as a widely recognized approach, exploits the lossy nature of compression to remove gradient-based perturbations, albeit with potential loss of critical semantic content [17]. Similarly, bit-depth reduction seeks to minimize sensitivity to finely tuned adversarial perturbations through reduced numerical precision. While these techniques are computationally efficient, rigid encoding mechanisms can inadvertently make transformed data susceptible to new adaptive attacks [55].

Feature-space transformations represent a more sophisticated category of defenses, leveraging learned representations to neutralize adversarial artifacts while preserving semantic content. Defense-GAN, for instance, reconstructs inputs by mapping them to the latent space of clean data and regenerating plausible non-adversarial samples [27]. This type of approach is particularly appealing due to its potential to robustly counter diverse perturbations. However, challenges such as computational overhead during inference and the susceptibility of learned latent spaces to sophisticated adversaries continue to limit practical adoption [55].

Despite their utility, preprocessing-based defenses encounter inherent vulnerabilities against adaptive attacks, which are designed to exploit or circumvent the transformations being applied. Methods specifically targeting the transformations, such as those based on distributional shifts or transfer-based strategies, highlight the brittleness of these defenses under dynamic attack conditions [56]. Additionally, many techniques rely on predefined parameters (e.g., compression ratios, filter strengths) that restrict generalizability and effectiveness across diverse threat models.

In response to these challenges, emerging research explores hybrid strategies that synthesize preprocessing with other defense paradigms. For example, integrating preprocessing transformations with adversarial training has demonstrated promise in reducing perturbations while concurrently enhancing model robustness to transformed inputs [57]. Advances in neural perceptual threat models, which guide transformations by approximating human judgments of perceptual distortions, also present an exciting direction for optimizing preprocessing techniques [58].

In conclusion, input preprocessing and transformation-based defenses provide a valuable but incomplete solution for safeguarding machine learning systems from adversarial attacks. Bridging the gap between these methods and other complementary strategies, such as adversarial training or gradient-based techniques, is critical to achieving robust defense capabilities. Future work should focus on adaptive parameterization, efficient implementation, and threat model generalization to refine the practical applicability of these transformations within an integrated adversarial defense framework.

3.3 Gradient-based and Regularization Techniques

Gradient-based and regularization techniques are pivotal in addressing the adversarial vulnerabilities tied to excessive sensitivity of machine learning models to input perturbations. By influencing the way gradients are computed or utilized, these methods aim to either limit adversarial exploitability or enforce structured learning paradigms that inherently promote robustness. This subsection explores the theoretical underpinnings, diverse approaches, and challenges associated with these strategies.

At the forefront of gradient-based defenses is *gradient masking*, a technique designed to obscure or distort gradient information, making it more challenging for adversaries to generate effective perturbations. Representative methods include gradient clipping and shattering, where gradients are modified to reduce their usability for attacks. However, such methods often suffer from adaptive attacks that can circumvent the obfuscation by leveraging alternate model properties, as highlighted in [19]. Moreover, gradient masking may inadvertently introduce brittleness by causing failures under unexpected attack strategies or during normal inference conditions [20].

Beyond masking, *robust optimization techniques* take a more theoretically principled approach to mitigate adversarial vulnerabilities. These methods involve crafting loss functions that optimize both clean accuracy and adversarial robustness. A notable example is Cross-Lipschitz regularization, which penalizes large variations in model output relative to changes in inputs, effectively bounding adversarial gradients. Additionally, adversarial training—the controlled integration of adversarial examples during model training—can improve gradient stability. Despite its empirical success, adversarial training often incurs significant computational costs and exhibits limited generalizability across unseen attack scenarios, as examined in [59].

Regularization methods focused on enforcing structured learning, such as constraining intra-class feature compactness and maximizing inter-class separability, also hold promise for countering adversarial effects. These strategies reduce reliance on non-robust features—a recurring vulnerability exploited by adversarial examples, as discussed in [60]. Such approaches, while theoretically sound, often struggle to balance robustness with clean accuracy, posing an ongoing trade-off in practical implementations.

A critical weakness across gradient-based defenses lies in their susceptibility to adaptive adversaries capable of exploiting alternative pathways, e.g., model architecture or decision boundaries, as demonstrated explicitly in [61]. Furthermore, these defenses often operate within a narrow threat model, limiting their resilience to evolving or unforeseen attack strategies.

Emerging research suggests that future advancements in gradient-based defenses will likely require integration with complementary paradigms such as formal verification and interpretability [11]. By combining robust regularization with certifiable bounds, researchers may address robustness limitations while minimizing performance trade-offs. Emphasis should also be placed on designing computationally efficient methods to reduce prohibitive training costs, ensuring feasibility at scale. Greater exploration of hybrid strategies uniting adversarial perturbation analysis and gradient-regularized training could further propel advancements in this domain, rendering gradient-based defenses more versatile and resilient under diverse attack scenarios.

3.4 Detection and Runtime Mitigation Mechanisms

Adversarial detection and runtime mitigation mechanisms serve as dynamic safeguards in machine learning, focusing on identifying adversarial inputs during inference and diminishing their impact in real-world scenarios. These techniques play a vital role in ensuring the reliability of machine learning systems, especially when deployed in adversarial environments such as cybersecurity, autonomous systems, and critical infrastructure. Positioned alongside gradient-based and certifiability strategies, detection and mitigation form a complementary layer of protection, addressing limitations inherent in static defenses while responding adaptively to emerging threats.

Detection techniques primarily harness statistical properties or model behavior to distinguish adversarial inputs from benign data. Statistical anomaly detection methods, such as Mahalanobis distance-based approaches, capitalize on the statistical distribution of feature representations to detect atypical inputs. By measuring the deviations of inputs within the learned feature space, these techniques often use pre-trained generative models or density estimators to assign likelihoods, flagging suspicious inputs [27]. Though effective in capturing out-of-distribution data points, these methods are susceptible to high false-positive rates, partic-

ularly when natural variations of benign inputs lie outside the training data distribution.

Complementary to statistical approaches, behavioral monitoring methods examine the model's internal dynamics, such as gradients or activation patterns, to detect adversarial activity. Divergence in gradient magnitudes or shifts in confidence scores across defense layers can act as indicators of adversarial perturbations [62]. Confidence thresholding, which identifies inputs with abnormally low confidence scores, is computationally efficient and applicable to both black-box and white-box attacks. However, these mechanisms face difficulties against adaptive adversaries who craft perturbations intentionally designed to evade such thresholds [12].

Once adversarial inputs are detected, runtime mitigation mechanisms offer real-time countermeasures. Input transformations, like feature reconstruction through autoencoders or GANs, project adversarial inputs onto low-dimensional representations to neutralize perturbations [27]. Simultaneously, ensemble-based defenses aggregate predictions from multiple independently trained models to dilute the effectiveness of adversarial attacks [63]. Although robust, these approaches often entail significant computational overhead, including latency and resource consumption, thereby limiting their scalability in high-throughput systems.

Despite advances, adversarial detection and runtime defenses encounter several critical challenges. Adaptive attacks designed to evade detection mechanisms, such as query-efficient methods based on Bayesian optimization, reveal vulnerabilities in current strategies [56]. Likewise, high false-positive rates and resource-intensive mitigation hinder the practical deployment of these techniques in mission-critical applications such as network intrusion detection or autonomous navigation systems [64]. These limitations underscore the need for lightweight yet effective solutions that can balance robustness with efficiency.

Moving forward, recent innovations inspire promising pathways for addressing these challenges. For instance, stateful defenses leverage historical query information to identify patterns associated with adversarial behavior, providing an additional layer of detection [65]. Similarly, hybrid strategies that blend statistical anomaly detection with behavioral analysis could achieve improved detection accuracy while reducing false positives. Integrating these mechanisms with formal verification techniques offers the potential for constructing comprehensive and certifiable adversarial defense frameworks, bridging the gap between empirical effectiveness and theoretical guarantees. As adversarial machine learning continues to evolve, the interplay between runtime detection, mitigation, and certifiability holds immense promise for robustifying AI systems across diverse and dynamic environments.

3.5 Certifiable Robustness and Formal Verification

Certifiable robustness and formal verification stand at the frontier of adversarial machine learning, aiming to provide provable guarantees of a model's robustness against adversarial attacks within a predefined threat model. Unlike heuristic defenses that often succumb to adaptive attacks, these approaches offer mathematical certainty, thereby addressing the limitations posed by empirical evaluations. This subsection examines the theoretical foundations, practical advancements, and existing challenges in this promising domain.

Certifiable robustness techniques seek to quantify worst-case guarantees under specific adversarial threat sets, typically expressed as perturbations constrained by ℓ_p norms, such as ℓ_∞, ℓ_2, or ℓ_1. This often hinges on constructing robustness bounds that certify whether a model's predictions remain invariant across all possible perturbations within a defined radius. Methods like convex relaxations and interval bounds have achieved notable success in this regard. Semidefinite programming and convex adversarial polytope formulations have enabled tighter bounds by approximating or modeling linear and quadratic constraints in multi-layer neural networks [4]. However, these approaches often face scalability issues due to the computational complexity involved, especially when applied to deep networks or large datasets.

Robustness verification frameworks play a pivotal role in automating the certification process. Tools such as ERAN and AI2 implement techniques based on abstract interpretation, symbolic interval propagation, and linear relaxation to verify robustness properties for neural networks [39]. While these tools have reduced the computational burden to some extent, the trade-off between tight robustness certificates and scalability remains a persistent challenge. For instance, while symbolic interval propagation is computationally efficient, it occasionally yields loose bounds, thereby reducing its certifiability against stronger adversarial attacks.

The integration of certified robustness into training pipelines, often termed certified training, represents an active area of research. This approach couples robustness certification with loss functions optimized during training, thereby directly enhancing the resilience of models. Techniques like randomized smoothing, which leverages Gaussian noise to create provably robust predictions, have garnered attention due to their simplicity and applicability across various architectures [15]. Despite its practical success, this method assumes smooth decision boundaries, which can be problematic for highly non-linear neural networks.

Nevertheless, critical limitations and open challenges persist. First, scalability to large-scale models and high-dimensional data, as seen in computer vision systems, remains a bottleneck. Optimizations that reduce computational overhead while maintaining tight certification bounds are imperative [21]. Furthermore, existing methods tend to underperform in multi-modal or dynamic environments, such as those involving autonomous systems or vision-language models, where robust certification across modalities poses unique challenges [66]. Extending certifiable robustness frameworks to handle these scenarios is a pressing need for real-world applicability.

