Computer Science Review 61 (2026) 100963

Contents lists available at ScienceDirect

Computer Science Review

journal homepage: www.elsevier.com/locate/cosrev

Beyond vulnerabilities: A comprehensive survey of adversarial attacks across domains☆

Dimitrios Christos Asimopoulos[a, b], Panagiotis Radoglou-Grammatikis[c, d,*], Georgios Th. Papadopoulos[e], Panagiotis Sarigiannidis[c]

[a] Department of Information and Electronic Engineering, International Hellenic University, Sindos Campus, Thessaloniki, Greece
[b] MetaMind Innovations, Kila, Kozani, Greece
[c] Department of Electrical and Computer Engineering, University of Western Macedonia, Campus ZEP, Kozani, Greece
[d] K3Y Ltd, Sofia, Bulgaria
[e] Department of Informatics and Telematics, Harokopio University of Athens, Athens, Greece

HIGHLIGHTS

- **Comprehensive Analysis:** Reviews adversarial attacks across multiple data types and their impact on machine learning models.
- **Taxonomy Development:** Proposes a structured taxonomy of adversarial attacks aligned with the MITRE ATLAS framework.
- **Vulnerability Identification:** Identifies vulnerabilities across data modalities exploited by adversarial attacks.
- **Attack Categorization:** Categorizes adversarial attack techniques across different data types and ML systems.
- **Domain-Specific Taxonomy:** Examines adversarial attacks across domains including IoT, healthcare, NLP, speech, and LLMs.

ARTICLE INFO

Keywords:
Artificial intelligence
Adversarial attacks
Cybersecurity
Internet of things
Machine learning
White-box
Black-box

ABSTRACT

Adversarial attacks present significant risks to machine learning (ML) systems, exploiting model vulnerabilities and threatening the integrity, security, and trustworthiness of applications across multiple sectors. This paper provides a comprehensive review of adversarial attack types—white box, black box, and other type of attacks—and examines tailored attacks and defense mechanisms across domains such as Internet of Things (IoT), healthcare, industrial control systems, autonomous vehicles, speech recognition, natural language processing (NLP), finance, and Large Language Models (LLMs). Each domain introduces unique adversarial challenges and demands specific countermeasures, from anomaly detection to adversarial training and robust model architectures. By systematically categorizing both attack methodologies and defense strategies, this survey offers a holistic understanding of adversarial dynamics across fields, highlighting critical areas for further research and the development of resilient, cross-domain ML defenses.

1. Introduction

Artificial intelligence (AI) and machine learning (ML) have advanced dramatically in recent years, leading to the widespread use of these technologies in a variety of fields, including healthcare, finance, autonomous systems, and cybersecurity. While these breakthroughs have brought several benefits, they have also exposed inherent flaws in AI and ML systems.

AI attacks encompass a wide range of malicious activities that exploit vulnerabilities in AI and ML systems, leading to potentially harmful outcomes in various domains. These attacks can manifest in various forms,

☆ This work has received funding from the European Union and the Swiss State Secretariat for Education, Research and Innovation (SERI) under the grant agreement No 101192749 (XTRUST-6G). Disclaimer: Funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

* Corresponding author at: Department of Electrical and Computer Engineering, University of Western Macedonia, Campus ZEP, Kozani, Greece.
Email addresses: dimiasim3@ihu.gr; dasimopoulos@metamind.gr (D.C. Asimopoulos), pradoglou@uowm.gr; pradoglou@k3y.gr (P. Radoglou-Grammatikis), g.th.papadopoulos@hua.gr (G.T. Papadopoulos), psarigiannidis@uowm.gr (P. Sarigiannidis).
URL: https://metamind.gr/ (D.C. Asimopoulos).

https://doi.org/10.1016/j.cosrev.2026.100963
Received 15 January 2025; Received in revised form 9 March 2026; Accepted 9 March 2026
Available online 19 March 2026

including data poisoning, model inversion, evasion, adversarial attacks, and denial of service (DoS) attacks, all of which compromise the integrity, security, and reliability of AI applications. In addition, cyber attacks have become increasingly prevalent and sophisticated in recent years, targeting various sectors and causing significant disruption.

Several high-profile attacks occurred in 2023, including the Hot Topic attack in August, in which an unauthorized third party attempted to log into customer accounts using credentials obtained from an unknown source. Similarly, Prospect Medical Holdings faced a severe ransomware attack that took some of its hospitals and outpatient facilities offline, forcing medical staff to revert to manual operations. The broader cyber threat landscape shows that attackers are exploiting vulnerabilities at an alarming rate, as highlighted in the FortiGuard Labs 2023 Global Threat Landscape Report. These incidents follow a pattern observed in 2022, where politically motivated attacks were particularly prominent. For example, Russian-backed groups targeted critical infrastructure, such as the Finnish parliament and Ukraine's state-owned nuclear power company, with DoS and bot attacks. Other sectors, such as natural gas distributors in Greece and water companies in the United Kingdom, faced ransomware and data breaches. These examples underscore the growing threat of cyber attacks across industries, and highlight the need for robust cybersecurity measures to mitigate risk and protect critical services. As such, ensuring robust defenses is critical to securing AI technologies and maintaining the security of these systems. As AI continues to play an increasingly important role in key sectors, it is imperative to protect against the wide range of attack vectors targeting AI systems.

A key challenge is the vulnerability of systems to adversarial attacks. Adversarial attacks exploit the high-dimensional and complicated nature of AI models, especially deep neural networks (DNNs). The concept of adversarial attacks was introduced in [1] by Szegedy et al. This paper argued that marginally altered inputs can easily mislead neural networks. These attacks use perturbations to the input data that are often undetectable to human observers, but can significantly alter model performance. The perturbations can be as small as a few pixels in an image or minor changes in a text string, but they can cause the model to misidentify or misinterpret the input. Since then, several types of adversarial attacks have been identified, including but not limited to: evasion attacks, poisoning attacks, and model inversion. Adversarial attacks have far-reaching consequences for the reliability, security, and trustworthiness of AI systems. Adversarial attacks can be catastrophic in safety-critical applications such as self-driving cars, healthcare diagnostics, and financial fraud detection. In self-driving cars, for example, an adversarially perturbed stop sign could be misinterpreted as a yield sign, leading to unsafe driving decisions. Similarly, in healthcare, adversarial attacks on medical image classifiers can lead to inaccurate diagnoses, posing a serious threat to patient safety. In addition, adversarial attacks pose a significant barrier to the use of AI in cybersecurity. Attackers can use adversarial approaches to bypass intrusion detection systems, evade malware detection, and compromise authentication processes. Such attacks can have a significant economic impact, with potential losses in the millions. Overall, adversarial attacks represent a critical vulnerability in AI and ML systems, with far-reaching implications for various applications. As the use of AI technologies continues to grow, it is important to ensure their robustness and security against adversarial threats.

2. Background of adversarial attacks

Adversarial attacks are a type of cyber threat in which attackers use manipulated inputs to deceive ML models. These inputs, known as adversarial examples, are specifically designed to cause the ML model to make errors while appearing normal to human observers. This technique explores the vulnerabilities of ML algorithms, particularly in their interpretative layers.

Adversarial examples challenge the robustness of AI systems in real-world scenarios. They exploit the way algorithms process data, often taking advantage of the model's inherent biases or lack of ability to generalize from training data to new, unseen examples. This vulnerability is particularly alarming in critical systems such as autonomous vehicles, medical diagnostic tools, and security surveillance systems, where decision accuracy is paramount. There are three main types of adversarial attacks based on their knowledge access: white-box, black-box, and gray-box attacks. White-box attacks require full access to the model's architecture and parameters. In contrast, black-box attacks require no specific knowledge of the model's inner workings and use output data to iteratively modify inputs. Gray-box attacks fall in between, where the attacker has partial knowledge of the model.

Defending against adversarial attacks involves various strategies to increase the robustness of AI models. These include adversarial training, where the model is trained with a mixture of clean and adversarial examples to learn to ignore the perturbations, and defensive distillation, where a model is trained to produce softer outputs, making it harder for attackers to exploit precise model responses. Despite these defenses, the arms race between adversarial attacks and defensive measures continues to evolve. Researchers are actively exploring more sophisticated techniques for both attacking and defending AI systems. Understanding and mitigating these threats is critical to safely and ethically advancing AI technology and ensuring it remains resilient to evolving cyber threats. This understanding helps develop safer AI systems that can better withstand the complexities of operational environments.

2.1. Objectives

The primary objective of this survey is to analyze different types of adversarial attacks across different data modalities, with the goal of creating a comprehensive taxonomy based on the results of a selection of research papers. By systematically categorizing the characteristics and methodologies of adversarial attacks, this survey aims to identify commonalities and differences in how these attacks are implemented and defended in different domains such as image, audio, and text.

The taxonomy developed from this analysis will serve as a structured framework that outlines the scope and impact of adversarial threats on ML models. It will highlight specific vulnerabilities in different types of data and provide insights into how these attacks exploit inherent weaknesses in algorithms. This structured classification will help researchers and practitioners better understand the landscape of adversarial techniques and guide the development of more robust defensive mechanisms.

In addition, this survey aims to fill gaps in the current literature by integrating different approaches and results into a unified reference that can be used for future research efforts. The ultimate goal is to improve the security measures within AI systems, ensuring that they are better equipped to handle sophisticated adversarial inputs, thereby contributing to the advancement of safer, more reliable technology.

2.2. Contribution

This survey makes several important contributions to the field of ML security by providing an in-depth analysis of adversarial attacks across different data types. By systematically cataloging the diverse range of attack methodologies and their impact on various AI systems, it establishes a clear and structured taxonomy aligned with structured threat modeling principles and informed by the MITRE ATLAS framework. This taxonomy not only sheds light on specific vulnerabilities exploited by these attacks, but also serves as a critical resource for researchers. Furthermore, the survey makes the following contributions:

- **Comprehensive Analysis:** Provides a detailed examination of different types of adversarial attacks across multiple data types, enhancing the understanding of how these attacks work and how they affect different ML models.
- **Taxonomy Development:** Establishes a structured taxonomy that categorizes the various methods and characteristics of adversarial attacks, incorporating MITRE ATLAS-aligned

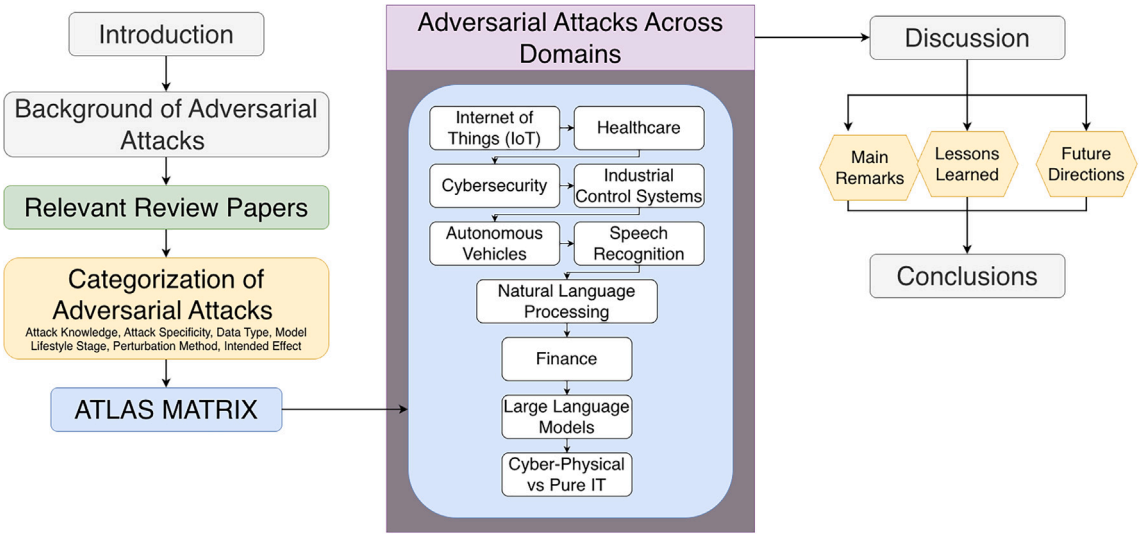

Fig. 1. Logical structure and content flow of the survey. The diagram illustrates how foundational concepts lead to a unified taxonomy, operationalized through the ATLAS matrix and applied across multiple domains, before synthesizing insights and future research directions.

adversarial tactics and techniques to support cross-domain comparison and standardized threat interpretation.

