Analysis of Adversarial Attacks and Defense Mechanisms on Machine Learning Models

Wakkmubura M.M.S.R
Faculty of Information Technology
University of Moratuwa
Wakumburammsr.21@uom.lk

Abstract – **Adversarial Machine Learning (AML) is a recently introduced technique that aims to deceive Machine Learning (ML) models by providing falsified inputs to render those models ineffective. The adoption of Artificial Intelligence (AI) in critical technological domains, such as smart cities and self-driving cars, makes mitigating these adversarial threats a fundamental security challenge. Consequently, while many researchers concentrate on developing new AML attacks, the significance of robust defense strategies is often overlooked.This article constitutes a comprehensive survey of existing literature focused primarily on AML defense solutions. This introduce a large-scale, domain-agnostic taxonomy of recent AML defense techniques applied across various fields, including audio, cyber-security, Natural Language Processing (NLP), and computer vision. The methodological analysis systematically explores the effectiveness of these solutions, comparing them based on whether they are attack- and/or domain-agnostic, utilize appropriate AML evaluation metrics, and share their source code or evaluation datasets.The goal is to organize existing knowledge and extract insights about the current AML defense state. To the best of our knowledge, this work represents the first survey attempting to systematize existing knowledge focusing solely on defense solutions against AML attacks. By elaborating on the findings of this taxonomy and analysis, the survey identifies several gaps in the literature and provides innovative directions for future research aimed at proposing more efficient and robust defense solutions to address the ongoing threat of AML**

Keywords - Adversarial Machine Learning, Adversarial Examples, Evasion Attacks, Poisoning Attacks, Adversarial Training, Defense Mechanisms, Detection-Based Defenses,

I INTRODUCTION

The implementation of the Artificial Intelligence (AI) principles and the use of the Machine Learning (ML) mechanisms have brought the revolution in a range of technological fields [1]. ML is a subdiscipline of AI and is aimed at adaptation to new situations and the application of patterns on data to find solutions. This has introduced emerging trends of smarter cities, cars, self-driving cars, and autonomous systems [2].

Despite the excellent results of deep learning (DL) models, many products and services whose operations are software intensive, including the security-critical sectors of industry, manufacturing, and medicine, remain vulnerable [3]. More recent work on applications such as image classification has proved that they are susceptible to strategic adversarial examples (AEs). These adversarialized inputs exploit weaknesses in model decision boundaries.

AML (Adversarial Machine Learning) is a fairly recently developed technique that aims to provide deception of ML models [4]. These are perturbations, usually small or invisible to human beings, but may have serious consequences on the outputs of the model and lead to inaccurate estimations, which is detrimental to the strength of DL systems [5]. There are real repercussions associated with the exploitation of ML vulnerabilities, such as the wrong classification of ransomware by anti-malware scanners or the false diagnosis of medical imaging systems [6].

In turn, it is mandatory to assure the strength, dependability, and reliability of ML systems in operational environments [7]. Nonetheless, the majority of researchers usually pay attention to identifying new AML attacks that can undermine existing ML infrastructures, but do not pay as much attention to the importance of defense strategies [4].

The recent surveys of literature tend to focus on a particular field (e.g., computer vision or cybersecurity) and discuss the attacks on AI systems in detail but rarely describe potential defenses [1]. This has introduced a gap in literature in terms of complete surveys that detail only the available defense solutions to AML attacks. Moreover, the research has been criticized for having a high score on rigor (well-designed and reported methods) but a low score on relevance (lacking practical applicability for industry practitioners). There are still numerous defense measures that are said to be in their infancy [4].

In an attempt to fill these important knowledge gaps and support the work of the scientific community in offering more resilient defense measures, the article presents a domain-agnostic and large-scale survey of defense strategies aimed at addressing AML attacks [4].

The primary objective of the literature review is to determine and evaluate the various adversarial threats to both ML models and the existing methods and strategies to prevent such attacks.