Recent works point towards promising directions. Advances in neural architecture design, such as integrating certifiability constraints during model development, and the use of neural symbolic methods offer potential solutions to address non-linearities in network behavior. Additionally, hybrid approaches that blend robust optimization with for-

mal verification could bridge the gap between theoretical guarantees and efficiency.

In conclusion, certifiable robustness and formal verification provide an essential safeguard in adversarial machine learning, offering trustworthiness through mathematical rigor. However, achieving widespread adoption will require addressing challenges in scalability, multi-modal contexts, and generalization to new adversarial paradigms. Progress in this domain promises to not only fortify AI systems against adversarial threats but also advance the fundamental understanding of robustness in machine learning.

4 Applications and Domain-Specific Dynamics

4.1 Adversarial Threats and Defenses in Computer Vision Applications

Adversarial threats present significant challenges to computer vision applications by exploiting the vulnerabilities of deep neural networks (DNNs), such as their sensitivity to input perturbations and reliance on fragile high-dimensional representations. These vulnerabilities span across core tasks like image classification, object detection, facial recognition, and medical imaging, impeding their reliability in critical real-world scenarios. A prominent example is the susceptibility of image classification models to adversarial perturbations, as high-dimensional input spaces allow for imperceptible modifications that radically alter predictions [3], [20]. Gradient-based attacks such as FGSM, PGD, and multi-step optimization methods have demonstrated consistent efficacy under white-box settings, while black-box attacks leverage transferability to exploit vision systems without direct model access [2], [6].

Object detection systems, often used in autonomous vehicles and surveillance, face increasingly complex adversarial scenarios. Attackers can manipulate bounding boxes or occlude key features with crafted perturbations, severely degrading performance under real-world conditions. Physical attacks, such as altering stickers or patterns on stop signs, have highlighted the practical risks to object detectors in safety-critical systems [1], [22]. Robust defenses have emerged, such as adversarial training with augmented datasets and transformation-invariant feature extraction techniques, although a trade-off between clean accuracy and robustness persists [6], [23].

Facial recognition models used in authentication and surveillance systems are uniquely vulnerable to identity manipulation. Techniques like adversarial face swaps and patching have demonstrated success in creating imperceptible yet highly effective adversarial samples. Countermeasures, including multimodal fusion frameworks and ensemble methods, have provided incremental improvements in robustness but remain susceptible to adaptive, unseen attack strategies [21], [64].

Medical imaging introduces additional challenges due to its high-stakes implications. Adversaries targeting diagnostic systems, particularly using perturbations on X-rays or CT scans, jeopardize patient safety by leading to misdiagnoses or overlooked conditions. Defense mechanisms, such as adversarial denoising autoencoders and certifiably robust training, attempt to secure these models. However, the scarcity of labeled adversarial medical datasets limits systematic evaluations [18], [54].

Despite significant advancements, challenges persist in balancing scalability, generalization, and defense efficacy. Recent research emphasizes the integration of certifiable robustness methods, adversarial benchmarking, and domain-specific data augmentations to advance defenses further [9], [18]. Looking forward, addressing the inherent trade-offs between interpretability, computational efficiency, and robustness across diverse adversarial scenarios will remain critical for the safe deployment of computer vision systems in real-world applications.

4.2 Adversarial Dynamics in Natural Language Processing Systems

Adversarial attacks on Natural Language Processing (NLP) systems present unique challenges due to the discrete and linguistic nature of textual data, which imposes constraints on the types of perturbations that can be applied. Unlike continuous domains such as images, text modifications must not only remain imperceptible but also preserve grammatical structure and semantic coherence, ensuring the text appears natural to human readers. This subsection examines adversarial attacks targeting key NLP tasks, such as text classification, sentiment analysis, and large language models (LLMs), while highlighting defense mechanisms tailored to address the distinctive nature of these attacks.

Adversarial perturbations in NLP typically fall into categories such as character-level manipulations (e.g., typos, substitutions) and word- or sentence-level transformations that aim to minimally alter semantics while misleading models. Papernot et al. first demonstrated the vulnerability of text classifiers to strategically crafted attacks using localized edits like synonym substitutions or word deletions, which preserve the overarching intent of the text while significantly shifting predictions [57]. More recently, nuanced strategies have leveraged semantic understanding via embedding-based metrics to craft perturbations that effectively target advanced NLP models [57]. However, the contextual and syntactic dependencies inherent in state-of-the-art models, such as BERT or GPT-3, increase the complexity of generating adversarial examples that remain both valid and effective.

While adversarial methods adapted to NLP demonstrate considerable success, defenses tailored to this unique domain have proven challenging, given the necessity of balancing adversarial robustness with linguistic integrity. Techniques such as adversarial training have been explored, where models are exposed to adversarially perturbed samples during training to improve their resistance. Additionally, input preprocessing methods, such as adversarial word-level filtering or embedding denoising techniques, have been proposed to mitigate attacks before they impact downstream predictions. These approaches, while promising, often face trade-offs in computational efficiency or generalization to unseen attack types, underscoring the need for more scalable, domain-specific solutions.

As NLP systems increasingly power critical applications such as conversational agents, automated content moderation, and large-scale document analysis, adversarial vulnerabilities introduce far-reaching implications. Malicious

perturbations targeting these systems could influence policymaking, disrupt communication platforms, or lead to misinformation propagation, amplifying their societal impact. Therefore, advancing the robustness of NLP systems will require continuous refinement of both attack understanding and the development of novel defenses capable of addressing the evolving threat landscape. The demand for adversarial benchmarking tailored to diverse NLP tasks and the integration of threat-resistant design principles early in model development aligns with broader trends in adversarial machine learning research. Moving forward, fostering resilience in NLP systems while preserving linguistic fidelity will be critical for their reliable deployment across real-world applications.

4.3 Autonomous Systems and Cyber-Physical Applications

The integration of machine learning into autonomous systems and cyber-physical applications has introduced transformative capabilities but also exposed these systems to unique adversarial vulnerabilities. Safety-critical systems, such as autonomous vehicles (AVs), robotics, and industrial control systems (ICS), operate in environments where adversarial attacks can have severe consequences, necessitating robust strategies for mitigating these risks. This subsection explores adversarial threats in these contexts, emphasizing physical-world challenges, system-level robustness requirements, and emerging research directions.

Autonomous vehicles rely heavily on perception modules powered by machine learning, such as vision and LiDAR-based object detection. Adversarial attacks targeting these modules can result in dangerous misclassifications or object occlusions, leading to system failures. For instance, perturbations designed to manipulate traffic signs can cause AVs to misread a stop sign as speed limit signage, endangering lives [23], [33]. Attacks against LiDAR sensors, such as injecting spoofed point clouds, disrupt depth perception and obstacle detection, further complicating autonomous navigation [67]. Adaptive physical attacks exploiting environmental conditions (like lighting changes) amplify the challenge, as adversarial perturbations need to remain effective under varying physical constraints [23].

While adversarial robustness in trajectory prediction has seen progress through adversarial training and ensemble defenses, these solutions come with trade-offs. For example, adversarial training often increases robustness at the cost of clean performance and higher computational demands [20]. Moreover, trajectory perturbations targeting motion planning models can lead to unsafe path selection, requiring the integration of robust optimization methodologies to stabilize predictions [32].

Emerging cybersecurity threats in cyber-physical systems, such as ICS used in energy grids and manufacturing, present further concerns. Poisoning attacks on these systems compromise data integrity, enabling adversaries to manipulate predictions for critical applications like demand forecasting or anomaly detection in industrial workflows [68]. Defenses leveraging robust statistics, such as model prefiltering based on outlier detection, have shown promise, though these remain limited by scalability and dynamic adversarial strategies [69].

Physical constraints and real-time decision-making exacerbate adversarial challenges in these systems. For example, adversaries targeting time-sensitive applications like drone navigation or robotic surgery exploit spatio-temporal dynamics for adversarial attacks that are imperceptible in the physical domain but catastrophic in execution [50]. Defending against these threats requires multimodal fusion strategies, combining sensor data with redundancy and leveraging robust model architectures that can generalize across attack surfaces.

Future research should address the intersection of adversarial robustness and explainability in autonomous systems. Ensuring trust in these systems will necessitate interpretable defenses that provide insights into adversarial vulnerabilities while maintaining operational safety. By investigating formal verification techniques and certifiable robustness frameworks, researchers can establish theoretical guarantees for models deployed in safety-critical scenarios. Additionally, integrating adversarial learning into design pipelines could enable proactive identification of vulnerabilities during the development phase, fostering long-term system resilience.

In conclusion, while adversarial machine learning has exposed critical vulnerabilities in autonomous and cyber-physical systems, these challenges present opportunities for innovation. Enhanced robustness through adaptive defenses, explainability, and cross-modal resilience promises to redefine the security standards of these transformative technologies.

4.4 Emerging Threats in Multimodal and Cross-domain Systems

Multimodal and cross-domain systems, which integrate diverse modalities (e.g., vision and language) or adapt across domains (e.g., medical imaging to natural scenes), present new adversarial challenges, amplifying the complexity of ensuring robustness in adversarial machine learning. These systems, by nature, rely on correlated feature spaces and inter-modality dependencies, which inherently expand the landscape of potential attack vectors. Moreover, fusion mechanisms that align features across modalities or domains amplify vulnerabilities, exposing novel attack surfaces that adversaries can exploit.

A growing concern in this area is the transferability of adversarial examples across modalities and domains. For instance, adversarial perturbations originally crafted in one modality, such as images, can transfer to another, such as textual data, by leveraging shared or aligned intermediate representations, as observed in vision-language transformers [15]. This threat is especially pronounced in systems utilizing joint embedding spaces, where perturbations designed in the vision space disrupt textual predictions and vice versa. Vision-language models such as CLIP demonstrate this vulnerability, showing that adversarial examples can selectively manipulate object descriptions or captions, further underscoring the difficulty of defending alignment-based architectures [16].

Beyond transferability, fusion mechanisms intrinsic to multimodal systems exacerbate adversarial vulnerabilities. For example, models designed for vision-language tasks,

such as visual question answering or caption generation, heavily depend on multimodal attention mechanisms to aggregate and fuse information across modalities. Attacks targeting a single modality often propagate errors throughout the fusion pipeline, leading to cascading failures in downstream tasks. Notably, recent research highlights the potency of joint perturbation frameworks, which simultaneously create adversarial perturbations across modalities to exploit multimodal dependencies, outperforming single-modality attacks [15].