- **Vulnerability Identification:** Highlights specific vulnerabilities in various data modalities, providing critical insights into the weaknesses that adversarial attacks exploit, which can aid in the development of targeted defensive strategies.
- **Presentation and Categorization of Different Attack Types:** Analyzes and categorizes a variety of adversarial attack methodologies across different data types to improve understanding of their specific impact and the vulnerabilities they exploit in AI systems. This taxonomy serves as a critical resource for researchers and practitioners, fostering the development of innovative defensive strategies against these sophisticated threats.
- **Domain-Specific Taxonomy:** Provides a deep taxonomic view specific to important domains such as autonomous vehicles, healthcare, finance, IoT devices, speech recognition systems, natural language processing (NLP) tasks, and large language models (LLMs), with cross-domain mapping guided by MITRE ATLAS tactics and techniques. This taxonomy helps understand the unique threats and vulnerabilities each domain faces and helps formulate domain-appropriate defense strategies.

2.3. Structure

The remainder of this paper is organized as follows: Section 3 reviews similar survey papers to provide a foundation for the current work. Section 4 categorizes the different types of adversarial attacks and methodologies. Section 5 provides an overview of the ATLAS matrix, the categorization framework upon which this paper is based. Subsequently, Section 6 analyzes research in different domains and examines how adversarial attacks are uniquely managed within each domain. Section 7 provides a summary of the findings from the analysis, highlighting key lessons learned and outlining future research directions. Finally, Section 8 concludes the paper by summarizing the main contributions and findings (Fig. 1).

3. Relevant review papers

The study of adversarial attacks and defensive mechanisms in ML has resulted in a large body of literature, including several review papers that summarize and evaluate the current state of the art. This section provides an overview of similar review publications listed in Table 1 that have made significant contributions to the understanding of adversarial attacks, their consequences, and defense tactics. These studies serve as a foundation for this work and help identify gaps that must be addressed.

In Chakraborty et al. [6], the authors investigate adversarial examples that exploit deep learning systems by causing incorrect predictions while remaining indistinguishable to humans. The study classifies attacks according to different threat models, surveys existing defense strategies, and highlights their challenges. It also examines how various attack techniques interact with learning models and assesses system robustness across multiple adversarial scenarios.

In Xu et al. [2], the authors present a comprehensive review of adversarial attacks and defense mechanisms across multiple data domains. The paper covers well-known image-based attacks such as Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini–Wagner (CW), highlighting how small input perturbations can cause neural networks to misclassify. It also examines defense approaches including adversarial training and gradient masking, discussing both their strengths and limitations. Beyond images, the study explores adversarial threats to graph-structured data, text classifiers, and speech recognition systems, and emphasizes the ongoing challenges in building robust deep learning models.

In Chakraborty et al. [3], Chakraborty et al. provide a comprehensive taxonomy of adversarial attack strategies and corresponding defenses in ML, focusing primarily on deep learning but also addressing models such

Table 1
Overview of relevant review literature surveyed in this study, organized by title reference, authors, and year of publication.

Title	Authors	Year
[2]	XU, Han, et al.	2020
[3]	Chakraborty, Anirban, et al.	2021
[4]	Akhtar, Naveed, et al.	2021
[5]	Li,Yao, et al.	2022
[6]	Chakraborty, Anirban, et al.	2018
[7]	Akhtar, Naveed; Mian, Ajmal	2018
[8]	Kong, Zixiao, et al.	2021
[9]	Long, Teng, et al.	2022
[10]	Khamaiseh, Samer Y., et al.	2022
[11]	Shayegani, Erfan, et al.	2023
[12]	Yan, Senming, et al.	2022
[13]	Wang, Donghua, et al.	2022
[14]	He, Ke; Kim, et al.	2023
[15]	Z Zhang, et al.	2024

as support vector machines. Adversarial threats are categorized into evasion attacks, where inputs are manipulated at inference time, poisoning attacks that corrupt the training process, and exploratory attacks aimed at understanding a model's behavior. The paper clearly distinguishes between white-box and black-box attack scenarios and reviews several defense mechanisms, including adversarial training, distillation, and feature squeezing. It concludes that although numerous countermeasures exist, none are fully effective, leaving robust defense against adversarial attacks an open research problem.

Furthermore, a similar research is conducted specifically in the computer vision area. In Akhtar et al. [4], the authors present a thorough review of state-of-the-art adversarial attacks and defenses in computer vision, emphasizing the security vulnerabilities of deep learning models, especially in safety-critical applications like autonomous driving and face recognition. The paper surveys early adversarial methods such as Limited Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS), FGSM, Basic Iterative Method (BIM), DeepFool, and Carlini–Wagner, explaining their principles and impact on model robustness. It also discusses how these attack techniques have evolved over time, offering valuable insights into the growing sophistication of adversarial threats.

Also in the field of computer vision, in [7], Akhtar et al. provide a comprehensive overview of adversarial attacks on deep learning models and the corresponding defense mechanisms. They classify attacks into evasion attacks, which manipulate inputs at inference time, poisoning attacks that corrupt training data, and exploratory attacks that probe model behavior. The paper reviews adversarial techniques targeting both support vector machines and neural networks, and explores multiple defenses such as adversarial training, input compression, data randomization and augmentation, defensive distillation, foveation-based methods, and GAN-based approaches, concluding that no single method is sufficient and continued research is required.

Deep learning has been widely adopted in fields such as computer vision, NLP, and data mining, but its vulnerability to adversarial attacks limits its use in security-critical applications. In the context of computer vision, Long et al. in [9] survey classic and recent adversarial attacks using a structured taxonomy, analyze research trends through a large-scale knowledge graph built from over 5900 Scopus articles, and identify future research directions based on keyword trend analysis.

In addition, Wang has also made a significant contribution to the field of computer vision by presenting a survey paper on physical adversarial attacks in computer vision [13]. This work provides a comprehensive review and taxonomy of physical adversarial attacks on deep learning–based vision systems, emphasizing their robustness in real-world settings. It details various attack strategies and methodologies, discusses both digital and physical adversarial examples with a stronger focus on physical attacks, and explores different deployment environments. The survey also outlines future research directions, making it a valuable reference for understanding current adversarial threats and developing stronger defenses.

Li et al. [5] review the vulnerabilities of ML systems, particularly DNNs, to adversarial examples. The paper surveys methods for generating adversarial attacks and the defenses designed to counter them, classifying attacks into textual-only, multimodal, and complex system attacks. It examines techniques and defenses for each class using examples from images, text, and malicious code, and proposes a common framework to engage the statistical community. The study also highlights the need for strong defenses in security-critical applications such as healthcare and autonomous vehicles.

Similar work is also presented by Khamaiseh in [10]. In this paper, the authors survey recent advances in adversarial attacks and defense strategies for DNN-based image classification. They highlight the vulnerability of DNNs to small input perturbations that can cause misclassification and introduce key concepts such as adversarial examples and perturbations. The proposed taxonomy includes white-box, black-box, and gray-box attacks. The paper reviews attack methods such as FGSM, BIM, DeepFool, C&W, and Universal Adversarial Perturbations, along with defense strategies including adversarial training, input transformations, and robust optimization. Overall, it provides insights into the effectiveness, limitations, and future directions of adversarial deep learning.

In addition, Yan et al. [12] provide an analysis of adversarial attacks on malware classification systems. It describes various adversarial strategies, categorized as white-box and black-box attacks, that aim to evade malware classifiers through manipulations that fool these systems without affecting the functionality of the malware. These strategies include gradient-based methods, the use of GANs, and other optimization techniques. The paper also explores the robust defenses that can be employed to counter these attacks, such as adversarial training, model distillation, and the integration of novel techniques such as random feature nullification, which increase the resilience of classifiers against these sophisticated attacks. This comprehensive review not only discusses the current methodologies and their implications, but also points out future research directions to further strengthen the security of malware detection systems.

Another relevant and significant paper that provides insights into adversarial attacks in the era of AI is presented by Kong et al. [8]. This paper attempts to interrogate the very important and critical growing concern of adversarial attacks in the AI domain with key thrusts toward the need for robust security. The paper is a detailed review of various methods used in adversarial attacks, including images, text, and malicious code. It classifies these attacks into text-only, multi-modal, and complex system attacks. Each is elaborated with the typical algorithms and defense techniques. This provides a very structured framework for understanding the AI security landscape, and outlines the associated threats and defenses. The discussion continues with open questions and comparisons to other literature, positioning this work as a solid resource for those entering or engaged in adversarial attack research.

As LLM continue to grow and become more widely used, especially with the introduction of ChatGPT, vulnerabilities in this area are increasing. In [11], Shayegani et al. attempt to provide an overview of vulnerabilities in LLM models revealed by adversarial attacks. More specifically, this paper highlights the rapid development of LLM architectures and their embedding into complex systems, which increase security risks. Despite the security precautions put in place through instruction tuning and reinforcement learning, LLMs remain vulnerable to attacks such as the well-known jailbreak incidents on models such as ChatGPT and Bard. The adversarial threats are classified, and the common difficulties as they relate to text-only, multi-modal, and system-specific attacks with corresponding defenses are detailed, to provide a deeper understanding of adversarial threats and further enhance future security.