The objectives of this survey, in particular, are:

- The paper should give a general review of such attacks and their effects on the various kinds of machine learning models.
- Attacker knowledge and attack methodology classification.
- Research into the areas and applications of ML systems most susceptible to adversarial attacks, e.g., audio, cybersecurity, NLP, and computer vision.
- Survey and analysis of existing defense strategies, such as input preprocessing, adversarial training, model regularization, ensemble strategies, and detection strategies.
- Recognizing gaps, problems, and limitations in the existing methods of adversarial defense, and recommending future research programs to create more robust and generalizable defenses.

As far as we know, this work is the first survey that attempts to systematize the available knowledge only on the issue of defense solutions to AML and also includes a comparison of the academic soundness and applicability of published solutions [4]. The review can be used to improve the work on developing and addressing the current and growing threat of AML.

II CLASSIFICATION OF ADVERSARIAL ATTACKS

Adversarial Machine Learning (AML) attacks can be classified across multiple dimensions: attack phase, adversary knowledge, and attack objectives. This framework helps researchers and practitioners understand threat vectors and develop targeted defenses [4, 8].

A Attack Phase: Poisoning vs Evasion

A.1 Poisoning Attacks (Training-Time / Causative Attacks)

Poisoning attacks target the training phase by injecting or modifying falsified training data, compromising the model's learning foundation. These causative attacks include label flipping, data injection, and backdoor attacks (stealthy attacks with hidden triggers) [8]. Once trained on poisoned data, the model remains fundamentally unreliable throughout its operational lifetime.

A.2 Evasion Attacks (Test-Time / Exploratory Attacks)

Evasion attacks target trained models during testing by manipulating individual samples without accessing training data. These exploratory attacks generate adversarial examples (AEs)- subtle perturbations that cause misclassification while remaining imperceptible to humans. Evasion attacks are the most practical type as they require neither model nor data access [1].

A.3 Industry vs. Academia

Academia (particularly computer vision) focuses heavily on adversarial examples and evasion attacks, often under controlled laboratory conditions [4]. In contrast, industry practitioners prioritize data poisoning, backdoor attacks, and model theft as more realistic threats to production systems [9]. Real-world evidence confirms that poisoning attacks and model manipulation occur in industrial settings [8].

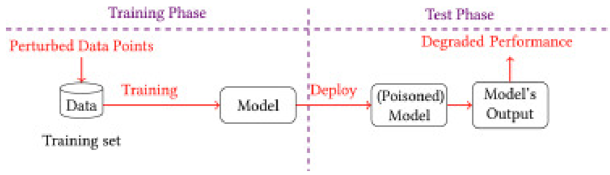


Figure 1: Temporal classification of adversarial attacks: poisoning attacks occur during the training phase, while evasion attacks occur during the testing phase of the ML pipeline [2].

B Adversary Knowledge: White-box, Black-box, Gray-box

Attack severity depends on the adversary's knowledge of the target model [1].

B.1 White-box Attacks (WBA)

In white-box attacks, the adversary has complete knowledge-access to the algorithm's architecture, hyperparameters, training data, and gradients. This represents the most potent attack type and is commonly used to evaluate upper bounds on robustness. White-box attacks operate under the principle that if a model resists them, it should also resist weaker black-box and gray-box attacks.

B.2 Black-box Attacks (BBA)

In black-box attacks, the adversary has no knowledge of the system, only observing inputs and outputs. Despite this limitation, black-box attacks are more realistic in practice, as they simulate real-world scenarios where attackers lack internal model access. Attacks on public APIs and deployed cloud services exemplify this scenario.

B.3 Gray-box Attacks (GBA)

Gray-box attacks assume partial knowledge-some model features, class labels, or hidden layer outputs. Also called semi-white-box attacks, they represent realistic scenarios with limited insider information or partial reverse engineering [2].

B.4 Transferability

Transferability is the property that adversarial examples crafted against one model often fool different models. This enables black-box attacks via substitute models: attackers train a substitute model and apply its adversarial examples to the target model. Effective defenses should aim to reduce transferability and make attacks less portable across architectures [4].

C Attack Goals: Targeted vs Untargeted

C.1 Untargeted Attacks

Untargeted attacks aim to produce any incorrect output, forcing the model to predict a label different from the true label. The attacker does not specify which wrong class is acceptable. Optimization minimizes the probability of the correct label, making untargeted attacks easier to generate than targeted variants [1].