Cross-domain adaptation, a core component of domain-agnostic learning, also faces significant adversarial challenges. Domain-adversarial training, commonly used for cross-domain tasks, remains vulnerable to attacks that exploit domain-specific weaknesses. For instance, adversarial examples targeting urban traffic datasets can effectively transfer to models fine-tuned for rural driving scenes, revealing critical insufficiencies in existing domain adaptation methods [64]. These findings emphasize the fragility of domain-adaptive systems when confronted with carefully engineered domain-specific perturbations.

Evaluating robustness in multimodal and cross-domain systems also remains a formidable challenge. Traditional adversarial norms, such as L_p perturbations, often fail to capture the nuanced complexities of attacks in these settings, necessitating the development of innovative evaluation metrics and frameworks [70]. Promising directions for future research include designing adversarial certifications tailored to cross-modal and cross-domain systems, as well as developing adaptive, modality-focused defenses. Hybrid architectures that integrate adversarial training across modalities, combined with disentangled, modality-specific representations, hold particular promise for addressing the inherent complexities of these systems.

As multimodal and cross-domain systems continue to proliferate across diverse applications, the urgency to address these adversarial vulnerabilities intensifies. Ensuring their reliability in real-world deployments will require not only tackling foundational challenges but also advancing adversarially robust fusion mechanisms and cross-modal resilience strategies. This would align closely with the broader progress in adversarial machine learning for safety-critical and specialized domains, driving the next wave of innovation in secure, adaptable systems.

4.5 Specialized Domains and Application-Specific Impacts

Adversarial machine learning (AML) poses unique challenges within specialized domains like fraud detection, network security, and military or defense applications where failure often has substantial societal or economic consequences. Unlike generalized application areas, these domains exhibit highly contextualized vulnerabilities, distinct threat models, and intricate system dynamics that demand tailored adversarial strategies and countermeasures.

In fraud detection, adversarial attacks often exploit feature sparsity and the inherent imbalance of fraudulent versus legitimate data in training sets. Sophisticated attackers generate data modifications that leverage slight anomalies undetectable to standard fraud detection models, causing systems to classify fraudulent transactions as legitimate. Techniques such as adversarial perturbations in transactional sequences have proven particularly effective due to the subtlety of deviations required. Counterstrategies in this domain increasingly leverage robust feature engineering, such as graph-based feature representations, in combination with adversarial training [15]. A notable trade-off is the tension between model interpretability—key for regulatory compliance—and robustness against adversarial examples. Future advancements should aim to balance these priors with transformer-based architectures or hybrid anomaly detection methods [39].

Network security presents another focal domain for AML. Systems like intrusion detection models are susceptible to targeted evasion and poisoning attacks. For instance, adversarially-crafted network packets have been demonstrated to bypass machine learning-based intrusion detection systems (IDS) by subtly modifying non-critical fields without degrading their functional plausibility [71]. Furthermore, the evolution of targeted adversarial approaches, such as "realistic botnet" traffic generation using algorithms like Adv-Bot, emphasizes the increasing sophistication of these attacks in obscuring detection [31]. Defensive strategies predominantly revolve around adaptive perturbation models (e.g., A2PM), fortified training with adversarial examples, or leveraging fine-grained real-time behavioral analysis to address evasion [50]. However, scaling such defenses to large, heterogeneous network pipelines remains an unresolved challenge.

In high-stakes defense systems, including autonomous surveillance and military applications, attacks focus on physical-world adversarial examples to impair real-time decision-making. For example, adversarial camouflage such as pattern overlays or adversarial light projections has been shown to effectively mislead autonomous drones or object recognition systems in both white-box and black-box settings [28], [72]. Similarly, LiDAR-based systems for autonomous vehicles in defense undergo spoofing attacks that misrepresent object distances, increasing risks of system failure [73]. Critical advancements here involve integrating multimodal sensor feedback and adopting self-healing adversarial resistance frameworks, which dynamically reject anomalous sensor signals. Nonetheless, ensuring robust operation under adaptive real-world perturbations remains a priority; techniques like adversarial-aware reinforcement learning could provide promising avenues for future research [74].

Despite domain-specific progress, open issues remain pervasive across specialized applications. Many defenses, while successful in isolated contexts, falter under unforeseen or cross-domain transferability scenarios—evidenced by attacks like cross-modal perturbations on multimodal systems [52]. Advancing robust adversarial mitigation will require collaborative efforts to standardize threat models, define comprehensive benchmarks, and develop domain-aware evaluation metrics to capture the complex, context-specific requirements of these high-impact fields.

5 Evaluation Mechanisms and Benchmarks

5.1 Metrics for Adversarial Robustness Evaluation

Evaluating adversarial robustness is critical for understanding a machine learning model's resilience against adversarial perturbations and ensuring its reliability in real-world deployments. This subsection delves into the core metrics used to quantify adversarial robustness, analyzing their technical foundations, practical implications, and emerging innovations.

The most fundamental metric in this context is **Robust Accuracy**, which measures the proportion of adversarially perturbed inputs for which the model maintains correct predictions, effectively representing the trade-off between robustness and clean accuracy. While clean accuracy assesses performance on natural, unperturbed inputs, robust accuracy provides insight into defensive strength under attacks. However, achieving high robust accuracy often results in diminished clean accuracy, a trade-off extensively discussed in [2], [8]. Balancing these metrics remains an active area of research.

Adversarial perturbations are typically evaluated using **perturbation norms** such as L_0, L_2, and L_∞. These norms quantify the magnitude of the adversarial perturbation, constraining it to remain imperceptible to humans. Specifically, L_0 measures sparsity (number of perturbed features), L_2 measures Euclidean distance, and L_∞ captures the maximum change in any single input feature. While L_∞ has been widely adopted due to its simplicity, real-world scenarios often necessitate norms that align with perceptual similarity, an issue discussed in [1], [4]. These metrics are central in defining threat models and designing adversarial training procedures.

Another crucial development is **Certified Robustness**, which provides theoretical guarantees of a model's performance under bounded adversarial perturbations. This is often quantified by the **robustness radius**, which represents the maximum perturbation size ϵ that ensures correct classification for all inputs within this boundary. Methods relying on certified robustness, such as convex relaxations, have demonstrated strong theoretical grounding but limited scalability to high-dimensional tasks [9], [75].

Emerging advancements include **adversarial hypervolume** and **robustness curves**, which visualize model robustness across varying levels of perturbation intensities and attack types [18], [76]. These novel metrics capture more nuanced behaviors, moving beyond binary success/failure evaluation to assess the model's ability to withstand uncertainty across threat scenarios.

Despite these advancements, challenges persist. Many existing metrics assume static, well-defined threat models and fail to generalize to unforeseen or multi-modal attacks [12], [76]. Moreover, the computational cost of robustness evaluation, especially in large-scale models, remains prohibitive, underscoring the need for lightweight yet rigorous testing frameworks. As the field progresses, integrating perceptual metrics, incorporating domain-specific evaluation, and aligning robustness evaluation with application-level requirements will be necessary to holistically assess robustness.

In conclusion, while adversarial robustness metrics have evolved significantly, their limitations necessitate greater alignment with real-world constraints and adaptive threat models. Future research efforts should focus on standardized multi-faceted evaluation methodologies that extend beyond current norms, bridging theoretical guarantees and practical applicability.

5.2 Evaluation Scenarios and Standardized Datasets

The evaluation of adversarial robustness relies heavily on carefully constructed scenarios and standardized datasets, enabling fair and consistent benchmarking across a broad spectrum of attack and defense methodologies. These elements play a pivotal role in advancing the field by fostering comparative analyses and unveiling nuanced adversarial dynamics. Given the escalating pace of adversarial machine learning research, the demand for unified, reproducible evaluation standards is more urgent than ever.

Traditional datasets such as MNIST, CIFAR-10, and ImageNet have become mainstays in robustness evaluations, with their adversarially adapted versions (e.g., adversarial CIFAR-10 examples) laying the groundwork for extensive experimental studies [17], [57]. These datasets have ensured consistency in experimental setups while simultaneously exposing limitations, including their relatively straightforward nature compared to the challenges posed by real-world applications. Recently, alternative benchmarks like ImageNet-Patch and domain-specific adversarial datasets (e.g., for medical imaging or autonomous driving simulations) have emerged to bridge this gap, offering more contextually relevant scenarios that align with practical constraints [22], [77]. By simulating conditions such as physical disturbances or adversarial object manipulations, these datasets facilitate evaluations closely mirroring real-world adversarial challenges.

Additionally, curated benchmarks such as MultiRobustBench have introduced comprehensive multi-attack evaluation pipelines, enabling consistent assessments across white-box, black-box, and hybrid models while uncovering algorithmic trade-offs [23]. These efforts are bolstered by datasets tailored for evaluating cross-domain adversarial transferability, which reflects scenarios where adversarial perturbations are designed to transfer across varying tasks, modalities, or architectures. For instance, task-generalized datasets support insights into universality-based attacks like universal adversarial perturbations [13], [78], providing broader perspectives on attack generalization.

Despite these advancements, significant challenges remain. Real-world adversarial conditions often involve dynamic factors such as variable environmental lighting, occlusions, or sensor inconsistencies—dimensions that are insufficiently represented in synthetic datasets or controlled experimental setups [72], [79]. Current evaluation protocols also tend to predominantly focus on static data like images or text, overlooking domains such as video processing, multimodal systems, or sequential datasets [56], [77]. Furthermore, ethical considerations and safety-critical constraints are frequently underrepresented, yet they are becoming increasingly relevant in fields like autonomous driving and intrusion detection [22], [80].

Looking ahead, the development of datasets and benchmarks must emphasize adaptability to evolving adversarial tactics and application-specific demands. Benchmarks should integrate multi-dimensional robustness metrics, encompassing perturbation norms (e.g., L_2, L_∞), transferability of adversarial examples, and operational constraints such as computational efficiency and practicality in real-world deployments [81], [82]. By embedding these considerations within evaluation frameworks, the field can achieve greater rigor, foster innovation, and ready itself for the robust deployment of machine learning systems in complex, real-world adversarial landscapes.

5.3 Frameworks and Tools for Evaluation

Modern adversarial machine learning research necessitates robust, standardized frameworks and tools to evaluate the effectiveness of both attack strategies and defense mechanisms. These tools play a critical role by facilitating reproducibility, enabling rigorous quantitative assessments, and fostering insight into model vulnerabilities under various adversarial settings. This subsection examines the state-of-the-art evaluation frameworks and tools available for studying adversarial robustness, discussing their capabilities, limitations, and broader implications for benchmarking adversarial machine learning.