An exemplary work in this area is [14], which provides a detailed analysis of adversarial attacks targeting network intrusion detection systems (NIDS). The authors examine adversarial ML from the perspective of NIDS and highlight how deep learning–based detection models can be misled by subtle perturbations that result in incorrect classifications. The paper proposes a structured taxonomy that includes evaluation datasets for both general and IoT networks, feature extraction methods, feature reduction techniques, and different detection paradigms. It also reviews common deep learning architectures used in NIDS, such as DNNs, Convolutional Neural Networks (CNNs), Recurrent Neural Networks(RNNs), and autoencoder-based models. Furthermore, the survey discusses white-box and black-box adversarial attacks along with relevant defense strategies, stressing the importance of developing practical solutions suited to the unique constraints of NIDS environments.

A domain-specific perspective on adversarial machine learning is presented by Zhang et al. in [15], where the authors provide a comprehensive survey of vulnerabilities associated with machine learning approaches applied in IoT-based smart grid environments. Unlike general adversarial ML surveys, this work emphasizes the cyber-physical characteristics of power systems, highlighting how adversarial attacks

must respect physical constraints, system dynamics, and operational limitations inherent to energy infrastructures. The study reviews adversarial attacks across different stages of power system operation, including generation, transmission, distribution, and consumption scenarios, and analyzes both model-centric vulnerabilities and system-level risks. Furthermore, it discusses defense strategies tailored to smart grid environments, such as resilient learning models and power-system-aware protection mechanisms. By focusing on the intersection of adversarial ML and cyber-physical energy systems, this work provides valuable insights into domain-specific challenges that differ from traditional adversarial settings in computer vision or natural language processing.

Although the aforementioned review papers provide valuable insights into adversarial attacks and defenses across various domains, they exhibit several limitations that motivate the need for the present work. Many existing surveys focus either on specific application areas, such as computer vision, malware detection, smart grids, or large language models, or emphasize particular attack methodologies without establishing a unified analytical framework. For example, domain-specific studies such as [15] analyze adversarial threats within cyber-physical smart grid environments by considering physical constraints and system-level dynamics, while other surveys primarily categorize attacks based on model architecture or data modality. In contrast, our work introduces a unified, multi-dimensional threat modeling perspective that systematically organizes adversarial attacks according to adversary capabilities, attack objectives, lifecycle stages, perturbation strategies, and intended operational effects. Furthermore, by aligning adversarial analysis with structured threat intelligence frameworks such as MITRE ATLAS, this survey moves beyond traditional taxonomy-driven reviews and provides a cross-domain analytical foundation that enables consistent comparison across heterogeneous application domains. This holistic perspective highlights recurring adversarial patterns, bridges gaps between domain-specific studies, and supports the development of more generalizable and operationally relevant defense strategies.

4. Categorization of adversarial attacks

Adversarial attacks pose a significant threat to the integrity, reliability, and trustworthiness of ML systems. Given the diversity of attack strategies, threat assumptions, and operational contexts, isolated or parallel categorization schemes often obscure the intrinsic relationships between adversarial behaviors. To address this limitation, this section adopts a *unified adversarial threat modeling perspective*, in which adversarial attacks are described through multiple complementary dimensions rather than independent taxonomies.

Rather than introducing independent or parallel taxonomies, adversarial behavior is modeled through five complementary dimensions that collectively capture attacker assumptions, operational mechanisms, and intended outcomes. Building upon system-level modeling approaches commonly used in cybersecurity and cyber-resilience literature, the proposed framework organizes adversarial threats according to the **adversary model**, which characterizes attacker knowledge and capabilities; the **attack model**, which defines objectives and specificity; the **attack stage**, which introduces the temporal dimension within the ML lifecycle; the **perturbation strategy**, describing how adversarial inputs are generated; and the **intended effect**, capturing the operational impact on model behavior and decision outcomes. Together, these dimensions provide a unified analytical structure that supports consistent interpretation and cross-domain comparison of adversarial threats while avoiding fragmented classifications.

Fig. 2 illustrates the operational relationships between training-time poisoning, inference-time evasion, and iterative model probing within this unified framework.

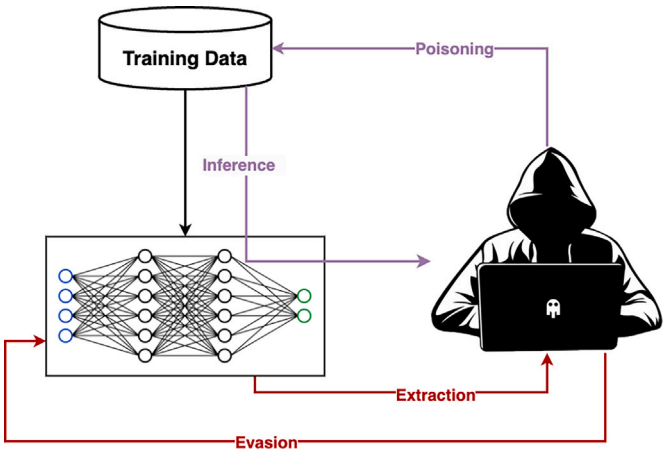

Fig. 2. Methodological overview of adversarial attack vectors, illustrating training-time poisoning, inference-time evasion, and iterative model extraction through querying.

4.1. Adversary model: attacker knowledge

Within the adversary model dimension, attacker knowledge characterizes the extent of information available about the target system. **White-box attacks** assume full access to the model architecture, parameters, and training procedures, enabling precise gradient-based optimization of adversarial perturbations. In contrast, **black-box attacks** operate without internal model knowledge and rely on input–output queries, surrogate modeling, or decision-boundary approximation techniques. The level of attacker knowledge directly influences both attack feasibility and perturbation efficiency, and reflects realistic deployment constraints across domains.

4.2. Attack model: objective and specificity

The attack model captures the adversary's intended outcome and level of control over model behavior. **Targeted attacks** aim to induce a specific attacker-chosen misclassification, requiring fine-grained manipulation of the decision boundary. **Untargeted attacks**, by contrast, seek to cause any incorrect prediction, exposing general robustness weaknesses while often requiring fewer optimization constraints. These objectives represent distinct threat severities and operational goals.

4.3. Attack stage in the model lifecycle

The attack stage introduces the temporal dimension of adversarial behavior. **Poisoning attacks** occur during the training phase, where malicious samples or labels are injected to compromise model behavior after deployment. **Evasion attacks** operate at inference time, crafting adversarial inputs that mislead deployed models without modifying internal parameters. This lifecycle-centric perspective is essential for aligning threat analysis with appropriate defense mechanisms.

4.4. Perturbation strategy

Adversarial perturbation strategies describe how malicious inputs are generated. **Gradient-based methods** leverage exact or estimated gradients and are typically associated with white-box settings. **Score-based methods** exploit confidence scores or probability outputs to guide perturbation generation under partial knowledge. **Decision-based methods** rely solely on final model decisions and iterative querying, making them suitable for strict black-box environments where gradient information is unavailable.

Table 2
Unified adversarial threat model for machine learning systems.

Model Dimension	Attribute	Description
Adversary Model	Attacker Knowledge	White-box, black-box, or partial knowledge of the target model
	Capabilities	Query access, computational resources, surrogate models
Attack Model	Attack Objective	Targeted or untargeted misclassification
	Intended Effect	Misclassification, confidence reduction, source-to-target mapping
	Perturbation Method	Gradient-based, score-based, decision-based
Attack Stage	Training Phase	Data poisoning, backdoor injection
	Inference Phase	Evasion and model extraction attacks

Recent studies have also explored adversarial perturbations in the frequency domain, revealing that deep neural networks may exhibit sensitivity to specific spectral components. Instead of directly manipulating pixels in the spatial domain, these approaches exploit the spectral characteristics of images to generate more transferable and effective adversarial examples. For instance, Qian et al. proposed LEA2, a lightweight ensemble adversarial attack that identifies non-overlapping vulnerable frequency regions across multiple models and leverages them to craft adversarial perturbations with improved cross-model transferability [16]. Similarly, mixed-frequency input transformations have been introduced to enhance attack effectiveness by incorporating high-frequency components from different images during gradient computation, resulting in more stable optimization directions and improved transferability [17]. Furthermore, recent work has investigated multimodal adversarial attacks that combine frequency-domain enhancement with fine-grained cross-modal guidance, enabling the generation of stronger adversarial examples by jointly exploiting spectral vulnerabilities and multimodal correlations [18]. These findings indicate that frequency-aware perturbation strategies represent an emerging direction for adversarial attack design, particularly for improving transferability in realistic black-box scenarios.

4.5. Intended effect

Finally, attacks can be distinguished according to their operational impact on model behavior. **Misclassification attacks** directly induce incorrect predictions, while **source-to-target attacks** enforce specific label mappings. **Confidence-reduction attacks** degrade prediction certainty without necessarily altering the predicted class, undermining trust in model outputs and decision reliability.

By organizing adversarial attacks through this unified multidimensional threat modeling framework, the taxonomy clarifies the intrinsic relationships between attacker capabilities, attack objectives, perturbation mechanisms, lifecycle stages, and operational outcomes. Such a structured representation provides a principled foundation for systematic defense design, robustness evaluation, and cross-domain comparison, while also facilitating alignment with structured threat frameworks such as MITRE ATLAS.

Table 2 consolidates the proposed categorization under a unified threat modeling framework. Rather than enumerating multiple parallel classifications, adversarial behavior is organized across five complementary dimensions, enabling consistent analysis across heterogeneous application domains and providing a coherent bridge to the domain-specific analyses that follow.

5. ATLAS matrix

Building upon the unified adversarial threat modeling framework introduced in the previous section, the MITRE ATLAS (Adversarial Threat Landscape for AI Systems) is adopted as an operational reference model to contextualize adversarial behavior across domains. Rather than serving as a descriptive taxonomy, ATLAS provides a threat-centric analytical structure that aligns adversary goals, operational stages, and attack mechanisms within the lifecycle of AI systems.

MITRE ATLAS organizes adversarial activity through a matrix of adversary *tactics* (high-level objectives), *techniques* (concrete attack mechanisms), and *procedures* (real-world execution patterns). Within this matrix, adversarial behavior is structured across tactic-oriented stages including reconnaissance, initial access, model access, execution, persistence, defense evasion, discovery, collection, and impact. Each tactic is instantiated through specific techniques such as adversarial example generation, model probing, prompt manipulation, data poisoning, and model extraction. This structured representation enables consistent interpretation of adversarial strategies under realistic threat assumptions and facilitates mapping between conceptual adversarial models and practical attack implementations.

In this work, ATLAS is used as an operational backbone that bridges the multidimensional threat modeling taxonomy introduced previously with domain-specific adversarial studies. Rather than analyzing domains independently, adversarial attacks are interpreted through ATLAS tactic–technique relationships, enabling systematic cross-domain comparison and highlighting recurring adversarial patterns across heterogeneous application environments. Collaboration and community engagement are central to the evolution of MITRE ATLAS, incorporating contributions from academic, industrial, and governmental stakeholders. The ATLAS Navigator tool further supports dynamic visualization of tactic prevalence and technique relationships, assisting analysts in identifying dominant threat patterns and prioritizing defensive strategies.