C.2 Targeted Attacks

Targeted attacks force the model to output a specific target label predetermined by the attacker. This stricter constraint makes targeted attacks harder to generate but substantially more dangerous in safety-critical domains. In medical imaging, targeted attacks can deliberately cause misdiagnosis of critical diseases [6]. In autonomous driving, they can force recognition of a stop sign as a speed limit sign, causing specific traffic violations. These domain-specific dangers justify the additional optimization effort required.

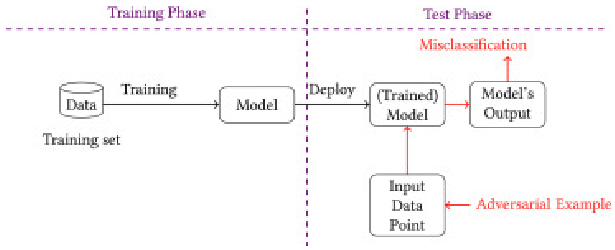


Figure 2: Conceptual illustration of data poisoning attacks: malicious samples are injected into the training dataset, corrupting the model's learning process and resulting in a compromised classifier with systematic vulnerabilities [8].

III COMMON ATTACK METHODS

Adversarial attacks take advantage of the model vulnerabilities in its decision boundaries to develop actively engineered inputs. Originally created to work with computer vision, these techniques can be effectively applied to cybersecurity, natural language processing, audio processing, and medical imaging, using their underlying principles, which are known as core principles of computer vision learning algorithms, as the foundations of their application to these domains [1,4]. This part elaborates the common attack methods and marks out their practical connotations..

A Fast Gradient Sign Method (FGSM)

The first and the least complicated gradient-based attack is FGSM seen by Goodfellow et al. [10], is the earliest and simplest gradient-based attack. It perturbs the input by taking a single step in the direction of the sign of the gradient of the loss function. Despite its simplicity, FGSM remains highly influential as:

- a fast and scalable attack for large datasets,
- a baseline tool for evaluating model robustness,
- a common component in adversarial training pipelines.

Its computational efficiency allows rapid generation of adversarial examples, demonstrating that even small, visually imperceptible changes can mislead deep neural networks.

B Basic Iterative Method (BIM) / Iterative FGSM

BIM extends FGSM by applying multiple smaller FGSM steps instead of one large perturbation [1]. BIM stays within the constraint set L_∞ by clipping pixel values after each step. Compared to FGSM, BIM:

- produces much stronger attacks,
- is particularly effective in white-box settings,
- has been used to break early defenses based on gradient masking.

Studies show that BIM can achieve near 100% attack success rates in undefended image classifiers [1].

C Projected Gradient Descent (PGD)

The most commonly used first-order adversarial attacks are iterated FGSM projectioned to the feasible region and PGD, which is the approximation of the worst-case loss in an L_infty ball, and is used frequently in robustness benchmarks. [11].

PGD has become the default attack used in adversarial training, where its strength helps improve robustness across vision, NLP, and cybersecurity systems [4].

D DeepFool (DF)

DeepFool is an L_2-based attack that iteratively pushes the input toward the closest decision boundary [1]. It is designed to produce minimal, near-imperceptible perturbations and is widely used to measure the vulnerability margin of classifiers.

DeepFool demonstrates that many models have extremely small classification margins, highlighting structural weaknesses in their learned decision surfaces.

E Carlini & Wagner (C&W)

The C&W attack [5] uses an optimization-based formulation and can operate under L_0, L_2, or L_∞ norms. It is known for:

- bypassing widely used defenses such as Defensive Distillation [12],
- achieving near 100% success rates across datasets,
- producing adversarial examples that remain effective even after transformations.

The C&W attack is considered a benchmark for evaluating strong, targeted attacks and has significantly influenced the development of robust defense strategies.

F Jacobian-based Saliency Map Attack (JSMA)

JSMA is an L_0 attack that uses saliency maps derived from the Jacobian to identify the most influential input features [3]. By modifying only a few features (e.g., pixels), it generates sparse adversarial examples.