Established open-source libraries like CleverHans, Foolbox, and the Adversarial Robustness Toolbox (ART) have emerged as essential resources for generating, analyzing, and testing adversarial examples across diverse machine learning models and threat scenarios. CleverHans enables systematic evaluations, supporting an array of evasion attacks such as FGSM, PGD, and C&W, while providing baseline implementations for defenses like adversarial training [20]. Similarly, Foolbox expands on adversarial benchmarking by offering flexibility in crafting white-box and black-box attacks, leveraging query-based methods and surrogate models [33]. ART further enriches this ecosystem with its focus on robustness verification under multi-domain contexts, including images, NLP, and graph-based systems [15].

Despite their ubiquity, these frameworks face several challenges. For instance, while CleverHans and Foolbox excel at generating adversarial perturbations in static scenarios, they often struggle with adaptive attacks that dynamically tailor perturbations to circumvent specific defenses [30]. Furthermore, the reliance on norm-based metrics (e.g., L_2 or L_∞ perturbation bounds) limits their applicability to real-world constraints like physical attacks or temporal adversarial patterns [83]. Addressing this limitation, specialized frameworks like DeepRobust focus on domain-specific challenges, such as graph poisoning and transferability analysis in multi-task systems [35].

Emerging tools like RobTest and MultiRobustBench have gained traction for their scalability and adaptability to increasingly complex threat models. These systems simulate adaptive attack scenarios and evaluate robustness using ensemble-based metrics that capture nuanced model behaviors across diverse levels of adversarial knowledge [32]. Visualization tools, such as perturbation visualizers and latent-space feature graphs, have also become integral for interpreting adversarial dynamics at multiple levels, providing insights into the underlying vulnerabilities that spatial and semantic perturbations exploit [20].

Despite these advances, certain gaps persist. Current frameworks often focus predominantly on single-modality models, with limited extensions to multimodal or cross-domain systems, such as vision-language tasks [15]. Moreover, evaluation methodologies frequently overestimate defense robustness due to improperly simulated adaptive adversaries or insufficient variation in attack scenarios [69]. Efforts to incorporate ethical constraints into evaluation tools are also scant, with limited consideration of broader societal implications under adversarial manipulation [19].

Future directions should prioritize unifying evaluation protocols across diverse data modalities, integrating domain-specific real-world constraints, and emphasizing adaptive attack simulations to better mimic adversaries in operational environments. As multi-modal systems and generative AI tools like Diffusion Models (DMs) and Large Language Models (LLMs) gain prominence, frameworks that assess robustness in increasingly hybrid contexts will be pivotal [48], [84]. Beyond technical robustness, fostering reproducible, open science through modular toolkits and shared benchmarks will remain a cornerstone for advancing the field of adversarial machine learning.

5.4 Methodological Best Practices and Guidelines

Ensuring rigorous and reproducible evaluation in adversarial machine learning is crucial for reliable research and the advancement of the field. This subsection explores methodological best practices to comprehensively assess adversarial attacks and defenses while mitigating common pitfalls, building on the tools and frameworks discussed earlier and paving the way for addressing key challenges highlighted subsequently.

A cornerstone of effective evaluation is the explicit simulation of adaptive adversaries, tailored to account for evolving threat landscapes. Robust defenses must demonstrate resilience against dynamic attacks that exploit specific countermeasures, thereby addressing practical adversarial scenarios rather than overestimating robustness through static methodologies [12]. Inadequate incorporation of adaptive strategies has been a recurring limitation, often leading to exaggerated claims of efficacy [12]. Therefore, adaptive evaluations are essential to substantiate defense mechanisms against realistic adversarial threats.

Another imperative practice is conducting cross-evaluation of defenses across diverse threat models and adversarial modalities. Relying exclusively on a single attack type or perturbation norm, such as L_2 or L_∞, can obscure vulnerabilities to alternative or unforeseen strategies [76]. Evaluating defenses under multi-attack scenarios or across varying modalities—including poisoning, backdoor, and evasion attacks—offers a more holistic perspective on robustness. Efforts like the Adversarial Robustness Toolbox have facilitated this by standardizing multi-threat evaluation pipelines [53], promoting consistent and transparent benchmarking.

Reproducibility underscores the integrity of adversarial research. Publicly accessible datasets, shared codebases,

and rigorous documentation of hyperparameters, attack settings, and defense implementations are imperative to ensure that experimental results are verifiable. Benchmarking initiatives, such as those outlined in [70], highlight the value of consolidated evaluation protocols to enhance transparency and foster collaborative progress. Minor variations in experimental setups can yield inconsistent results, further underscoring the importance of rigorously documenting methodologies and providing sufficient details to enable independent verification [62].

Avoiding false positives in robustness claims constitutes another critical best practice. Misleading methodologies, such as gradient obfuscation or incomplete assessment of detection mechanisms, have historically led to inflated claims of defense effectiveness [12], [62]. Strategies to counteract such flaws include employing gradient-based single-step attacks, iterative adaptive approaches, and avoiding exploitation of specific evaluation artifacts [85]. Furthermore, stateful mechanisms, as explored in [86], should undergo extensive validation to ensure robustness against novel adaptive adversarial algorithms.

Finally, aligning evaluation methodologies with emerging and practical use cases is critical for ensuring real-world relevance. Domain-specific challenges, such as network intrusion detection [87], and the growing prominence of multi-modal evaluations reflect the need for frameworks that extend well beyond static datasets and conventional attack strategies. Incorporating evaluations that simulate dynamic adversarial scenarios and cross-modal contexts will contribute to a more accurate reflection of real-world adversarial environments.

By addressing these gaps through adaptive adversary testing, multi-modal robustness evaluations, and rigorous research transparency, the field of adversarial machine learning can establish a solid empirical foundation. This harmonized approach will bridge methodological practices with the challenges posed by advancing attack strategies, fostering innovations that are both theoretically sound and practically robust.

5.5 Challenges and Open Questions in Robustness Evaluation

Robustness evaluation in adversarial machine learning encounters multifaceted challenges that hinder the development of comprehensive and reliable assessment mechanisms. At its core lies the issue of dynamic threat modeling, wherein the evolving sophistication of attack strategies frequently outpaces the adaptability of current evaluation frameworks. Attackers often exploit unforeseen vulnerabilities in defenses deemed robust under specific scenarios, raising concerns about overfitting to known attacks [4], [81]. This dynamic necessitates robustness evaluation frameworks capable of anticipating and countering adaptive attacks, a direction largely underexplored.

Another key challenge involves assessing multi-modal and cross-domain robustness. With the rise of multimodal models integrating vision, language, and other modalities, evaluating robustness across such heterogeneous domains remains complex [28], [51]. For instance, as demonstrated in attacks on vision-language tasks [52], perturbations seamlessly bridging modalities often evade detection. Similarly, frameworks assessing domain transferability and generalization under these multi-modal settings are still in infancy, limiting the ability to benchmark universal robustness [88]. Standardization in such scenarios is further complicated by the varying evaluation criteria across tasks and models.

Computational costs also impose significant constraints, especially in exhaustive robustness evaluations. State-of-the-art methodologies, such as those accounting for distributional shifts or employing large-scale datasets like DAmageNet [89], demand considerable compute resources. The creation of perturbations under diverse attack strategies, coupled with iterative testing on complex architectures, exacerbates this burden [90]. These requirements make scalability a critical bottleneck, particularly for resource-constrained practitioners.

Practical implications also extend to the absence of unified robustness metrics. As highlighted through discrepancies in norm-based metrics like L_0, L_2, and L_∞ [39], such measures often fail to capture the nuanced trade-offs between adversarial and clean accuracy. Moreover, frameworks often overlook robustness hypervolumes, designed to quantify robustness under varying perturbation intensities and real-world constraints [29]. These limitations call for metrics integrating theoretical guarantees, empirical resilience, and practical feasibility.

Ethical and safety concerns further complicate robustness evaluation. Physical-world attacks, such as adversarial camouflage [28] or adversarial laser attacks [72], raise the stakes for real-world deployments. Robustness metrics often underrepresent adversarial risks in operational contexts, such as autonomous vehicles and multimodal systems [91]. Incorporating system-level semantics into robustness evaluation, as suggested by SysAdv [22], provides a promising direction, though the practical scalability of such techniques remains uncertain.

Looking forward, tackling robustness evaluation requires frameworks that synergize adaptive resilience, computational scalability, and cross-domain transferability. Collaborative benchmarks, such as DAmageNet [89], offer encouraging progress in standardizing multi-attack scenarios. Additionally, leveraging advancements in explainable machine learning could enhance diagnostic capabilities, enabling the detection of subtle weaknesses in defenses. Addressing these open questions will pave the way for robust, scalable, and ethical machine learning systems.

6 Future Directions, Open Problems, and Broader Perspectives

6.1 Ethical Implications, Dual-Use Concerns, and Regulations

Adversarial machine learning (AML) presents profound ethical challenges due to its dual-use nature, where techniques designed for beneficial purposes can be exploited maliciously. Methods enabling the testing and fortification of model robustness against adversarial examples are instrumental for enhancing safety-critical systems like autonomous vehicles and medical diagnostics; however, these same techniques can also enable attackers to craft sophisticated adversarial inputs for malicious activities, such as de-

ceptive cyberattacks or bypassing facial recognition systems [3], [39], [54].

The dual-use paradigm is particularly concerning for applications with severe societal consequences. For instance, adversarial examples targeting cybersecurity systems, including intrusion detection and malware detection, demonstrate potential for catastrophic consequences when leveraged by malicious adversaries to exploit critical infrastructure [49], [92]. Similarly, physical adversarial attacks in domains such as autonomous driving highlight ethical dilemmas regarding the intentional manipulation of Stop signs or traffic objects, risking human lives and public safety [1], [22].

Effective governance frameworks must address the ethical safeguards required to balance such dual-use risks. Ethical oversight could come in the form of international standards that enforce transparency in research dissemination and regulate the development of adversarial technologies. For example, principles around responsible disclosure and ethical publishing practices may focus on delineating defensive contributions from offensive exploitation [19], [93]. Furthermore, organizations deploying machine learning solutions must establish robust validation pipelines combining ethical testing protocols with legal mechanisms to enforce accountability for misuse [9].