To operationalize this framework, the following subsections analyze representative adversarial studies under different attacker access assumptions—white-box, black-box, and hybrid scenarios—as practical manifestations of ATLAS threat behaviors. This organization reflects how core ATLAS tactics such as *Defense Evasion*, *Model Access*, *Model Extraction*, and *Impact* manifest under varying levels of attacker knowledge and system interaction.

5.1. ATLAS-aligned white-box threat scenarios

Within the MITRE ATLAS framework, white-box adversarial scenarios primarily align with tactics such as *Defense Evasion* and *AI Model Manipulation*, where direct access to model parameters enables gradient-based optimization and targeted perturbation design during inference or training stages. These tactics are commonly instantiated through techniques including gradient-based adversarial example generation, targeted perturbation crafting, and training-time manipulation.

In Ye et al. [19], a new adversarial attack method targeting neural networks demonstrates higher success rates compared to conventional first-order attacks while maintaining computational efficiency. Evaluations on ResNet18 highlight how full model access facilitates exploitation of intrinsic network properties, illustrating ATLAS-aligned evasion techniques. Similarly, the CosPGD attack proposed in Agnihotri et al. [20] leverages cosine similarity optimization to manipulate prediction distributions across tasks such as semantic segmentation and optical flow estimation. These results illustrate how task-aware optimization strategies operationalize ATLAS evasion tactics. Furthermore, studies such as Ayub et al. [21] and Apruzzese et al. [22] demonstrate how gradient-based attacks like JSMA exploit vulnerabilities in ML-based intrusion detection systems, exemplifying inference-time defense evasion under white-box conditions.

5.2. ATLAS-aligned black-box threat scenarios

Black-box adversarial scenarios typically correspond to ATLAS tactics including *Model Probing*, *Model Extraction*, and *Inference-time Evasion*, where adversaries iteratively query deployed systems to approximate decision boundaries under restricted knowledge assumptions.

Table 3
Mapping of domain-specific adversarial attacks to MITRE ATLAS tactics.

Domain	Representative attacks	ATLAS tactics
Network Security (NIDS)	JSMA, GAN-based evasion	Defense Evasion
Computer Vision	PGD variants, CMA-ES	Model Manipulation, Extraction
Malware Detection	Confidence manipulation	Model Probing, Defense Evasion
Natural Language Processing	TextAttack	Input Manipulation, Transferability
Steganography	EAST	Targeted Evasion

The work in Wu et al. [23] demonstrates the manipulation of confidence scores to bypass malware detection models, reflecting model probing techniques aligned with ATLAS threat behaviors. Evolutionary optimization strategies presented in Qiu et al. [24] further illustrate how adversaries exploit limited feedback channels to craft adversarial inputs across deep neural network architectures.

Attribute-based adversarial methods introduced in Wei et al. [25] highlight how perceptual feature manipulation achieves high fooling rates while maintaining realism, emphasizing the adaptability of ATLAS-aligned evasion techniques. Additional studies such as Apruzzese et al. [26] and Alshahrani et al. [27] demonstrate how adversaries exploit feedback mechanisms and generative models to perform evasion and poisoning attacks within network security environments.

5.3. ATLAS-aligned hybrid threat scenarios and transferability

Hybrid adversarial scenarios capture situations where partial knowledge, transferability, or adaptive querying enable attackers to combine white-box and black-box capabilities. Within ATLAS, these behaviors frequently correspond to tactics involving model extraction and inference-time evasion supported by surrogate modeling or adaptive optimization.

Platforms such as TextAttack [28] demonstrate how adversarial NLP workflows operationalize transferability through systematic benchmarking of multiple attack strategies. The EAST framework [29] illustrates targeted evasion techniques that embed hidden information within images, reflecting ATLAS-aligned attack patterns. Finally, the surrogate-assisted hybrid attack framework proposed in Asimopoulos et al. [30] exemplifies hybrid threat scenarios where surrogate models approximate inaccessible targets, enabling gradient-based attacks under realistic constraints. Table 3 summarizes representative adversarial attacks across domains and their alignment with MITRE ATLAS tactics..

Rather than treating domains as independent analytical units, the subsequent sections interpret adversarial attacks across application areas through the ATLAS threat perspective established above. This threat-centric organization enables consistent comparison across heterogeneous domains by aligning diverse adversarial behaviors under shared tactics and operational objectives.

6. Adversarial attacks in different domains

Based on the MITRE ATLAS matrix taxonomy, we conducted a comprehensive literature search to examine the evolution of adversarial research across multiple application domains. In total, 237 relevant publications were identified, comprising 153 journal articles and 84 conference papers as illustrated in Fig. 3. This distribution reflects the dual nature of the field: journal publications typically provide mature and in-depth analyses, whereas conference proceedings capture rapid advances in emerging attack mechanisms and evaluation practices.

The temporal distribution reveals a clear growth trend over the past decade (Fig. 4). Early studies between 2014 and 2016 were limited and mainly focused on foundational adversarial phenomena in neural networks. From 2017 onward, research activity increased substantially, with a pronounced surge between 2020 and 2023, during which annual

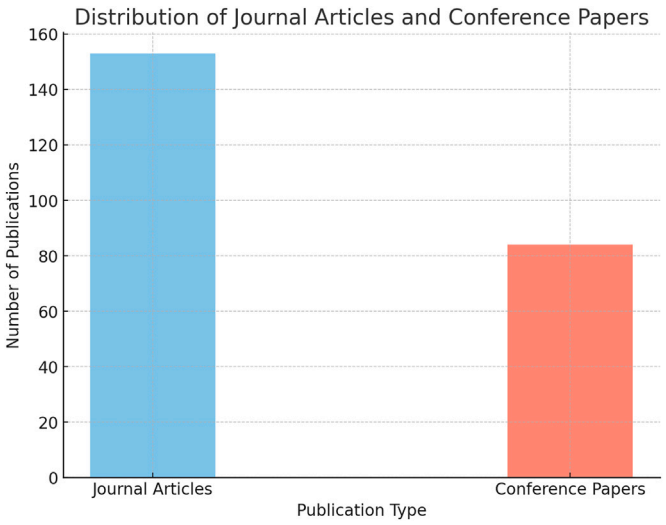

Fig. 3. Distribution of publications by type, comparing journal articles and conference papers in the analyzed dataset.

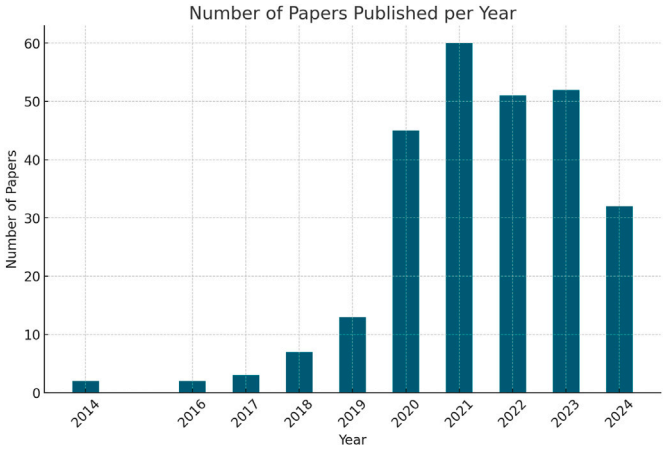

Fig. 4. Annual publication trends, showing the number of papers published per year over the studied period.

publications exceeded 40 and peaked in 2021. This growth coincides with the widespread deployment of ML in safety- and security-critical settings, motivating deeper investigation into adversarial vulnerabilities and robustness.

To avoid treating domains as isolated case studies, the remainder of this section adopts an *ATLAS-driven, threat-centric* interpretation of domain-specific literature. Concretely, for each domain we (i) identify the dominant **ATLAS tactics** reflected in real deployments (e.g., *Defense Evasion*, *Model Access/Probing*, *Extraction*, *Persistence/Poisoning*, and *Impact*), (ii) highlight the **techniques** that instantiate these tactics (e.g., gradient-based perturbation crafting, query-based boundary estimation, model extraction via surrogate learning, backdoor/label poisoning, and generative traffic synthesis), and (iii) discuss domain constraints (resource limitations, observability, feedback channels, and cyber-physical coupling) that shape feasible attacker procedures and defenses. This framing enables consistent cross-domain comparison while preserving domain-specific threat surfaces and operational realities.

6.1. Internet of things (IoT)

The expanding IoT ecosystem introduces heterogeneous ML deployments at the edge, in gateways, and in cloud-assisted pipelines, often under constrained resources, partial observability, and limited

feedback interfaces. From an ATLAS perspective, IoT adversarial research predominantly maps to (i) *Defense Evasion* against intrusion detection and device identification, (ii) *Model Access/Probing* and *Extraction* under black-box or API-limited settings, (iii) *Persistence* through training-time manipulation (e.g., poisoning in collaborative/federated learning), and (iv) *Impact* on detection reliability and safety-critical decision-making. Across these tactics, recurring techniques include gradient-based adversarial example crafting (FGSM/PGD/JSMA), query-based decision-boundary approximation, surrogate-assisted transferability, and GAN-driven synthesis of adversarial traffic.

ATLAS: defense evasion in IoT detection and identification. A dominant theme in IoT is inference-time evasion of ML-based detection and identification pipelines. For deep learning–based network intrusion detection in IoT, Qiu et al. [31] studied black-box evasion with minimal packet modification, while Zhou et al. [32] proposed a hierarchical black-box framework against GNN-based IoT IDS using a shadow model and saliency-guided perturbations. Papadopoulos et al. [33] and Du et al. [34] demonstrated that label poisoning and FGSM-style adversarial inputs can degrade performance on Bot-IoT, reinforcing that both inference-time and training-time manipulations can facilitate evasion. Additional work targets device identification: Bao et al. [35] analyzed targeted and untargeted attacks against CNN-based IoT device identification, showing accuracy degradation with increasing perturbation budgets.

IoT also includes vision-based and sensing-driven subsystems (e.g., UAV operations). Tian et al. [36] presented targeted and untargeted attacks against regression models used in UAV systems, demonstrating imperceptible adversarial images that can threaten navigation and control. Raja et al. [37] further examined adversarial attacks on AI-based UAV infrastructure inspection, showing that constraint-aware adversarial generation can cause missed high-risk areas, while adversarial training reduces vulnerabilities.

Finally, IoT privacy protection can be framed as adversarial manipulation to mitigate leakage or misuse: Ding et al. [38] proposed selective gradient sign iterative methods to generate highly similar adversarial examples that protect photo privacy on IoT-connected mobile devices while preserving usability.

ATLAS: model access/Probing and extraction under edge constraints. A second major axis is adversarial behavior enabled by limited feedback channels typical in edge deployments. Works such as Qiu et al. [31] and Zhou et al. [32] explicitly leverage black-box interactions and surrogate/shadow modeling to approximate decision boundaries. In federated/vertical settings, Yang et al. [39] proposed a black-box feature inference attack on vertical federated learning using zeroth-order gradient estimation, highlighting that restricted feedback can still be exploited to infer sensitive properties—an important operational risk in multi-party IoT learning pipelines.