Dense attacks are less detectable by conventional denoising methods, and hence, this is why JSMA is significant in security sensitive tasks like malware classification and intrusion detection.

G AutoAttack (AA)

AutoAttack is a parameter-free ensemble attack framework that combines four strong attacks, including FAB, APGD, and Square Attack [4]. It aims to eliminate researcher-induced bias from incorrectly tuned attacks.

Since 2020, AutoAttack has become a standard tool for reliable robustness evaluation, widely adopted in benchmarking competitions and academic studies.

H Cross-Domain Application

Although most attack methods originated in computer vision, research shows broad applicability across domains:

- Cybersecurity: Attacks manipulate API call sequences and malware features to evade detection systems [2, 13].
- NLP: Adversarial misspellings, synonym substitution, and embedding perturbations affect sentiment analysis and text classification [14].
- Audio: Perturbations on waveforms or spectrogram inputs mislead speaker recognition and keyword-spotting systems [15].
- Medical Imaging: Minor perturbations can cause misdiagnosis in X-ray, CT, or MRI models [6].

These developments confirm that adversarial vulnerabilities exist in nearly every modality, highlighting the need for domain-agnostic defense frameworks [7].

IV DEFENSE MECHANISMS AGAINST ADVERSARIAL ATTACKS

Defense mechanisms against Adversarial Machine Learning (AML) attacks are typically grouped into mitigation (proactive) and detection (reactive) strategies. Mitigation techniques enhance model robustness, while detection methods flag and reject adversarial inputs before classification. Survey results indicate that the majority of existing defenses focus on mitigation rather than detection [4].

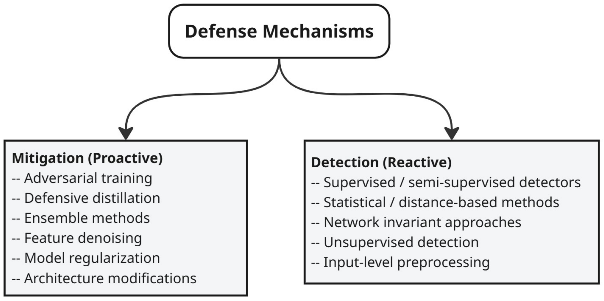

Defense Mechanisms
Mitigation (Proactive)
-- Adversarial training
-- Defensive distillation
-- Ensemble methods
-- Feature denoising
-- Model regularization
-- Architecture modifications
Detection (Reactive)
-- Supervised / semi-supervised detectors
-- Statistical / distance-based methods
-- Network invariant approaches
-- Unsupervised detection
-- Input-level preprocessing

Figure 3: High-level taxonomy of mitigation-based and detection-based defenses against adversarial attacks [4].

A Mitigation (Proactive Defenses)

Mitigation refers to proactive defenses that increase the intrinsic robustness of ML/DL models against AML attacks [4].

A.1 Adversarial Training

Adversarial training (AT) enhances robustness by training on datasets containing both benign and adversarial examples. The approach uses a min–max formulation: internal maximization seeks the most harmful adversarial examples (maximizing loss), while external minimization improves network parameters to mitigate this threat (minimizing expected maximum loss) [11].

Variations include standard AT with FGSM attacks, curriculum-based adversarial training, ensemble adversarial training (incorporating transferred perturbations for black-box robustness), and approaches such as MemLoss that reuse previously generated adversarial examples to improve both robustness and accuracy [4].

A.2 *Defensive Distillation*

Defensive distillation involves training a teacher model normally, then generating soft labels (class probabilities) using softmax at temperature T. A student model trains using these soft labels, also at temperature T. During testing, the student uses temperature 1.

The defense aims for gradient masking, where softmax smoothing trains the network to have gradients close to 0, reducing sensitivity to input perturbations. However, later research showed that defensive distillation is not robust to powerful attacks like C&W, which can find adversarial examples with high success rates.

A.3 *Ensemble Methods*

Ensemble learning combines multiple ML/DL algorithms, often providing more robustness than single-model defenses [1]. Examples include AppCon (ensemble of application-specific detectors), IoT attack detection systems, and hybrid approaches combining vocoders with randomized smoothing for audio system robustness [4, 15].