More broadly, the intersection of adversarial research with legal frameworks raises critical concerns regarding intellectual property, civil liberties, and surveillance tools. Data poisoning or adaptive adversarial attacks on privacy-preserving models could compromise user identities in sensitive applications, further embedding biases into systems or exacerbating inequalities under biased regulatory environments. Such concerns are aggravated in jurisdictions prone to authoritarian misuse of technology [47], [94].

An emerging trend is the role of regulatory frameworks to proactively address these misuse risks. The development of certifiably robust systems and the incorporation of audit mechanisms in adversarial pipelines offer promising directions for minimizing unintended consequences. Collaborative efforts between academia, industry, and policymakers can ensure that adversarial learning technologies evolve responsibly [62], [93]. These collaborations must also encompass adaptive ethical considerations for new paradigms such as multimodal and federated learning, where decentralized architectures present novel attack surfaces and governance challenges [16], [92].

Ultimately, the AML community faces a precarious balancing act between fostering innovation and mitigating abuse. Solutions demand embedding ethical safeguards at all levels, from the creation of shared research benchmarks under constrained threat models to the development of explainable defensive strategies that promote stakeholder trust. A unified global strategy—comprising responsible innovation, enforceable legal standards, and comprehensive ethical guidelines—remains imperative to ensure that adversarial advancements serve societal benefits far beyond their potential harms [11], [18].

6.2 Scalable and Generalizable Robustness

Scalable and generalizable robustness is a pivotal focus in advancing adversarial machine learning, emphasizing the need for defense mechanisms that balance efficiency with adaptability across diverse tasks, data distributions, and architectures. The fundamental challenge lies in translating theoretical robustness guarantees into practical solutions for real-world systems, where models must contend with dynamic, heterogeneous, and computationally intensive environments.

Adversarial training, a widely adopted framework for enhancing robustness, has demonstrated empirical success but remains constrained by its computational demands, particularly for large-scale datasets and high-dimensional models. Strategies aiming to mitigate these inefficiencies, such as gradient projection techniques [57], highlight promising developments but tend to be domain-specific, limiting their applicability in diverse contexts. Furthermore, adversarial training is often prone to overfitting to specific perturbation types, which restricts its effectiveness against novel attacks or unforeseen adversarial scenarios [76]. In versatile deployment environments, it is essential to advance adaptive defenses capable of addressing a variety of threat models.

Another critical frontier is cross-modal and transferable robustness, which examines whether defenses effective in one domain (e.g., vision models) can generalize seamlessly to another (e.g., text or multimodal systems). Universal perturbations crafted in one task often demonstrate significant transferability to others, yet existing defensive measures fail to maintain domain-agnostic efficacy [13]. Emerging techniques such as neural perceptual threat models have made strides in approximating perceptual distances, providing enhanced robustness against unforeseen attack modalities and achieving notable results across different scenarios [58]. However, establishing universal metrics that can quantify adversarial vulnerabilities consistently across modalities remains a formidable technical challenge.

In real-world applications, especially those with time-sensitive operations like autonomous systems or cybersecurity frameworks, efficiency at scale is an indispensable requirement for practical defenses. Sparse optimization frameworks such as Sparse-RS deliver significant advances by reducing computational overhead without compromising effectiveness [90]. Similarly, lightweight mechanisms like adversarial feature injection have shown promise by introducing perturbations in feature spaces during training, fostering scalable robustness in a computationally efficient manner [95].

Future advances in scalable and generalizable robustness must focus on harmonizing these diverse goals. This includes developing defenses adaptive to multi-attack contexts, establishing datasets that comprehensively capture real-world adversarial landscapes across domains [49], and employing robust optimization techniques to enhance multi-task resilience. Unified benchmarks and evaluation protocols remain critical to resolving inconsistencies in robustness assessments, ensuring scalable mechanisms that can keep pace with evolving adversarial landscapes.

6.3 Advancing Robustness in Decentralized and Federated Learning

Decentralized and federated learning frameworks have brought transformative potential in privacy-preserving machine learning by enabling collaborative modeling without

sharing raw data. However, these systems are inherently vulnerable to adversarial threats, including data and model poisoning, gradient manipulation, and collusion-based attacks, due to their distributed architectures and distinct privacy constraints. Robustness in this context requires tackling unique challenges posed by data heterogeneity, asynchronous updates, and the difficulty of detecting adversarial behavior across distributed participants.

A primary concern in federated learning is **poisoning attacks**, where adversaries modify local data or model updates to degrade the global model's performance or embed backdoors. For example, targeted clean-label poisoning attacks can compromise federated learning models with minimal perturbations [96], while backdoor attacks remain particularly challenging due to the diversity in data distributions and the lack of centralized oversight [97]. To counter these threats, defenses such as robust aggregation mechanisms (e.g., median or trimmed mean frameworks) have been proposed, which mitigate outliers introduced by malicious participants. However, these methods often assume aligned distributions across nodes, limiting their efficacy in real-world heterogeneous setups [94].

Another critical issue is **Byzantine adversaries**, who act maliciously to manipulate contributions in federated updates. Byzantine-resilient learning techniques like norm-based clipping or coordinate-wise median aggregation play a pivotal role in ensuring stability under such threats [24]. However, these approaches face trade-offs between robustness and performance, as overly defensive strategies can inadvertently discard valuable information from benign nodes.

Decentralized paradigms introduce further complexity by removing reliance on a central server, amplifying the challenge of cross-node communication and collusion detection. For example, in peer-to-peer systems, adversarial nodes can perform **gradient inversion attacks**, exposing private data while operating within the legitimate optimization pipeline [98]. Differentially Private mechanisms and adaptive noise injection have shown potential to limit these risks but struggle with maintaining model utility, especially in asynchronous updates [68].

A promising emerging direction is coupling adversarial robustness with **heterogeneity-aware defenses**, leveraging consensus-inspired protocols for node-specific reliability. Techniques such as hierarchical federated aggregation, where participant groups are dynamically pruned based on trustworthiness evaluations, show promise in mitigating both poisoning and privacy violations [99].

Moving forward, robust federated learning systems require principled approaches to align adversarial and privacy protections. For instance, unifying methods from privacy-enhanced machine learning, such as secure multi-party computation, with adaptive adversarial training frameworks offers a rich avenue for exploration [30]. Moreover, designing practical benchmarks and metrics that consistently evaluate robustness under diverse adversarial threat models is an open problem with strong implications for real-world deployment [83].

In synthesis, ensuring robustness in decentralized and federated learning demands solutions that account for the interplay of adversarial dynamics, heterogeneity, and structural constraints. Advancing research in this domain requires a balance between theoretical innovation and practical deployment strategies, with interdisciplinary collaborations crucial to addressing the escalating challenges.

6.4 Interpretability and Explainability in Adversarial Defenses

Adversarial defenses in machine learning (ML) must balance robustness with interpretability, as the opaque nature of many state-of-the-art strategies undermines stakeholder trust in real-world applications. Integrating interpretability into adversarial defense mechanisms represents a dual technical and philosophical challenge, requiring defenses not only to withstand attacks but also to elucidate why certain perturbations succeed or fail. This subsection explores the advancements and hurdles in aligning adversarial robustness with explainability, critically evaluating existing methodologies while identifying avenues for innovation.

Interpretability-driven approaches to adversarial robustness can be categorized into feature-level and model-level explainability strategies. Feature-level methods aim to illuminate how adversarial inputs compromise models by distorting high-dimensional features. For instance, saliency-map-based tools highlight regions that adversaries exploit, offering valuable insights for debugging models and refining preprocessing defenses [100]. Such tools facilitate isolating robust features while discarding vulnerable ones, as seen in methods like Defense-GAN [27]. However, these approaches struggle under adaptive attacks, where adversaries obscure interpretable patterns, and their sensitivity to noise limits their reliability in dynamic adversarial contexts.

Conversely, model-level interpretability seeks to unravel a model's decision-making process to ensure defense behaviors resonate with human intuition. Techniques like explainable regularization enforce simplicity in decision boundaries, reducing reliance on adversarially sensitive subspaces [100]. Another promising avenue involves using explainability metrics to identify components most susceptible to adversarial vulnerabilities—for instance, employing activation maximization methods to visualize how perturbations skew predictions [54]. However, dissecting complex architectures like vision transformers and large-scale language models remains computationally intensive, presenting barriers to scalability.

A significant trade-off exists between robustness and transparency. Robust models, such as those trained adversarially, often appear less interpretable, as their focus on ℓ_p-norm-constrained robustness tends to misalign with human-comprehensible features [62]. Meanwhile, interpretable-by-design architectures frequently underperform against transferable adversarial examples, where perturbations crafted for one model generalize effectively across others [34]. This tension underscores the need for approaches that harmonize these often-competing objectives.

Recent research emphasizes embedding interpretability into defense evaluation protocols. Diagnostic tools leveraging explainability can pinpoint the layers or modules most vulnerable to adversarial manipulation, thereby supporting targeted robustness improvements [100]. Similarly, explainability frameworks like InterpretML have demonstrated

potential in bridging user trust with adversarial defense strategies by elucidating interactions between perturbations and countermeasures across data modalities [100].

Future advancements must focus on creating universal interpretability protocols that adapt dynamically to evolving adversarial threats while ensuring computational efficiency. Hybrid techniques, such as explainability-aware adversarial training and inherently interpretable generative defenses like Defense-GAN, warrant further exploration [27]. Moreover, interdisciplinary collaboration encompassing ML, human-computer interaction, and cognitive neuroscience should be prioritized to unravel model vulnerabilities and enhance interpretability under adversarial conditions. Strengthening explainability in adversarial settings will not only bolster robustness but will also facilitate the ethical and transparent deployment of secure ML systems, acting as a critical bridge to future AML advancements.

6.5 Emerging Trends and Collaborative Research Directions

The dynamic field of adversarial machine learning (AML) is witnessing a transformation with emerging trends emphasizing broader scopes and synergies across disciplines. A crucial shift is the exploration of robust benchmarking and standardization efforts. While existing tools such as CleverHans and Foolbox have facilitated adversarial robustness evaluation [90], there remains a gap in establishing unified benchmarks that encompass real-world attack scenarios and multi-modal considerations. Recent works underline efforts towards comprehensive datasets, such as the creation of adversarial-specific benchmarks like DAmageNet, which evaluates adversarial robustness across multiple models and tasks using universal perturbations [89]. These efforts, although promising, highlight the need for broader standardization to ensure reproducibility and fair comparisons across methods.