ATLAS: persistence via training-time manipulation (Poisoning) in collaborative/Federated IoT. IoT deployments frequently rely on collaborative or federated learning, where training-time manipulation enables the persistence of malicious influence. Chen et al. [40] addressed the security and privacy of collaborative deep learning in IoT, focusing on defenses against GAN-based inference/data leakage and emphasizing participant isolation and layer-wise protections. Ferrag et al. [41] reviewed federated deep learning mechanisms for IoT cybersecurity and compared federated vs. centralized learning for intrusion and malware detection. Li et al. [42] proposed a federated framework for software-defined IIoT resilient to adaptive poisoning via grouping, malicious behavior detection, and privacy-preserving exchange. Poisoning dynamics are also studied more broadly: Dunn et al. [43] evaluated label poisoning across multiple ML models in IoT datasets (ToN_IoT, UNSW-NB15), showing severe degradation at higher poisoning rates. In cognitive radio IoT, Liu et al. [44] combined data poisoning with a jamming waveform to reduce sensing accuracy, demonstrating that cyber-physical signal channels can amplify the effectiveness of training-time and runtime manipulations.

ATLAS: defensive countermeasures and robustness-oriented techniques. A substantial subset of IoT work proposes deployable defenses, often under resource constraints. Qian et al. [45] proposed EI-MTD, a moving-target defense for edge intelligence that mitigates black-box transferability by dynamically scheduling lightweight robust models. Fu et al. [46] evaluated CNN/LSTM/GRU-based IoT IDS under FGSM, showing that adversarial or hybrid training can substantially improve robustness depending on the architecture. Jiang et al. [47] introduced FGMD, a fusion-based adversarial defense that integrates adversarial sample generation during training and improves robustness while preserving performance. Rashid et al. [48] reported large accuracy degradation under adversarial attack in smart-city IoT IDS and showed that adversarial retraining can restore high detection accuracy, while Anthi et al. [49] similarly demonstrated robustness gains from adversarial training in smart-home DoS detection.

Several works combine generative modeling with robustness and detection improvements. Nie et al. [50] proposed a GAN-based intrusion detection method for Social IoT edge computing, while Wu et al. [51] combined fuzzy rough set–based feature selection with CNN feature extraction and GAN-enhanced detection. Idrissi et al. [52] proposed an unsupervised GAN-based host IDS for IoT devices and reported promising performance on MQTTset. Hassan et al. [53] proposed a robust deep learning security framework for IIoT that improves the realism of adversarial/noisy samples via cooperative generation. Qian et al. [54] studied data insufficiency and poisoning risks in IIoT training by combining GAN-based generation with continual learning. Ferrag et al. [55] developed a GAN-based dual-detector framework for 6G-IoT leveraging adversarial training, while Benaddi et al. [56] proposed a conditional GAN–enhanced CNN–LSTM IDS addressing data imbalance. Nayak et al. [57] introduced a GAN–SVM framework with adversarial training for routing attack detection in RPL-based IIoT.

Beyond IDS, IoT security includes malware and code-integrity defenses. Yumlembam et al. [58] proposed a GNN-based Android malware detector and VGAE-MalGAN to generate adversarial graphs, showing that adversarial retraining improves robustness. Chaganti et al. [59] presented a Bi-GRU–CNN model for IoT malware detection over ELF byte sequences with strong generalization. Shrivastava et al. [60] proposed an ROP-based tamper-resistance mechanism for IoT devices with modest overhead.

Finally, trust and secure communication mechanisms complement robustness against adversarial manipulation. Hao et al. [61] proposed a blockchain-enabled zero-trust information-sharing protocol combining cryptography with GAN-assisted filtering of fabricated data, while Hu et al. [62] introduced a lightweight federated learning scheme for IIoT integrating multifactor authentication with low overhead. At the systems level, Liu et al. [63] analyzed adversarial attacks on DRL-based IIoT controllers across training and deployment, illustrating that control policies can be vulnerable even when model accuracy is high.

Table 4 consolidates the reviewed IoT literature under a unified taxonomy aligned with our attack categorization scheme, while also reflecting ATLAS-relevant behaviors. The surveyed works cluster into functional themes (collaborative/federated learning security, IDS and authentication, privacy/integrity, IIoT and smart-city deployments, GAN/hybrid security models, and robustness of detection/identification). Interpreted through ATLAS, IoT studies are dominated by *Defense Evasion* and *Model Probing/Access* in realistic black-box settings, whereas *Persistence/Poisoning* emerges primarily in federated/collaborative pipelines. This mapping explains why black-box threat models prevail in IoT—attackers often interact with deployed detectors through constrained APIs and feedback channels—while defenses increasingly shift toward adversarial (re)training, fusion-based detection, and privacy-preserving mechanisms suitable for resource-constrained environments.

Table 4
Systematic classification of adversarial attack methodologies in IoT applications aligned with MITRE ATLAS tactics and techniques.

Title	Year	White-box	Black-box	Other	Targeted	Untargeted	Defense	ATLAS Tactic(s)	ATLAS Technique Anchor
Collaborative Learning and Federated Systems									
[40]	2020		x		x		x	Persistence, Defense Evasion	Privacy inference / GAN leakage
[41]	2021		x	x			x	Persistence	Federated poisoning
[45]	2022		x	x			x	Defense Evasion	Moving-target defense
Device Authentication and Network Intrusion Detection									
[42]	2022		x	x			x	Persistence	FL poisoning mitigation
[31]	2020		x		x			Defense Evasion	Adversarial example crafting
[50]	2021	x	x				x	Defense Evasion	GAN-based IDS
[32]	2021	x		x	x			Model Access, Extraction	Shadow model probing
[33]	2021		x		x			Defense Evasion	FGSM evasion
Data Privacy and Integrity									
[38]	2020			x			x	Defense Evasion	Privacy adversarial perturbation
[34]	2020	x			x			Defense Evasion	FGSM IDS attack
[46]	2021	x			x	x		Defense Evasion	Adversarial training
[36]	2021		x		x	x		Impact	UAV perception manipulation
[51]	2021			x	x		x	Defense Evasion	GAN-enhanced detection
[35]	2021	x			x	x		Defense Evasion	Device ID perturbation
[52]	2022	x	x		x		x	Defense Evasion	GAN anomaly detection
[44]	2022		x	x				Persistence, Impact	Poisoning + jamming
[47]	2022			x			x	Defense Evasion	Fusion adversarial defense
Industrial IoT and Smart City IoT Systems									
[53]	2020	x		x	x			Defense Evasion	Robust adversarial training
[48]	2022			x	x		x	Defense Evasion	Adversarial retraining
[49]	2021		x		x	x		Defense Evasion	Packet feature perturbation
[43]	2020					x		Persistence	Label poisoning
[64]	2021	x			x		x	Defense Evasion	Smart meter attack
[63]	2021			x	x			Impact	DRL controller manipulation
Innovative GANs and Hybrid Models for IoT Security									
[54]	2022		x	x	x		x	Persistence	GAN data generation
[55]	2023	x			x	x		Defense Evasion	GAN adversarial training
[56]	2022		x		x	x		Defense Evasion	Conditional GAN IDS
[60]	2022		x		x			Impact	Code tampering detection
[57]	2021		x	x	x			Defense Evasion	GAN-SVM intrusion detection
[59]	2022	x	x		x			Defense Evasion	Malware detection adversarial
[61]	2021	x		x				Defense Evasion	Blockchain trust filtering
Securing Device Identification and Intrusion Detection Robustness									
[65]	2023		x		x		x	Defense Evasion	Hybrid SOCNN IDS
[66]	2023		x			x	x	Persistence	Edge FL threats
[58]	2022	x			x		x	Extraction	Adversarial graph generation
[39]	2023		x			x	x	Extraction	Feature inference attack
[37]	2022	x	x		x	x		Impact	UAV adversarial perception
[62]	2023	x			x		x	Persistence	Secure FL authentication

6.2. Healthcare

Healthcare is among the most safety-critical domains for machine learning deployment, where adversarial manipulation can directly affect diagnostic accuracy, treatment planning, and patient outcomes. In contrast to conventional ML applications, healthcare AI operates across heterogeneous data modalities (medical imaging, physiological time-series, EHR/clinical records) and socio-technical workflows (clinical decision support, smart healthcare infrastructures). As illustrated in Fig. 5, adversarial threats may emerge across multiple stages of the ML lifecycle and can be interpreted through MITRE ATLAS tactics such as inference-time defense evasion, training-time manipulation (poisoning), model probing/extraction, and high-impact operational disruption.

General security, privacy, and e-healthcare infrastructure. Early work emphasizes that healthcare adversarial risk is inseparable from broader e-health security and privacy considerations. Zeadally et al. [67] surveyed security attacks and mitigation practices in e-health systems and highlighted architectural weaknesses that can amplify ML-related threats. Privacy-preserving access control remains a primary requirement in EHR-centric deployments, as addressed by Kanwal et al. [68]. In distributed settings, integrity and trust of model updates become critical: Kalapaaking et al. [69] proposed blockchain-enabled federated learning with secure multiparty computation to verify updates and defend against poisoning, while Siniosoglou et al. [70] introduced a layered federated architecture for Medical Cyber-Physical Systems (MCPS), improving robustness against adversarial behaviors in collaborative environments. Broader security-oriented reviews further underline the high clinical and operational stakes of adversarial attacks in healthcare AI [71].

Medical imaging: attacks on diagnostic DNN pipelines. Medical imaging remains the dominant adversarial attack surface due to the widespread use of deep models in radiology, pathology, dermatology, and ophthalmology. Multiple surveys consolidate evidence that diagnostic DNNs are vulnerable under both white-box and black-box assumptions [72–76]. Attack studies show that both global and localized perturbations can induce clinically misleading outputs. For example, universal and minimal-perturbation attacks demonstrate that small or even single-pixel modifications can trigger misclassification [77–79]. Compound perturbation settings (e.g., simultaneous FGSM/PGD across full images and local regions) further amplify degradation and reflect realistic multi-step evasion behaviors [80]. During COVID-19, adversarial examples were shown to significantly degrade the reliability of DL-based diagnostic models when defenses were absent [81].