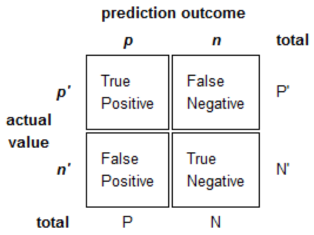
A.4 *Feature Denoising and Architectural Modifications*

Feature denoising removes noise from internal representations during inference, typically after specific model layers. High-Level Representation Guided Denoiser (HGD) uses a loss function comparing top-level outputs from original and adversarial examples, rather than pixel-level reconstruction.

In audio systems, denoisers are often combined with adversarial training [15]. Medical imaging defenses explore dual-batch normalization and replacing max-pooling with average-pooling to better capture global context and improve robustness. Overall, architectural modifications remain relatively under-investigated compared to attack-agnostic preprocessing or training-time defenses [4].

A.5 *Model Regularization*

Model regularization imposes adversarial examples as regularization during optimization, combining minimum loss with weighted regularization from adversarial losses [4]. Gradient masking attempts to maintain gradients near 0, reducing sensitivity to perturbations. However, while these ideas may provide partial robustness against some white-box attacks, they often fail against black-box and stronger adaptive attacks.



		prediction outcome		
		p	n	total
actual value	p'	True Positive	False Negative	P'
	n'	False Positive	True Negative	N'
total		P	N	

Figure 4: Confusion matrix used in evaluating adversarial example detection performance. [17]

B *Detection (Reactive Defenses)*

Detection methods flag and reject adversarial inputs before classification, acting as reactive defenses that determine whether an input is adversarial [1, 4].

B.1 *Supervised and Semi-supervised Detectors*

Supervised detectors deploy auxiliary ML/DL models trained to recognize AML attacks, relying on distinguishable features between clean and adversarial inputs. Detectors exploit features from behavioral differences, including softmax probabilities or neuron activation patterns.

RAID (Randomized Adversarial-Input Detection) and ESRM (Enhanced Spatial Rich Model) exemplify supervised detection. ESRM uses steganographic features and binary classification to detect perturbations in pixel space [16].

B.2 *Statistical and Distance-based Methods*

These methods calculate statistical properties (distribution/density metrics) to distinguish clean from adversarial samples, assuming adversarial samples lie outside the normal data manifold.

Maximum Mean Discrepancy (MMD) is a kernel-based statistical test distinguishing adversarial examples from training data. Principal Component Analysis (PCA) detects that later PCA components of adversarial examples show larger variance. Kernel Density (KD) estimation and histogram-based features from layer outputs can be used to train binary SVM classifiers [18, 19].

B.3 *Network Invariant Approaches*

These methods analyze violations of network invariants, assuming clean and adversarial samples produce different feature maps and activation values across layers. SafetyNet quantizes last ReLU activation layers and trains binary SVM classifiers based on activation patterns, while Neural-network Invariant Checking (NIC) monitors activation patterns and feature maps by building layer-specific models [4].

B.4 Unsupervised Detection

Unsupervised detectors train using only clean data, requiring no prior attack knowledge and effectively detecting unknown attacks, particularly black-box attacks [18].

Feature Squeezing destroys perturbations through bit-depth reduction and spatial smoothing. MagNet is a denoiser-based framework with detector and reformer components. Reconstruction-based detection like Image Reconstruction Differences (IRD) measures differences between original and reconstructed examples [20].

Local Intrinsic Dimensionality (LID) measures sample outlierness to distinguish adversarial from original samples. Gaussian Mixture Models (GMM) approximate hidden state distributions using clean training data. Deep Neural Rejection (DNR) and Selective and Feature-based Adversarial Detection (SFAD) combine multiple cues such as confidence, uncertainty, and mismatch detection [4]. Object-based methods like UnMask extract and compare object features with training data of the predicted class.

C Input-Level Preprocessing

Input-level preprocessing operates on data before classification, aiming to destroy or mitigate adversarial perturbations.