Additionally, the growing integration of adversarial learning in large vision-language models (VLMs) and multimodal architectures introduces new vulnerabilities and challenges. Studies have demonstrated targeted attacks exploiting vision and language fusion, significantly compromising aligned systems like GPT-4V and MiniGPT-4 [101], [102]. With pre-trained VLMs increasingly adopted, the field must pivot towards collaborative frameworks involving vision, language, and cybersecurity to design robust defenses that safeguard against multi-modal attack dynamics. Collaborative initiatives, such as blending robust feature denoising with adversarial training pipelines for cross-modal systems, have shown partial success in defending against transferability-driven attacks [41].

Another prominent area is the intersection of AML with decentralized and federated learning systems. These paradigms face additional challenges due to their heterogeneity, privacy-preserving requirements, and distributed nature, which attackers exploit through poisoning or collusion to disrupt global models [103]. Techniques such as Byzantine-resilient aggregation and cross-node robustness evaluations are emerging as crucial steps towards securing decentralized environments. Still, a holistic approach integrating privacy-preserving mechanisms with scalable adversarial defenses remains underexplored.

Interdisciplinary collaboration is paramount for accelerating progress in AML. Researchers advocate for leveraging domain-specific expertise to craft innovative solutions, as evidenced by advancements in adversarial camouflage for physical-world attacks [28] and adversarial botnet traffic generation to evaluate practical cybersecurity risks [31]. Collaborative synergies across machine learning, cybersecurity, and ethics communities could foster a more comprehensive framework to tackle real-world adversarial scenarios.

Future work should prioritize long-term strategies capable of mitigating adaptive attacks. This requires designing defense mechanisms that remain effective against evolving adversarial strategies over extended time horizons [76]. Furthermore, addressing the inherent trade-offs between robustness, generalizability, and computational efficiency is critical, particularly for resource-constrained applications like autonomous driving and IoT systems [104]. By leveraging cross-disciplinary expertise and advancing evaluation standards, adversarial machine learning can evolve into a more resilient and impactful discipline.

6.6 Leveraging Adversarial Learning for Positive Applications

Adversarial methodologies, traditionally viewed as tools for exploiting machine learning models, are now transforming into instruments for promoting model robustness, fairness, and interpretability. As adversarial techniques mature, their potential to generate synthetic data, uncover model biases, and enhance real-world performance is driving innovation across diverse applications. This subsection delves into these promising use cases, highlighting how adversarial learning can transcend its origins as a threat vector to become a constructive force in advancing AI systems.

One of the most impactful applications of adversarial learning is in *adversarial data augmentation*, wherein adversarial examples are strategically integrated into training pipelines to amplify data diversity. This approach increases robustness and enhances generalization, particularly in data-limited scenarios. Defense-GAN [27] exemplifies this paradigm by deploying generative adversarial models to produce adversarial perturbations that bolster defensive training, thereby fortifying the model against adaptive attacks. Moving beyond traditional tactics, frameworks such as LAS-AT [105] adopt a dynamic perspective by incorporating learnable adversarial attack strategies that fine-tune perturbation parameters, simultaneously enriching datasets and bolstering resilience. These advancements underscore adversarial learning's potential to transform vulnerability into resilience.

In addition to strengthening models, adversarial learning is unveiling its utility in *bias detection and fairness auditing*, where it acts as a diagnostic tool to identify hidden vulnerabilities. By targeting specific data subpopulations, adversarial attacks can spotlight patterns of errors that reveal biases and inequities. For example, adversarial perturbation pipelines have been harnessed to expose fairness violations by identifying vulnerabilities tied to protected attributes, ensuring a transparent evaluation of model behaviors [15]. Tools like ANTHRO [106] extend these efforts by leveraging human-crafted variations to create realistic adversarial

conditions, supporting robust fairness assessments. Despite the potential of these methods to improve inclusivity and integrity, challenges remain in scaling them effectively, especially across complex and multimodal datasets.

Adversarial methodologies also hold promise in *privacy and security frameworks*, where their mechanisms are repurposed to strengthen defenses rather than exploit weaknesses. For example, adversarial perturbations can be intentionally deployed to protect sensitive information by confusing adversaries while maintaining legitimate functionality for users. This concept has been explored in generative systems like PixelDefend [107], which introduces adversarially generated noise to obfuscate exploitable patterns while preserving model performance. Similarly, adversarial techniques are being applied to counter misinformation and deepfake media manipulation, with methods designed to disrupt the generation of highly deceptive multimedia content [108]. These innovations add a new dimension to the utility of adversarial learning, expanding its role from defensive counter-measures to proactive enhancers of cybersecurity.

However, realizing the full potential of adversarial learning for positive applications presents persistent challenges. Computational overheads in generating adversarial data, trade-offs between clean sample accuracy and adversarial robustness, and the difficulty of generalizing these improvements across diverse architectures remain significant roadblocks. As advancements in large-scale vision-language pretraining expand the scope of adversarial learning [51], [89], new opportunities arise to resolve these issues, opening pathways to extend adversarial strategies into domains such as multimodal alignment and decentralized systems.

Looking ahead, there is immense potential in merging adversarial learning with interpretable machine learning and formal verification frameworks to expand its positive impact across socially critical applications. By mastering this dual-use paradigm—leveraging adversarial techniques for both mitigation and constructive innovation—the field stands poised to shape robust, ethical AI systems tailored for deployment across diverse real-world environments. This evolution of adversarial learning will redefine its role, bridging the gap between adversarial resilience and constructive utility to meet the growing demands of trustworthy artificial intelligence.

7 Conclusion

The field of adversarial machine learning (AML) has undergone rapid development, revealing both profound vulnerabilities and significant opportunities within modern artificial intelligence systems. This survey has systematically explored adversarial threats and corresponding defense mechanisms, highlighting how these evolving dynamics intersect with critical domains such as computer vision, natural language processing, autonomous systems, and cybersecurity. A recurring theme throughout this survey has been the duality of advancements in attack strategies and defensive measures, creating what some describe as an "arms race" in adversarial research [67], [109].

On the attack side, techniques have grown more sophisticated, ranging from classic evasion attacks that exploit gradient-based perturbations [3] to advanced modalities such as physical-world adversarial examples [1] and targeted attacks in multimodal systems [15]. The universality and transferability of adversarial examples remain major areas of concern, as attacks crafted for one model can often fool multiple others, even across domains [7]. These challenges amplify the importance of developing robust and generalizable defenses that transcend isolated threat models.

Defense mechanisms have displayed parallel progress, with adversarial training emerging as one of the most effective strategies for improving model robustness [6], [8]. Other methods, such as input preprocessing, gradient regularization, and formal verification, provide complementary avenues for increasing resilience [9], [53]. However, this survey also notes persistent limitations, including the computational expense of adversarial training and the vulnerability of many defenses to adaptive attacks [12]. These findings underscore the importance of moving toward holistic and scalable defense strategies capable of addressing multi-modal, task-specific, and unforeseen attack scenarios [76], [87].

A critical takeaway from this body of literature is the necessity of standardized benchmarking and rigorous evaluation frameworks to measure robustness under realistic and adversarial conditions [18], [62]. Existing evaluation tools such as the Adversarial Robustness Toolbox [53] and multi-attack benchmarks [70] must evolve further to incorporate increasingly complex threat models while ensuring reproducibility. Moreover, formal discussions of practical applicability, such as in cybersecurity and autonomous systems, reinforce the real-world urgency of addressing adversarial vulnerabilities [22], [67].

Looking ahead, future research must emphasize several key areas. First, there is a need to systematically bridge adversarial robustness with explainability and interpretability [110]. Ensuring that stakeholders can diagnose and trust robust systems will be critical as these technologies integrate into high-stakes environments. Second, addressing adversarial robustness in decentralized and federated learning contexts remains a largely open challenge, particularly given the proliferation of distributed AI systems [8]. Finally, multi-disciplinary collaboration across machine learning, cybersecurity, ethics, and systems engineering is essential to foster a more comprehensive understanding of adversarial dynamics [93].

In conclusion, while substantial strides have been made in understanding and mitigating adversarial threats, the dynamic and evolving nature of attack strategies demands continual innovation. Embracing unified standards, interdisciplinary collaboration, and application-focused approaches will ensure that research in adversarial machine learning not only stays ahead of malicious actors but also advances the broader goal of developing secure, fair, and trustworthy AI systems.

References

[1] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *ArXiv*, vol. abs/1607.02533, 2016. 1, 2, 4, 8, 11, 14, 17

[2] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, pp. 2805–2824, 2017. 1, 2, 5, 8, 11

[3] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *CoRR*, vol. abs/1412.6572, 2014. 1, 2, 5, 8, 14, 17

[4] A. Fawzi, H. Fawzi, and O. Fawzi, "Adversarial vulnerability for any classifier," *ArXiv*, vol. abs/1802.08686, 2018. 1, 7, 11, 13

[5] A. Shafahi, W. R. Huang, C. Studer, S. Feizi, and T. Goldstein, "Are adversarial examples inevitable?" *ArXiv*, vol. abs/1809.02104, 2018. 1, 2

[6] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *ArXiv*, vol. abs/1611.01236, 2016. 1, 5, 8, 17

[7] J. Gu, X. Jia, P. de Jorge, W. Yu, X. Liu, A. Ma, Y. Xun, A. Hu, A. Khakzar, Z. Li, X. Cao, and P. Torr, "A survey on transferability of adversarial examples across deep neural networks," *ArXiv*, vol. abs/2310.17626, 2023. 1, 2, 17

[8] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent advances in adversarial training for adversarial robustness," in *International Joint Conference on Artificial Intelligence*, 2021, pp. 4312–4321. 1, 5, 11, 17

[9] M. H. Meng, G. Bai, S. Teo, Z. Hou, Y. Xiao, Y. Lin, and J. Dong, "Adversarial robustness of deep neural networks: A survey from a formal verification perspective," *ArXiv*, vol. abs/2206.12227, 2022. 1, 5, 8, 11, 14, 17

[10] Z. Yan, Y. Guo, and C. Zhang, "Deep defense: Training dnns with improved adversarial robustness," in *Neural Information Processing Systems*, 2018, pp. 417–426. 1, 2, 5

[11] B. Wu, S. Wei, M. Zhu, M. Zheng, Z. Zhu, M. Zhang, H. Chen, D. Yuan, L. Liu, and Q. Liu, "Defenses in adversarial machine learning: A survey," *ArXiv*, vol. abs/2312.08890, 2023. 1, 6, 14