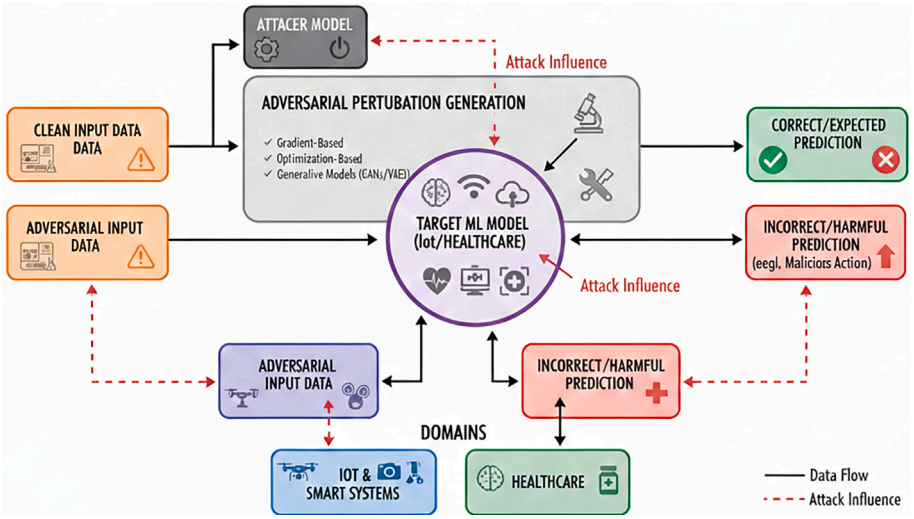

Fig. 5. Conceptual overview of an adversarial attack pipeline in IoT and healthcare ML systems, illustrating how an attacker model generates perturbations (gradient-, optimization-, or generative-based) that transform clean inputs into adversarial inputs, biasing the target ML model toward incorrect/harmful predictions (vs. correct/expected outputs); solid arrows denote data flow and dashed arrows denote attack influence.

Additional modality-specific threats include digital watermarking as an adversarial manipulation channel, which can lead to substantial performance drops across MRI, CT, and X-ray pipelines [82]. These results collectively align with ATLAS defense-evasion patterns, where inference-time input manipulation undermines downstream clinical decisions.

Ophthalmology and retinal imaging. Retinal analysis pipelines provide a representative case where small perturbations can compromise diabetic retinopathy screening systems. Lal et al. [83] proposed defenses combining adversarial training with feature fusion against speckle-noise attacks, while Daanouni et al. [84] demonstrated MobileNet vulnerability under FGSM even for very small perturbation budgets. Additional evaluations on retinal datasets further confirm that robustness must be explicitly engineered rather than assumed [85].

Physiological signals and time-series clinical models. Adversarial threats extend beyond imaging to physiological time-series such as ECG and other biosignals, where sequential dependencies allow subtle manipulations to propagate into high-confidence errors. Xue et al. [86] analyzed weight-manipulation attacks targeting RNN-based healthcare models and proposed the detection of compromised learning modules. Hard-label black-box attacks on ECG classifiers show that effective evasion is possible even when only final predicted classes are available [87]. Defense work includes adversarial distillation training for ECG classification, improving robustness against low-noise PGD and other adversarial conditions [88]. More broadly, monitoring-oriented healthcare models also motivate adversarial/anomaly detection mechanisms for reliability in IoHT contexts [89].

Electronic health records (EHR), clinical data poisoning, and feature manipulation. Structured clinical records introduce a distinct adversarial surface where attackers can target event sequences, coding patterns, and feature semantics. Mozaffari et al. [90] proposed an algorithm-independent poisoning framework that injects adversarial data into training sets, inducing targeted or arbitrary misclassification across multiple healthcare datasets. Sun et al. [91] used adversarial attack formulations to identify vulnerable/sensitive locations in EHR by targeting LSTM prediction models. Ye et al. [92] introduced a hierarchical reinforcement-learning approach for black-box attacks on healthcare risk prediction models by selecting attack positions and substitutes within EHR features. Explainability-driven analysis has also been used to reveal the footprint of adversarial perturbations in EMR/EHR models, supporting accountability and rapid triage in clinical settings [93].

Smart healthcare, IoHT, and malware/Workflow oriented threats. Smart healthcare systems and IoHT deployments expand the threat model to connected apps, automated prescriptions, and device-integrated decision pipelines. Newaz et al. [94] demonstrated adversarial attacks against ML classifiers in smart healthcare systems, spanning poisoning and inference-time manipulation. Selvaganapathy et al. [95] studied GAN-based malware attacks on healthcare applications, showing that adversarial samples can bypass anti-malware classifiers, while denoising defenses had limited success. Gaglio et al. [96] evaluated adversarial manipulation in a smart prescription system, demonstrating high attack success with minimal record alteration. These studies illustrate ATLAS-aligned behaviors involving evasion, impact, and attack staging within operational healthcare workflows.

ATLAS: defensive countermeasures and robustness-oriented techniques. Defensive research in healthcare increasingly targets ATLAS-aligned behaviors by combining prevention (robust training), detection (anomaly/explainability), and architectural hardening. For imaging, ensemble-based defenses and adversarial retraining reduce misclassification under FGSM and one-pixel attacks [97], while denoising and robustness modules improve resilience across modalities [98,99]. Unsupervised detection methods can identify diverse attacks without requiring attacker assumptions, improving deployability [100]. Robustness can also be enhanced through feature fusion/ensemble mechanisms such as MEFF [101], and explainability-based detection provides model-agnostic screening across datasets and attack families [102]. Efficiency-aware defenses (e.g., GPU/parallel optimization) improve the practicality of robustness pipelines in clinical environments [103]. Comparative evidence suggests that architecture choice matters: Vision Transformers may exhibit improved resilience relative to CNNs under certain pathology conditions [104]. Overall, the healthcare literature shows a transition from isolated attack demonstrations toward integrated resilience toolchains that explicitly address evasion, poisoning, probing, and operational risk (Table 5 and 6).

6.3. Cybersecurity

Adversarial attacks represent a major security challenge for machine learning systems deployed in cybersecurity applications, where models

operate in adversarial environments and interact with intelligent attackers. Unlike traditional domains, cybersecurity systems must continuously adapt to evolving threats, making them particularly susceptible to adversarial manipulation across multiple stages of the ML lifecycle. Following the MITRE ATLAS framework, adversarial research in cybersecurity can be systematically organized according to adversary tactics (e.g., persistence, defense evasion, discovery, extraction, and impact) and corresponding technical techniques such as surrogate modeling, transferability exploitation, adversarial sample crafting, and poisoning strategies.

ATLAS: model access, extraction, and transferability-based attacks. Early foundational work by Papernot et al. [109] demonstrated the feasibility of black-box adversarial attacks using substitute models trained

Table 5
Systematic classification of adversarial attack methodologies and defenses in healthcare AI systems aligned with MITRE ATLAS tactics and techniques (Part I).

Title	Year	White-box	Black-Box	Other	Targeted	Untargeted	Defense	ATLAS Tactic(s)	ATLAS Technique Anchor
General Security, Privacy, and Infrastructure in e-Healthcare									
[67]	2016			x			x	Reconnaissance, Defense Evasion	Threat modeling / healthcare security baseline
[73]	2021			x	x		x	Discovery, Defense Evasion	Survey of imaging manipulation / GAN-based attacks
[71]	2024			x	x		x	Impact, Defense Evasion	Risk mitigation / adversarial prevention in healthcare AI
[68]	2021			x			x	Credential Access, Defense Evasion	Privacy-preserving access control for EHR
[69]	2023	x	x		x		x	Persistence	Federated poisoning defense / update verification
[70]	2021			x	x		x	Persistence, Defense Evasion	Secure federated aggregation for MCPS
Adversarial Attacks and Defenses in Medical Imaging									
[74]	2022	x	x		x	x		Defense Evasion	Imaging attack/defense survey; inference-time perturbations
[75]	2024	x	x		x	x	x	Defense Evasion, Impact	Robustness benchmarking / adversarial training benchmarks
[72]	2019	x	x		x		x	Defense Evasion	Imaging evasion taxonomy + mitigation measures
[100]	2020	x	x		x		x	Defense Evasion	Unsupervised adversarial example detection
[101]	2024	x	x		x	x	x	Defense Evasion	Ensemble feature fusion robustness (MEFF)
Modality-Specific Medical Imaging Attacks									
[76]	2023	x	x		x	x		Defense Evasion	Radiology attacks (X-ray/MRI/etc.) under WB/BB
[81]	2020	x			x	x		Defense Evasion	COVID-19 imaging evasion (FGSM-like perturbations)
[87]	2020		x		x	x		Model Probing, Defense Evasion	Hard-label black-box decision-based evasion
[77]	2021	x	x		x	x	x	Defense Evasion	Universal adversarial perturbations (UAP)
[105]	2022		x		x	x		Model Probing, Defense Evasion	Black-box UAP generation for imaging
[82]	2022		x	x	x			Defense Evasion, Impact	Watermarking manipulation as adversarial channel
[106]	2024	x				x	x	Defense Evasion	White-box FGSM stress test + robustness enhancement
[99]	2022	x	x		x	x	x	Defense Evasion	Denoiser-based defense (HGD) vs FGSM/PGD
Ophthalmology & Retinal Imaging									
[83]	2021	x			x		x	Defense Evasion	Adversarial training + feature fusion
[84]	2022	x			x		x	Defense Evasion	FGSM vulnerability analysis (MobileNet DR)
[85]	2022	x			x		x	Defense Evasion	FGSM/L-BFGS-B evaluation + training/distillation
Cancer and Disease-Specific Prediction Models									
[107]	2021		x		x			Model Probing, Defense Evasion	Transfer-based black-box attacks; pretraining effects
[97]	2020	x			x	x	x	Defense Evasion	FGSM / one-pixel + ensemble adversarial training
ECG, Physiological Signals, and Time-Series Healthcare Data									
[86]	2018	x			x		x	Persistence, Defense Evasion	Weight manipulation / model tampering in RNN
[88]	2024	x	x		x	x	x	Defense Evasion	Adversarial distillation training for ECG
[89]	2023	x			x	x	x	Defense Evasion	ConvLSTM detection of adversarial/ anomalous signals

Table 6
Systematic classification of adversarial attack methodologies and defenses in healthcare AI systems aligned with MITRE ATLAS tactics and techniques (Part II).

Title	Year	White-Box	Black-Box	Other	Targeted	Untargeted	Defense	ATLAS Tactic(s)	ATLAS Technique Anchor
Electronic Health Records (EHR) and Clinical Data									
[90]	2014	x			x		x	Persistence	Training-time poisoning in healthcare datasets
[91]	2018			x	x			Discovery, Model Probing	Identify sensitive EHR events via adversarial optimization
[92]	2022		x		x			Model Probing, Defense Evasion	Hierarchical RL black-box EHR manipulation
[93]	2022			x	x		x	Defense Evasion	XAI-guided analysis of CW perturbations on EMR
Specialized Adversarial Techniques									
[108]	2020	x	x		x		x	Defense Evasion	Neural style transfer attack on diagnostics
[79]	2023	x			x			Defense Evasion	Pixel-level manipulation attacks
[78]	2021			x	x			Defense Evasion	One-pixel minimal perturbation
[80]	2024	x	x		x	x		Defense Evasion	Combined FGSM + PGD compound perturbations
Malware, Smart Healthcare, and IoHT Systems									
[94]	2020	x	x		x	x		Impact, Defense Evasion	Adversarial ML in smart healthcare classifiers
[95]	2021			x	x	x		AI Attack Staging, Defense Evasion	GAN-based malware evasion
[96]	2021			x	x			Impact, Defense Evasion	Smart prescription manipulation
Explainability, Robustness, and Performance Optimization									
[98]	2022	x			x		x	Defense Evasion	Sparsity denoising operators in CNNs
[103]	2024			x	x	x	x	Defense Evasion	GPU/parallel optimization for defense pipelines
[102]	2021	x			x		x	Defense Evasion, Discovery	Explainability-based adversarial detection
[104]	2022	x	x		x	x	x	Defense Evasion	Robustness comparison: CNN vs ViT in pathology

through query access alone. Their work introduced transferability as a core adversarial tactic aligned with ATLAS model extraction and discovery behaviors, showing that attackers can manipulate remote classifiers without internal knowledge. Subsequent research expanded this direction, including brute-force black-box evaluation methods [110] and lightweight ensemble-based transferability attacks such as LEA^2 [16], which exploit heterogeneous surrogate models to maximize cross-model adversarial effectiveness while reducing computational cost. Empirical transferability studies further confirmed that adversarial perturbations generated from surrogate models remain highly effective against unseen targets [111]. These works collectively emphasize that realistic cybersecurity attackers rely primarily on partial information and exploit transferability rather than direct gradient access.