C.1 Transformation-based Defenses

Transformation-based defenses apply operations to destroy perturbations.

Bit-depth reduction reduces color bit depth (e.g., 24-bit to 8-bit) to squeeze unnecessary features; spatial smoothing uses blur or non-local smoothing to destroy adversarial noise. Obfuscation methods like Protective Operation Chain randomly mix features, increasing attacker uncertainty [21].

Spectrogram pre-processing transforms audio waveforms to the frequency domain for sound classification, while sequence squeezing merges semantically similar features (e.g., API call sequences) for RNN classifiers [4]. Spell-checking corrects character-level adversarial text attacks before classification [14].

C.2 Dual Role: Detection and Mitigation

Many input-level preprocessing methods serve dual functions:

Mitigation (Proactive): Modifying input data removes perturbations before inference, hardening models while preserving clean inputs. *Detection (Reactive):* Methods exploit adversarial samples' sensitivity to transformations. If predictions differ significantly between original and transformed inputs, the sample is flagged as adversarial.

Feature Squeezing exemplifies this duality: it proactively reduces perturbations while simultaneously detecting adversarial inputs by comparing original and squeezed predictions using L_1 metrics. Detection Filter similarly classifies inputs as adversarial when labels differ between original and filtered versions [22].

V STRENGTHS AND WEAKNESSES OF ADVERSARIAL TRAINING

One of the most well-known and thoroughly studied defenses to adversarial attacks is Adversarial Training (AT). Being a mitigation-based method, AT makes models more robust by introducing adversarial examples during the training process, which typically is based on a min-max optimization formalized by Madry et al. However, being a popular field, AT has significant strengths and weaknesses in computer vision, cybersecurity, NLP, and medical imaging spheres.

A Strengths of Adversarial Training.

AT has proven to be very advantageous to the enhancement of model robustness:

Increased robustness. Crowding out adversarial examples in the training data also increases tolerance to perturbation and the worst case accuracy of ATP. Initial experiments had found that there was a significant error decrease when applying FGSM based adversarial examples on MNIST. Likewise, first-order attacks are well resisted on models that have been trained on PGD adversarial examples as well as vice versa [11].

Improvement within areas. There are computer vision, cybersecurity, and NLP where AT has been found to be effective. Min–max adversarial training in intrusion detection systems (IDS) enhanced resistance to evasion in attacks and a large decrease in success of attacks [13]. PGD-AT and MART variants of AT had superior classification and segmentation performance on various datasets in the medical imaging domain [6].

Potential of better generalization (in NLP). Some AT strategies with an NLP focus, especially those implemented on BERT and RoBERTa embeddings, have demonstrated both robustness and generalization improvements at the same time-breaking the longstanding robustness accuracy trade-off.

B B. Weaknesses and Limitations

Despite its strengths, AT faces several known drawbacks:

Trade-off with clean accuracy. AT often reduces performance on clean, non-adversarial inputs. This degradation has been observed in computer vision classifiers and deep learning IDS models, where adversarially trained detectors experienced slight but consistent decreases in natural accuracy [4].

High computational cost. Generating adversarial examples for every training step is computationally expensive, especially when using iterative attacks such as PGD or C&W. Although gradient-recycling techniques reduce overhead, AT remains significantly more resource-intensive than standard training.

Limited generalization to unseen attacks. Models trained on a specific attack type often fail to defend against unseen or adaptive attacks. Studies in cybersecurity showed that AT improved resilience only against the exact attack used during training, with limited transfer to other evasion strategies [21].

Mixed robustness against black-box attacks. Performance against black-box attacks varies across domains. While Ensemble Adversarial Training improves resistance to transfer-based attacks, AT can still be bypassed by strong black-box and gray-box adversaries in real-world scenarios [4].

Dependence on large adversarial datasets. Effective AT requires diverse and high-quality adversarial examples; however, increasing the quantity of adversarial samples does not guarantee improved robustness. Several studies observed diminishing returns when scaling adversarial augmentation [1].

Overall, Adversarial Training remains the foundational defense strategy due to its substantial robustness improvements across domains. However, the robustness–accuracy trade-off, high computational demands, weak generalization to unseen attacks, and inconsistent protection against black-box threats highlight its limitations. AT is effective, but far from a universal or complete solution.