[12] F. Tramèr, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," *ArXiv*, vol. abs/2002.08347, 2020. 1, 5, 7, 11, 12, 13, 17

[13] C. Zhang, P. Benz, C. Lin, A. Karjauv, J. Wu, and I. Kweon, "A survey on universal adversarial attack," in *International Joint Conference on Artificial Intelligence*, 2021, pp. 4687–4694. 1, 3, 11, 14

[14] W. Jin, Y. Li, H. Xu, Y. Wang, and J. Tang, "Adversarial attacks and defenses on graphs: A review and empirical study," *ArXiv*, vol. abs/2003.00653, 2020. 1, 5

[15] H. Xu, Y. Ma, H. Liu, D. Deb, H. Liu, J. Tang, and A. K. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," *International Journal of Automation and Computing*, vol. 17, pp. 151 – 178, 2019. 1, 4, 7, 9, 10, 12, 16, 17

[16] W. Zhang, Q. Sheng, A. Alhazmi, and C. Li, "Adversarial attacks on deep learning models in natural language processing: A survey," *arXiv: Computation and Language*, 2019. 1, 4, 9, 14

[17] A. Serban, E. Poll, and J. Visser, "Adversarial examples on object recognition," *ACM Computing Surveys (CSUR)*, vol. 53, pp. 1 – 38, 2020. 1, 5, 11

[18] Y. Dong, Q.-A. Fu, X. Yang, T. Pang, H. Su, Z. Xiao, and J. Zhu, "Benchmarking adversarial robustness on image classification," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 318–328, 2020. 1, 2, 8, 11, 14, 17

[19] J. Gilmer, R. P. Adams, I. Goodfellow, D. G. Andersen, and G. E. Dahl, "Motivating the rules of the game for adversarial example research," *ArXiv*, vol. abs/1807.06732, 2018. 1, 6, 12, 14

[20] R. Wiyatno, A. Xu, O. A. Dia, and A. D. Berker, "Adversarial examples in modern machine learning: A review," *ArXiv*, vol. abs/1911.05268, 2019. 2, 6, 8, 9, 12

[21] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14 410–14 430, 2018. 2, 4, 7, 8

[22] N. Wang, Y. Luo, T. Sato, K. Xu, and Q. A. Chen, "Does physical adversarial example really matter to autonomous driving? towards system-level effect of adversarial object evasion attack," *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4389–4400, 2023. 2, 4, 8, 11, 13, 14, 17

[23] N. Akhtar, A. Mian, N. Kardan, and M. Shah, "Advances in adversarial attacks and defenses in computer vision: A survey," *IEEE Access*, vol. PP, pp. 1–1, 2021. 2, 5, 8, 9, 11

[24] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," *2018 IEEE Symposium on Security and Privacy (SP)*, pp. 19–35, 2018. 2, 15

[25] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *International Conference on Machine Learning*, 2012, pp. 1467–1474. 2

[26] L. Muñoz-González, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E. C. Lupu, and F. Roli, "Towards poisoning of deep learning algorithms with back-gradient optimization," *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017. 2

[27] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-gan: Protecting classifiers against adversarial attacks using generative models," *ArXiv*, vol. abs/1805.06605, 2018. 2, 3, 5, 6, 7, 15, 16

[28] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yang, "Adversarial camouflage: Hiding physical-world attacks with natural styles," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 997–1005, 2020. 2, 10, 13, 16

[29] L. Nguyen, S. S. Arora, Y. Wu, and H. Yang, "Adversarial light projection attacks on face recognition systems: A feasibility study," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3548–3556, 2020. 2, 13

[30] A. Demontis, M. Melis, M. Pintor, M. Jagielski, B. Biggio, A. Oprea, C. Nita-Rotaru, and F. Roli, "Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks," in *USENIX Security Symposium*, 2018, pp. 321–338. 2, 12, 15

[31] I. Debicha, B. Cochez, T. Kenaza, T. Debatty, J. Dricot, and W. Mees, "Adv-bot: Realistic adversarial botnet attacks against network intrusion detection systems," *Comput. Secur.*, vol. 129, p. 103176, 2023. 3, 4, 10, 16

[32] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndic, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," *ArXiv*, vol. abs/1708.06131, 2013. 3, 9, 12

[33] N. Papernot, P. Mcdaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 2016. 3, 9, 12

[34] N. Papernot, P. Mcdaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *ArXiv*, vol. abs/1605.07277, 2016. 3, 15

[35] A. Bojchevski and S. Günnemann, "Adversarial attacks on node embeddings via graph poisoning," in *International Conference on Machine Learning*, 2018, pp. 695–704. 3, 12

[36] S. Goldwasser, M. P. Kim, V. Vaikuntanathan, and O. Zamir, "Planting undetectable backdoors in machine learning models : [extended abstract]," *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 931–942, 2022. 3

[37] A. Salem, R. Wen, M. Backes, S. Ma, and Y. Zhang, "Dynamic backdoor attacks against machine learning models," *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pp. 703–718, 2020. 3, 4

[38] L. Song, R. Shokri, and P. Mittal, "Privacy risks of securing machine learning models against adversarial examples," *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019. 3

[39] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial attacks and defences: A survey," *ArXiv*, vol. abs/1810.00069, 2018. 4, 7, 10, 13, 14

[40] T. Zheng, C. Chen, and K. Ren, "Distributionally adversarial attack," in *AAAI Conference on Artificial Intelligence*, 2018, pp. 2253–2260. 4

[41] J. Tu, H. Li, X. Yan, M. Ren, Y. Chen, M. Liang, E. Bitar, E. Yumer, and R. Urtasun, "Exploring adversarial robustness of multi-sensor perception systems in self driving," in *Conference on Robot Learning*, 2021, pp. 1013–1024. 4, 16

[42] F. Yin, Y. Zhang, B. Wu, Y. Feng, J. Zhang, Y. Fan, and Y. Yang, "Generalizable black-box adversarial attack with meta learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, pp. 1804–1818, 2023. 4

[43] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song, "Robust physical-world attacks on deep learning models," *arXiv: Cryptography and Security*, 2017. 4, 5

[44] J. Zheng, C. Lin, J. Sun, Z. Zhao, Q. Li, and C. Shen, "Physical 3d adversarial attacks against monocular depth estimation in autonomous driving," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24 452–24 461, 2024. 4

[45] N. Boucher, I. Shumailov, R. Anderson, and N. Papernot, "Bad

characters: Imperceptible nlp attacks," *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1987–2004, 2021. 4

[46] S. Eger, G. G. Şahin, A. Rücklé, J.-U. Lee, C. Schulz, M. Mesgar, K. Swarnkar, E. Simpson, and I. Gurevych, "Text processing like humans do: Visually attacking and shielding nlp systems," *ArXiv*, vol. abs/1903.11508, 2019. 4

[47] E. Shayegani, M. A. A. Mamun, Y. Fu, P. Zaree, Y. Dong, and N. B. Abu-Ghazaleh, "Survey of vulnerabilities in large language models revealed by adversarial attacks," *ArXiv*, vol. abs/2310.10844, 2023. 4, 14

[48] A. G. Chowdhury, M. M. Islam, V. Kumar, F. H. Shezan, V. Jain, and A. Chadha, "Breaking down the defenses: A comparative survey of attacks on large language models," *ArXiv*, vol. abs/2403.04786, 2024. 4, 12

[49] G. Apruzzese, M. Andreolini, L. Ferretti, M. Marchetti, and M. Colajanni, "Modeling realistic adversarial attacks against network intrusion detection systems," *Digital Threats: Research and Practice (DTRAP)*, vol. 3, pp. 1 – 19, 2021. 4, 14

[50] J. Vitorino, N. Oliveira, and I. Praça, "Adaptative perturbation patterns: Realistic adversarial learning for robust intrusion detection," *Future Internet*, vol. 14, p. 108, 2022. 4, 9, 10

[51] J. Zhang, Q. Yi, and J. Sang, "Towards adversarial attack on vision-language pre-training models," *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 4, 13, 17

[52] Z. Yin, M. Ye, T. Zhang, T. Du, J. Zhu, H. Liu, J. Chen, T. Wang, and F. Ma, "Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models," *ArXiv*, vol. abs/2310.04655, 2023. 4, 10, 13

[53] M.-I. Nicolae, M. Sinn, M.-N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. Molloy, and B. Edwards, "Adversarial robustness toolbox v1.0.0," *arXiv: Learning*, 2018. 5, 12, 17

[54] J. Zhang and C. Li, "Adversarial examples: Opportunities and challenges," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, pp. 2578–2593, 2018. 5, 8, 14, 15

[55] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *International Conference on Machine Learning*, 2018, pp. 274–283. 5

[56] D. Lee, S. Moon, J. Lee, and H. O. Song, "Query-efficient and scalable black-box adversarial attacks on discrete sequential data via bayesian optimization," in *International Conference on Machine Learning*, 2022, pp. 12 478–12 497. 6, 7, 11

[57] X. Wang, Y. Yang, Y. Deng, and K. He, "Adversarial training with fast gradient projection method against synonym substitution based text attacks," in *AAAI Conference on Artificial Intelligence*, 2020, pp. 13 997–14 005. 6, 8, 11, 14

[58] C. Laidlaw, S. Singla, and S. Feizi, "Perceptual adversarial robustness: Defense against unseen threat models," *ArXiv*, vol. abs/2006.12655, 2020. 6, 14

[59] J. Geiping, L. H. Fowl, G. Somepalli, M. Goldblum, M. Moeller, and T. Goldstein, "What doesn't kill you makes you robust(er): How to adversarially train against data poisoning," 2021. 6

[60] Z. Niu, Y. Sun, Q. Miao, R. Jin, and G. Hua, "Towards unified robustness against both backdoor and adversarial attacks," *IEEE transactions on pattern analysis and machine intelligence*, vol. PP, 2024. 6

[61] L. H. Fowl, M. Goldblum, P. yeh Chiang, J. Geiping, W. Czaja, and T. Goldstein, "Adversarial examples make strong poisons," in *Neural Information Processing Systems*, 2021, pp. 30 339–30 351. 6

[62] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, "On evaluating adversarial robustness," *ArXiv*, vol. abs/1902.06705, 2019. 7, 13, 14, 15, 17

[63] F. Tramèr, A. Kurakin, N. Papernot, D. Boneh, and P. Mcdaniel, "Ensemble adversarial training: Attacks and defenses," *ArXiv*, vol. abs/1705.07204, 2017. 7