ATLAS: defense evasion and operational security attacks. A significant portion of cybersecurity research focuses on evasion-oriented tactics targeting intrusion detection systems (IDS), malware classifiers, and botnet detection frameworks. Apruzzese et al. [112,113] demonstrated that both poisoning and evasion attacks can significantly degrade detection performance in real network traffic environments. Similar findings were reported in intrusion detection studies evaluating neural architectures under adversarial manipulation [114]. GAN-driven adversarial feature generation has also emerged as a powerful evasion mechanism, enabling attackers to craft malicious traffic that mimics legitimate behavior to bypass anomaly detection [115]. Surveys focusing on adversarial NIDS vulnerabilities further highlight how minimal perturbations can evade defensive systems in realistic black-box scenarios [116]. Within the ATLAS taxonomy, these attacks align primarily with defense evasion and impact tactics, reflecting real-world attacker objectives.

ATLAS: explainability exploitation and information leakage. Recent work has identified explainable AI (XAI) mechanisms as an additional attack surface in cybersecurity. Kuppa et al. [117,118] showed that explanations can be manipulated or exploited to reveal sensitive information about underlying models, enabling model extraction, membership inference, and poisoning attacks. These approaches align with ATLAS discovery and credential access tactics, illustrating how explainability introduces new vulnerabilities beyond classification outputs. Broader surveys analyzing adversarial learning from a cybersecurity perspective emphasize lifecycle-oriented threat modeling and the need to account for explanation-driven risks [119,120].

ATLAS: generative models and adversarial sample synthesis. Generative adversarial networks (GANs) play a dual role in cybersecurity as both offensive and defensive tools. Reviews by Yinka et al. [121] and Dutta et al. [122] have highlighted how GANs enable realistic adversarial traffic generation capable of bypassing security models while simultaneously supporting robustness improvements through adversarial training. Such techniques correspond to ATLAS persistence and defense evasion tactics, where adversarial samples are crafted to maintain long-term stealth within monitored systems.

ATLAS: defensive countermeasures and robustness-oriented techniques. In response to increasing adversarial threats, research has shifted toward practical defense mechanisms and realistic threat modeling. Frameworks for modeling attacker capabilities and constraints [26] help align evaluations with operational cybersecurity environments. Defense strategies include adversarial detection algorithms [123], adversarial training and preprocessing pipelines [124], and lifecycle-based defense–attack–enhanced-defense frameworks [12]. Additional work explores explainability-driven detection methods and standardized evaluation

Table 7
Systematic classification of adversarial attack methodologies in cybersecurity aligned with the MITRE ATLAS taxonomy.

Title	Year	White-box	Black-box	Other	Targeted	Untargeted	Defense	ATLAS Tactic(s)	ATLAS Technique Anchor
Adversarial Attacks and Vulnerabilities in Cybersecurity Systems									
[109]	2017		x		x			Discovery, Extraction	Substitute model transferability
[112]	2019	x	x		x	x		Defense Evasion	IDS evasion / poisoning
[113]	2019		x		x	x		Defense Evasion	Botnet detector evasion
[114]	2019		x			x		Defense Evasion	IDS perturbation attack
[110]	2020		x		x			Discovery	Query-based adversarial generation
[116]	2021		x			x		Defense Evasion	NIDS evasion
[115]	2022		x			x		Persistence, Defense Evasion	GAN adversarial traffic
[111]	2024	x	x			x		Discovery	Transferability attack
[16]	2023		x		x			Persistence	Ensemble surrogate attack
Adversarial Defenses, Robustness, and Secure System Design									
[26]	2022		x		x	x		Defense Evasion	Realistic threat modeling
[127]	2024		x			x		Defense Evasion	Quantum ensemble defense
[123]	2020		x			x	x	Defense Evasion	Adversarial detection
[124]	2024		x			x	x	Defense Evasion	Adversarial training
Frameworks, Explainability, and Systematic Analyses									
[128]	2018			x			x	Discovery	Survey taxonomy
[117]	2020		x		x			Discovery	XAI manipulation
[118]	2021		x		x			Discovery, Extraction	Explanation-based attack
[119]	2021			x			x	Discovery	Lifecycle taxonomy
[120]	2022			x	x	x		Discovery	Attack lifecycle analysis
[121]	2020		x			x		Persistence	GAN adversarial generation
[122]	2020		x		x			Persistence	GAN adversarial modeling
[12]	2022		x		x	x		Defense Evasion	Defense lifecycle
[125]	2024		x				x	Defense Evasion	Defense evaluation framework
[126]	2024		x				x	Defense Evasion	Network security defenses

frameworks for robust adversarial defense [125,126]. These approaches map primarily to ATLAS defensive countermeasure tactics aimed at reducing attacker success while maintaining system reliability.

Overall, the cybersecurity literature summarized in Table 7 reveals a progression from early feasibility demonstrations toward realistic threat modeling and defense-aware system design. Black-box attack models dominate due to operational constraints in deployed systems, while emerging trends include explainability-aware attacks, transferability-driven adversarial strategies, and lifecycle-oriented defense frameworks aligned with MITRE ATLAS.

6.4. Industrial control system (ICS)

In the context of an Industrial Control Systems (ICS), adversarial attacks can have devastating consequences for cybersecurity. For instance, an attacker could slightly modify malicious payloads or packet structures to bypass detection, allowing unauthorized access, malware infiltration, or data exfiltration to proceed without triggering an alert. Conversely, an attacker could flood the system with crafted traffic that generates numerous false positives, causing system administrators to overlook real threats amidst the noise. This tactic, known as “adversarial evasion,” can allow attackers to operate stealthily within a network while compromising critical assets. Additionally, adversarial attacks might target anomaly-based detection systems by gradually shifting the baseline of normal behavior, rendering the IDS less sensitive to real attacks over time.

Defending against adversarial attacks on ICS requires robust strategies, including adversarial training to harden models against manipulation, implementing ensemble methods that use multiple detection models for cross-verification, and integrating human oversight to analyze anomalies and suspicious patterns that automated systems might miss. Another approach is to leverage hybrid systems combining signature-based and anomaly-based detection to enhance resilience. As cyber-attacks become more sophisticated, ensuring that IDS can withstand adversarial manipulation is crucial for maintaining the integrity and security of modern network infrastructures. Strengthening these systems is essential to prevent adversaries from exploiting vulnerabilities that could lead to significant breaches and disruptions. Adversarial attacks have become a major threat to ICS, with various studies aiming to assess vulnerabilities and propose defenses. A foundational study by Line et al. [129] evaluated the preparedness of the power industry against targeted attacks, focusing on well-known attack vectors. They developed a taxonomy of these attacks and assessed the situational awareness of power distribution system operators, revealing that many sophisticated attacks, especially those involving social engineering, remain inadequately addressed. They offered guidelines to enhance the detection and response capabilities of these operators, laying the groundwork for improving ICS resilience.

Building on the understanding of targeted attacks, Urbina et al. [130] examined how adversaries can manipulate sensor or control signals within ICS environments. Their comprehensive review of attack detection mechanisms highlighted a significant gap in mitigating stealthy attacks. By introducing a new metric to assess the impact of these attacks, they demonstrated that the proper combination of detection schemes could effectively limit the damage caused by such covert activities. Their findings emphasized the need for robust configuration of ICS defenses. Expanding on these defensive strategies, Paridari et al. [131] proposed a novel cyber-physical security framework tailored to ICSs. Their framework incorporates an advanced analytics tool for detecting attacks and an estimation-based attack-resilient control policy that maintains system stability even under attack. They validated the effectiveness of this approach using simulations on a real energy management system, demonstrating the potential for practical implementation in ICS environments.

While the above works focused on detection and response, Feng et al. [132] took a different approach by exploring the potential for adversarial deep learning. They developed a framework that allows attackers to conduct stealthy attacks with minimal prior knowledge of the target ICS. By intercepting sensor and control signals, attackers can autonomously generate sophisticated attacks capable of bypassing anomaly detectors. Their research, which included real-world case studies, underscored the escalating complexity of adversarial attacks on ICSs. Complementing

this work, [133] explored evasion attacks targeting dynamic industrial control systems. His research examined attacks that manipulated sensor readings within the physical constraints of the system, demonstrating that these manipulated readings could evade detection algorithms. By comparing the cost and efficiency of these evasion attacks to traditional replay attacks, Erba highlighted the growing need for more sophisticated defenses. He introduced both white-box attacks using optimization methods and black-box attacks based on autoencoders, showing how adversaries can disguise anomalous data as normal. Continuing the exploration of stealthy attacks, Chen et al. [134] investigated a variety of cyberattacks against IDS in ICS environments, including injection attacks, function code attacks, and reconnaissance attacks. Their research introduced two attack strategies: the optimal solution and the GAN-based approaches, both of which proved effective on real-world ICS testbeds. By incorporating adversarial training, they were able to enhance the robustness of ML-based IDS models, further emphasizing the importance of adapting detection systems to the evolving nature of adversarial attacks. Sarkar et al. [135] expanded on these ideas by investigating an automated end-to-end attack framework designed to exploit unknown control processes within ICSs. They trained ML models to fingerprint ICS sectors and reverse engineer PLC binaries, allowing them to manipulate control processes with precision. Their work demonstrated the feasibility of conducting advanced attacks even in constrained environments, offering insights into how attackers can leverage ML to bypass ICS defenses.

In contrast to these offensive studies, Alabugin et al. [136] explored the use of GANs for detecting anomalies in industrial processes. By implementing the BiGAN architecture, they proposed an anomaly detection system that performed well when tested on the Secure Water Treatment Dataset. This approach highlights the potential for using adversarial methods in a defensive capacity, turning the tables on attackers by leveraging GANs to identify unusual behavior in ICS data. Further advancing this line of inquiry, Erba et al. [137] investigated how adversaries could evade reconstruction-based anomaly detectors by manipulating sensor data. Using learned physical constraints, they showed that both white-box and black-box attackers could hide anomalies within subsets of sensor readings, thus significantly reducing detection accuracy. This research underscored the importance of building detection systems that can adapt to more sophisticated evasion tactics.