VI GAPS AND CHALLENGES IN ADVERSARIAL DEFENSE RESEARCH

The literature presents significant gaps and issues in the Adversarial Machine Learning (AML) defenses, with references to the constraints of architecture, domain differences, and the new vulnerabilities of neural systems with complex AI systems. These holes demonstrate that the development of theory is not connected to the demands of practical implementation in the real world. Such gaps demonstrate the lack of connection between theoretical work and the requirements of practice in the field of deployment into reality. [4, 7].

A Architecture-Level and System-Level Gaps

The AML research tends to concentrate on the defense of individual ML models and ignores the wider *ML system architecture* into which these models are embedded. The current literature lacks an approach to AML through the lens of software architecture, although it has been proven that system-level designs can have a significant effect on security posturee [7]. he lack of system-wide threat modelling models adds to the current security arms race in machine learning and constrain the applicability of most of the suggested defenses. Future studies must focus on the use of system-level threats models of the ML instead of independent model defenses.

Zero-day, backdoor, and model theft have been under-researched in the literature, despite industry players considering these forms of attacks more applicable and harmful than conventional adversarial examples [7]. ittle research has offered viable countermeasures against these threats of high risks.

B Domain Maturity Gaps

The computer vision field is overrepresented in AML defense research. On the contrary, NLP and audio are still at the youthful stages of maturity, with fewer datasets and fewer sound evaluation procedures available to them [1]. Adversarial text or audio input design is more difficult to constrain by semantics and perception, by nature.

Recent architectures-GANs, VAEs, and transformer-based models-have been popular in practice, but have not been studied with regards to AML attacks and defenses. Surveys also report that only recently diffusion models and Large Language Models (LLMs) have become the focus of the attention of the AML community [4].

C LLM and Multimodal Vulnerabilities

Contemporary multimodal AI systems, especially vision-language model combination systems, have weaknesses that AML defenses cannot mitigate. Recent findings show that adversarially crafted or steganographically modified images can embed hidden commands that propagate through image-processing pipelines into LLMs, bypassing safety mechanisms [23]. These vulnerabilities expose risks such as data leakage, covert communication, and manipulation of downstream reasoning.

Attacks specific to LLM like hidden triggers embedded in data, steganography, or multimodal jailbreaks are sparsely studied, although the use of LLMs in actual applications is fast expanding.

D Domain-Specific Challenges

Medical Imaging.

Internet of Things (IoT). IoT environments demand real-time, cross-layer detection methods capable of operating under severe resource constraints (processing power, memory, bandwidth). The heterogeneity and scale of IoT deployments introduce additional attack surfaces, yet AML defenses for IoT remain limited [7].

E Summary

Despite a variety of AML defenses suggested, these do not bridging the gap between the theoretical strength and the reality, practical, scalable level of security. The absence of diversity in the domain, the inability to include the architectural and systemic views, the inadequate defense against zero-day and the LLM-based challenges, and the lack of assessment frameworks introduce the necessity to adopt more holistic measures to protect the modern ML systems.

VII DISCUSSION

This was a systematic literature review of the field of Adversarial Machine Learning (AML), which developed a systematic taxonomy of attacks categorized by target phase (poisoning vs. evasion), knowledge of the adversary (white-box, black-box, gray-box), and purpose (targeted vs. untargeted). We have found three major categories of defense mechanisms, which include mitigation (proactive hardening), detection (reactive rejection), and input-level preprocessing.

Although Adversarial Training (AT) is by far the most actively researched defense mechanism, which provides enhanced robustness in various domains, it is not a panacea. These inherent limitations encompass trade-offs against clean accuracy, high complexity of computation, and bad generalization to unknown attacks. Together with the existence of critical gaps in evaluation practices and the lack of scalable, globalized solutions to defense, these findings highlight why there is an urgent need to develop holistic evaluation frameworks on AML defenses.

This review is the first systematic survey aimed at studying AML defense solutions and their domain-neutral taxonomy in detail, which is a necessary basis of future studies and effective implementation of effective defenses.