[64] G. R. Machado, E. Silva, and R. Goldschmidt, "Adversarial machine learning in image classification: A survey toward the defender's perspective," *ACM Computing Surveys (CSUR)*, vol. 55, pp. 1 – 38, 2020. 7, 8, 10

[65] S. Chen, N. Carlini, and D. Wagner, "Stateful detection of black-box adversarial attacks," *Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*, 2019. 7

[66] Y. Zhao, T. Pang, C. Du, X. Yang, C. Li, N.-M. Cheung, and M. Lin, "On evaluating adversarial robustness of large vision-language models," *ArXiv*, vol. abs/2305.16934, 2023. 7

[67] I. Rosenberg, A. Shabtai, Y. Elovici, and L. Rokach, "Adversarial machine learning attacks and defense methods in the cyber security domain," *ACM Computing Surveys (CSUR)*, vol. 54, pp. 1 – 36, 2021. 9, 17

[68] Y. Ma, X. Zhu, and J. Hsu, "Data poisoning against differentially-private learners: Attacks and defenses," in *International Joint Conference on Artificial Intelligence*, 2019, pp. 4732–4738. 9, 15

[69] A. Paudice, L. Muñoz-González, A. György, and E. C. Lupu, "Detection of adversarial training examples in poisoning attacks through anomaly detection," *ArXiv*, vol. abs/1802.03041, 2018. 9, 12

[70] Z. Jin, J. Zhang, Z. Zhu, and H. Chen, "Benchmarking transferable adversarial attacks," *ArXiv*, vol. abs/2402.00418, 2024. 10, 13, 17

[71] E. Anthi, L. Williams, M. Rhode, P. Burnap, and A. Wedgbury, "Adversarial attacks on machine learning cybersecurity defences in industrial control systems," *J. Inf. Secur. Appl.*, vol. 58, p. 102717, 2020. 10

[72] R. Duan, X. Mao, A. K. Qin, Y. Yang, Y. Chen, S. Ye, and Y. He, "Adversarial laser beam: Effective physical-world attack to dnns in a blink," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16 057–16 066, 2021. 10, 11, 13

[73] Y. Cao, C. Xiao, B. Cyr, Y. Zhou, W. Park, S. Rampazzi, Q. A. Chen, K. Fu, and Z. M. Mao, "Adversarial sensor attack on lidar-based perception in autonomous driving," *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019. 10

[74] I. Ilahi, M. Usama, J. Qadir, M. Janjua, A. I. Al-Fuqaha, D. Hoang, and D. Niyato, "Challenges and countermeasures for adversarial attacks on deep reinforcement learning," *IEEE Transactions on Artificial Intelligence*, vol. 3, pp. 90–109, 2020. 10

[75] V. Voráček and M. Hein, "Provably adversarially robust nearest prototype classifiers," in *International Conference on Machine Learning*, 2022, pp. 22 361–22 383. 11

[76] S. Dai, S. Mahloujifar, C. Xiang, V. Sehwag, P.-Y. Chen, and P. Mittal, "Multirobustbench: Benchmarking robustness against multiple attacks," in *International Conference on Machine Learning*, 2023, pp. 6760–6785. 11, 12, 14, 16, 17

[77] D. Wang, W. Yao, T. Jiang, G. Tang, and X. Chen, "A survey on physical adversarial attack in computer vision," *ArXiv*, vol. abs/2209.14262, 2022. 11

[78] K. R. Mopuri, A. Ganeshan, and R. V. Babu, "Generalizable data-free objective for crafting universal adversarial perturbations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 2452–2465, 2018. 11

[79] Y. Zhong, X. Liu, D. Zhai, J. Jiang, and X. Ji, "Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15 324–15 333, 2022. 11

[80] E. Alhajjar, P. Maxwell, and N. D. Bastian, "Adversarial machine learning in network intrusion detection systems," *ArXiv*, vol. abs/2004.11898, 2020. 11

[81] F. Tramèr, J. Behrmann, N. Carlini, N. Papernot, and J. Jacobsen, "Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations," in *International Conference on Machine Learning*, 2020, pp. 9561–9571. 12, 13

[82] Y. Li, M. Cheng, C.-J. Hsieh, and T. C. Lee, "A review of adversarial attack and defense for classification methods," *The American Statistician*, vol. 76, pp. 329 – 345, 2021. 12

[83] A. E. Cinà, K. Grosse, A. Demontis, S. Vascon, W. Zellinger, B. Moser, A. Oprea, B. Biggio, M. Pelillo, and F. Roli, "Wild patterns reloaded: A survey of machine learning security against training data poisoning," *ACM Computing Surveys*, vol. 55, pp. 1 – 39, 2022. 12, 15

[84] V. T. Truong, L. B. Dang, and L. B. Le, "Attacks and defenses for generative diffusion models: A comprehensive survey," *ArXiv*, vol. abs/2408.03400, 2024. 12

[85] Y. Liu, Y. Cheng, L. Gao, X. Liu, Q. Zhang, and J. Song, "Practical evaluation of adversarial robustness via adaptive auto attack," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15 084–15 093, 2022. 13

[86] R. Feng, A. Hooda, N. Mangaokar, K. Fawaz, S. Jha, and A. Prakash, "Stateful defenses for machine learning models are

not yet secure against black-box attacks," *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023. 13

[87] J. Vitorino, I. Praça, and E. Maia, "Sok: Realistic adversarial attacks and defenses for intelligent network intrusion detection," *ArXiv*, vol. abs/2308.06819, 2023. 13, 17

[88] H. Fang, J. Kong, W. Yu, B. Chen, J. Li, S.-T. Xia, and K. Xu, "One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models," *ArXiv*, vol. abs/2406.05491, 2024. 13

[89] S. Chen, Z. He, C. Sun, and X. Huang, "Universal adversarial attack on attention and the resulting dataset damagenet," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, pp. 2188–2197, 2020. 13, 16, 17

[90] F. Croce, M. Andriushchenko, N. D. Singh, N. Flammarion, and M. Hein, "Sparse-rs: a versatile framework for query-efficient sparse black-box adversarial attacks," in *AAAI Conference on Artificial Intelligence*, 2020, pp. 6437–6445. 13, 14, 16

[91] Y. Deng, X. Zheng, T. Zhang, C. Chen, G. Lou, and M. Kim, "An analysis of adversarial attacks and defenses on autonomous driving models," *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–10, 2020. 13

[92] Y. Wang, T. Sun, S. Li, X. Yuan, W. Ni, E. Hossain, and H. Poor, "Adversarial attacks and defenses in machine learning-powered networks: A contemporary survey," *ArXiv*, vol. abs/2303.06302, 2023. 14

[93] R. Kumar, M. Nyström, J. Lambert, A. Marshall, M. Goertzel, A. Comissoneru, M. Swann, and S. Xia, "Adversarial machine learning-industry perspectives," *2020 IEEE Security and Privacy Workshops (SPW)*, pp. 69–75, 2020. 14, 17

[94] M. A. Ramírez, S.-K. Kim, H. A. Hamadi, E. Damiani, Y.-J. Byon, T.-Y. Kim, C. Cho, and C. Yeun, "Poisoning attacks and defenses on artificial intelligence: A survey," *ArXiv*, vol. abs/2202.10276, 2022. 14, 15

[95] A. Jeddi, M. Shafiee, M. Karg, C. Scharfenberger, and A. Wong, "Learn2perturb: An end-to-end feature perturbation learning to improve adversarial robustness," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1238–1247, 2020. 14

[96] "Deep k-nn defense against clean-label data poisoning attacks," in *ECCV Workshops*, 2019, pp. 55–70. 15

[97] Y. Gao, B. G. Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal, and H. Kim, "Backdoor attacks and countermeasures on deep learning: A comprehensive review," *ArXiv*, vol. abs/2007.10760, 2020. 15

[98] A. Salem, A. Bhattacharyya, M. Backes, M. Fritz, and Y. Zhang, "Updates-leak: Data set inference and reconstruction attacks in online learning," in *USENIX Security Symposium*, 2019, pp. 1291–1308. 15

[99] J. Hayase, W. Kong, R. Somani, and S. Oh, "Spectre: Defending against backdoor attacks using robust statistics," *ArXiv*, vol. abs/2104.11315, 2021. 15

[100] S. Han, C. Lin, C. Shen, Q. Wang, and X. Guan, "Interpreting adversarial examples in deep learning: A review," *ACM Computing Surveys*, vol. 55, pp. 1 – 38, 2023. 15, 16

[101] X. Qi, K. Huang, A. Panda, M. Wang, and P. Mittal, "Visual adversarial examples jailbreak aligned large language models," in *AAAI Conference on Artificial Intelligence*, 2023, pp. 21 527–21 536. 16

[102] E. Shayegani, Y. Dong, and N. B. Abu-Ghazaleh, "Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models," in *International Conference on Learning Representations*, 2023. 16

[103] J. Li, J. Y. Lee, Y. Yang, J. Sun, and K. Tomsovic, "Conaml: Constrained adversarial machine learning for cyber-physical systems," *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, 2020. 16

[104] Z. Cheng, J. Liang, H. Choi, G. Tao, Z. Cao, D. Liu, and X. Zhang, "Physical attack on monocular depth estimation with optimal adversarial patches," *ArXiv*, vol. abs/2207.04718, 2022. 16

[105] X. Jia, Y. Zhang, B. Wu, K. Ma, J. Wang, and X. Cao, "Las-at: Adversarial training with learnable attack strategy," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13 388–13 398, 2022. 16

[106] T. Le, J. Lee, K. Yen, Y. Hu, and D. Lee, "Perturbations in the wild: Leveraging human-written text perturbations for realistic adversarial attack and defense," *ArXiv*, vol. abs/2203.10346, 2022. 16

[107] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, "Pixeldefend: Leveraging generative models to understand and defend against adversarial examples," *ArXiv*, vol. abs/1710.10766, 2017. 17

[108] N. Ruiz, S. A. Bargal, and S. Sclaroff, "Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems," *ArXiv*, vol. abs/2003.01279, 2020. 17

[109] D. Li, Q. Li, Y. Ye, and S. Xu, "Arms race in adversarial malware detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 55, pp. 1 – 35, 2020. 17

[110] H. Baniecki and P. Biecek, "Adversarial attacks and defenses in explainable artificial intelligence: A survey," *ArXiv*, vol. abs/2306.06123, 2023. 17