References

1 S. Qiu, Q. Liu, S. Zhou, and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," *Appl. Sci.*, vol. 9, p. 909, 2019.

2 K. Sadeghi, A. Banerjee, and S. K. Gupta, "A system-driven taxonomy of attacks and defenses in adversarial machine learning," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, pp. 450–467, 2020.

3 N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European symposium on security and privacy (EuroS&P)*, 2016, pp. 372–387.

4 P. Bountakas, A. Zarras, A. Lekidis, and C. Xenakis, "Defense strategies for Adversarial Machine Learning: A survey," *Computer Science Review*, 2023, preprint submitted on August 15, 2023.

5 N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 39–57. [Online]. Available: https://arxiv.org/abs/1608.04644

6 G. W. Muoka, D. Yi, C. C. Ukwuoma, A. Mutale, C. J. Ejiyi, A. K. Mzee, E. S. A. Gyarteng, A. Alqahtani, and M. A. Al-Antari, "A comprehensive review and analysis of deep learning-based medical image adversarial attack and defense," *Mathematics*, vol. 11, no. 20, p. 4272, 2023.

7 F. V. Jedrzejewski, L. Thode, J. Fischbach, T. Gorschek, D. Mendez, and N. Lavesson, "Adversarial Machine Learning in Industry: A Systematic Literature Review," *Computers & Security*, vol. 145, p. 103988, 2024.

8 A. E. Cinà, K. Grosse, A. Demontis, S. Vascon, W. Zellinger, B. A. 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, 2023.

9 G. Apruzzese, H. S. Anderson, S. Dambra, D. Freeman, F. Pierazzi, and K. Roundy, ""real attackers don't compute gradients": Bridging the gap between adversarial ml research and practice," in *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2023, pp. 339–364. [Online]. Available: https://arxiv.org/abs/2212.14315

10 I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *International Conference on Learning Representations (ICLR)*, 2015, arXiv preprint arXiv:1412.6572 (2014).

11 A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *6th International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://arxiv.org/abs/1706.06083

12 N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symposium on Security and Privacy (SP)*, 2016, pp. 582–597.

13 G. Apruzzese, M. Andreolini, M. Colajanni, and M. Marchetti, "Hardening random forest cyber detectors against adversarial attacks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, pp. 427–439, 2020.

14 D. Pruthi, B. Dhingra, and Z. C. Lipton, "Combating Adversarial Misspellings with Robust Word Recognition," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

15 M. Pal, A. Jati, R. Peri, C.-C. Hsu, W. AbdAlmageed, and S. Narayanan, "Adversarial defense for deep speaker recognition using hybrid adversarial training," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6164–6168.

16 J. Liu, W. Zhang, Y. Zhang, D. Hou, Y. Liu, H. Zha, and N. Yu, "Detection Based Defense Against Adversarial Examples from the Steganalysis Point of View," in *Proceedings of the 2019 IEEE/CVF Conference on Computer*

Vision and Pattern Recognition (CVPR), 2019, pp. 4820–4829.

17 W. Xu, Y. Qi, and K. Evans, "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks," *arXiv preprint arXiv:1710.08290*, 2017.

18 X. Ma, B. Li, Y. Wang, S. M. Erfani, S. N. R. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, and J. Bailey, "Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality," in *6th International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://arxiv.org/abs/1801.02613v3

19 K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Advances in Neural Information Processing Systems*, 2018.

20 J. Sun and M. Yi, "Detecting adversarial examples using image reconstruction differences," *Soft Computing*, vol. 27, pp. 7863–7877, 2023.

21 G. Apruzzese and V. S. Subrahmanian, "Mitigating Adversarial Gray-Box Attacks Against Phishing Detectors," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2022.

22 B. Liang, H. Li, M. Su, X. Li, W. Shi, and X. Wang, "Detecting Adversarial Image Examples in Deep Neural Networks with Adaptive Noise Reduction," *IEEE Trans. Dependable Secur. Comput.*, 2021.

23 N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea, and C. Raffel, "Extracting training data from large language models," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2633–2650.