Anthi et al. [49] expanded the scope of adversarial learning research by focusing on the generation of adversarial samples using the Jacobian-based Saliency Map Attack. Their experiments on Random Forest and J48 classifiers demonstrated how these adversarial samples could significantly degrade classification accuracy, with decreases of up to 11 percentage points. Despite the vulnerabilities exposed, their research also illustrated how adversarial training could improve the robustness of these models, offering a clear path to mitigating such attacks.

Kravchik et al. [138] turned their attention to poisoning attacks, targeting neural network-based detection systems in ICS environments. They proposed two innovative poisoning strategies: interpolation-based and backgradient-based, both of which proved highly effective against synthetic and real-world ICS data. Their findings highlighted significant weaknesses in neural network detection systems, while also suggesting mitigation strategies to counteract these vulnerabilities. Building on the broader challenges facing ICS security, Makrakis et al. [139] conducted a comprehensive survey of the prominent threats to ICS. They categorized vulnerabilities in operational technology-specific network protocols and devices, offering a framework for understanding how adversaries can exploit systemic weaknesses. This survey provided essential insights into the growing threat landscape and outlined potential future defensive measures.

Following this, Gómez et al. [140] introduced a novel adversarial attack technique, the Selective and Iterative Gradient Sign Method (SIGSM). Their experiments, using the Electra dataset from an Electric Traction Substation, revealed how these adversarial samples could bypass intermediate network devices, creating new attack vectors that threaten ICS infrastructure. This study emphasized the importance of adapting to evolving adversarial techniques to protect critical systems. Umer et al. [141] contributed to the ongoing discussion on ICS security assessment by proposing an attack generation technique based on association rule mining. Tested on data from a Secure Water Treatment plant, their method generated over 110,000 unique attack vectors, most of which had never been encountered before. This highlighted the unpredictable nature of adversarial threats and the need for more comprehensive assessment techniques.

Figueroa et al. [142] further explored adversarial attacks by studying the impact of three specific techniques—FGSM, DeepFool, and Jacobian-Based Saliency Map Attacks—on ML models in power systems. Their findings revealed significant performance degradation in DNNs, reinforcing the urgency of robust defense mechanisms to protect critical infrastructure against adversarial threats. Returning to the topic of poisoning attacks, Kravchik et al. [143] proposed two new attack algorithms, specifically designed for ICS detectors based on neural networks. Building on earlier research, they demonstrated the broad applicability of these attacks across various ICS testbeds and processes, while also proposing mitigation strategies to secure neural network-based detection systems.

Yao et al. [144] took a proactive approach by developing a resilient ML framework designed to resist adversarial attacks. Their framework dynamically anonymizes feature spaces and randomizes models during runtime, preventing adversaries from manipulating ML operations. Validated using ICS datasets, this approach demonstrated a promising method for safeguarding ICS environments against adversarial manipulation. Duan et al. [145] tackled the challenge of attacking black-box models with limited queries, proposing the Attack Contrastive Learning Network (ACL-Net) as an efficient solution. Their model generated substitute models capable of conducting highly effective black-box attacks with fewer queries, outperforming previous methods in both query efficiency and attack performance.

Furthermore, Pozdnyakov et al. [146] examined the vulnerabilities of deep learning models in fault diagnosis for automated control systems. Subjecting these models to six different types of adversarial attacks, they explored five defense methods, demonstrating the strong vulnerability of these models to adversarial manipulation. Their novel approach, which combined multiple defense methods, proved highly effective in mitigating adversarial threats in ICS environments. Finally, Liu et al. [147] concluded this body of work by addressing a significant challenge in adversarial defense: designing models without prior knowledge of adversarial samples. They proposed an LSTM-ED-based method for generating adversarial samples that conform to protocol specifications and physical constraints. Their research demonstrated the substantial impact these adversarial samples could have on model performance in real-world ICS environments, while their LSTM-FWED defense method effectively mitigated these risks.

Table 8 summarizes the ICS literature according to the taxonomy dimensions adopted throughout this survey (attacker knowledge, perturbation/attack type, targeting, and defensive intent), and organizes the reviewed works into three dominant research streams. First, *Stealthy & Signal-Level Attacks* capture the core ICS threat surface where adversaries manipulate sensor/actuator signals and protocol-level measurements to remain within physical constraints while evading anomaly detectors, spanning both black-box and white-box settings and often focusing on targeted disruption. Second, *Adversarial ML & Poisoning Attacks* highlight the fragility of learning-based IDS and fault-diagnosis pipelines when attackers exploit training- or inference-time ML vulnerabilities (e.g., adversarial examples and poisoning), typically leading to untargeted degradation of detection accuracy and reliability. Third, *Defense, Resilience & Detection Frameworks* groups works that shift the focus from attack construction to operational robustness, including resilient control policies, GAN/LSTM-based detection schemes, and learning frameworks that explicitly aim to sustain performance under adversarial conditions. Overall, the table makes clear that ICS research

Table 8
ATLAS-aligned systematic taxonomy of adversarial attacks and defenses in industrial control systems (ICS).

Title	Year	White-box	Black-box	Other	Targeted	Untargeted	Defense	ATLAS Tactic(s)	ATLAS Technique Anchor
Stealthy Cyber-Physical Signal Manipulation & Evasion									
[126]	2014			x	x			Reconnaissance	Social engineering / targeted attack planning
[127]	2016			x			x	Defense Evasion	Stealthy signal manipulation
[129]	2017		x	x	x			Evasion	Autonomous adversarial generation
[130]	2019	x	x	x		x		Evasion	Sensor data manipulation
[131]	2020			x	x			Evasion	GAN-based adversarial injection
[132]	2020			x			x	Model Discovery	Process fingerprinting
[134]	2020	x	x	x		x		Defense Evasion	Physically constrained perturbations
[137]	2021			x		x		Evasion	Gradient-sign manipulation (SIGSM)
[138]	2021			x	x			Reconnaissance	Association-rule attack generation
[142]	2024		x	x		x		Model Extraction	Substitute model learning
[139]	2022			x		x		Evasion	FGSM / DeepFool / JSMA attacks
Adversarial ML & Poisoning Attacks									
[45]	2021			x		x		Evasion	JSMA adversarial generation
[135]	2021			x		x		Model Manipulation	Training data poisoning
[140]	2022			x		x		Model Manipulation	Backgradient poisoning
[143]	2024			x		x	x	Defense	Multi-defense ensemble
Defense, Resilience & Detection Frameworks									
[128]	2017			x			x	Defense	Resilient control policy
[133]	2020		x	x			x	Detection	GAN-based anomaly detection
[136]	2021			x			x	Risk Assessment	Threat landscape modeling
[141]	2023		x	x		x	x	Defense	Randomization-based resilience
[144]	2024		x	x		x	x	Defense	LSTM-based adversarial mitigation

is increasingly converging toward *attack realism* (physical feasibility and low observability) and *system-level resilience*, since purely model-centric defenses are often insufficient in safety-critical cyber-physical environments.

6.5. Autonomous vehicles

Adversarial attacks represent a major threat to autonomous vehicle (AV) ecosystems, where machine learning models operate as safety-critical components responsible for perception, planning, and decision-making. Unlike traditional IT environments, AV systems integrate cyber-physical sensing pipelines combining cameras, LiDAR, radar, and vehicle control modules, meaning adversarial manipulation can directly translate into unsafe physical behavior, as shown in Fig. 6. Within the MITRE ATLAS perspective, adversarial threats in autonomous driving primarily align with *Defense Evasion*, *Impact*, *Model Manipulation*, and *Discovery* tactics, reflecting how attackers exploit perception vulnerabilities to influence downstream control decisions.

Adversarial perturbations in AV environments frequently target visual recognition tasks such as traffic sign detection, object classification, and lane segmentation. Physical-world attacks are particularly concerning because they remain effective outside laboratory settings. For example, Qian et al. [148] demonstrated localized physical evasion attacks against license plate recognition systems by introducing realistic spot-like perturbations optimized via genetic algorithms, achieving success rates above 93%. These results highlight that perception models can be compromised through subtle environmental modifications without direct system access.

ATLAS: white-box perception and control manipulation. White-box adversarial scenarios correspond to ATLAS tactics such as model manipulation and execution-phase evasion, where attackers exploit gradient information or system transparency. Xu et al. [149] demonstrated iterative gradient-based perturbations capable of misleading urban scene segmentation models, while Jiang et al. [150] explored poisoning and evasion strategies targeting traffic sign recognition using particle swarm optimization. Similarly, Li et al. [151] introduced adversarial trajectory perturbation affecting LiDAR perception indirectly by manipulating vehicle motion dynamics. Universal adversarial perturbations described by Zhang et al. [152] further revealed cross-model vulnerabilities affecting multiple object detection architectures. Beyond perception attacks, Won et al. [153] addressed cryptographic attack surfaces in unmanned vehicle systems, proposing shuffled lookup tables to mitigate reverse engineering threats aligned with ATLAS persistence and credential protection strategies. Sobh et al. [154] showed that both white-box and black-box attacks can degrade multi-task perception pipelines simultaneously, illustrating systemic AV vulnerability.

ATLAS: black-box transferability and sensor manipulation. Black-box attacks map closely to ATLAS discovery and model probing tactics, relying on transferability or limited query access. Kumar et al. [155] proposed an efficient M-SimBA black-box attack accelerating adversarial convergence against traffic sign recognition models. Zhang et al. [156] demonstrated universal perturbations transferable across YOLOv3 and Faster R-CNN detectors, reinforcing the risk posed by model-agnostic attacks. LiDAR-based systems represent another critical attack surface; Sun et al. [157] manipulated point clouds to generate phantom obstacles, while Zhu et al. [158] showed that simple physical objects could deceive segmentation systems with success rates exceeding 90%. Additional stealthy black-box strategies include the Suspicion-Free Boundary Attack and context-aware adversarial approaches introduced by Sarker et al. [159,160], which minimize perturbation visibility while preserving attack effectiveness. Sensor fusion systems are similarly vulnerable; frustum-based attacks described by Hallyburton et al. [161] exploit inconsistencies between camera and LiDAR inputs, while Jakobsen et al. [162] highlighted weaknesses in multi-sensor fusion architectures.

ATLAS: defensive countermeasures and robustness-oriented techniques. Defensive research aligns with ATLAS mitigation and detection tactics, focusing on strengthening system resilience against adversarial manipulation. Deng et al. [163] evaluated adversarial training and model distillation for improving CNN-based driving models, while Liu and Park [164] proposed the LIFE framework to detect perception inconsistencies across sensing modalities. Dhawale et al. [165] introduced input transformation defenses for traffic sign recognition, and Shibly et al. [166] leveraged autoencoder-based memory mechanisms to prevent adversarial generalization. Complementary survey works such as

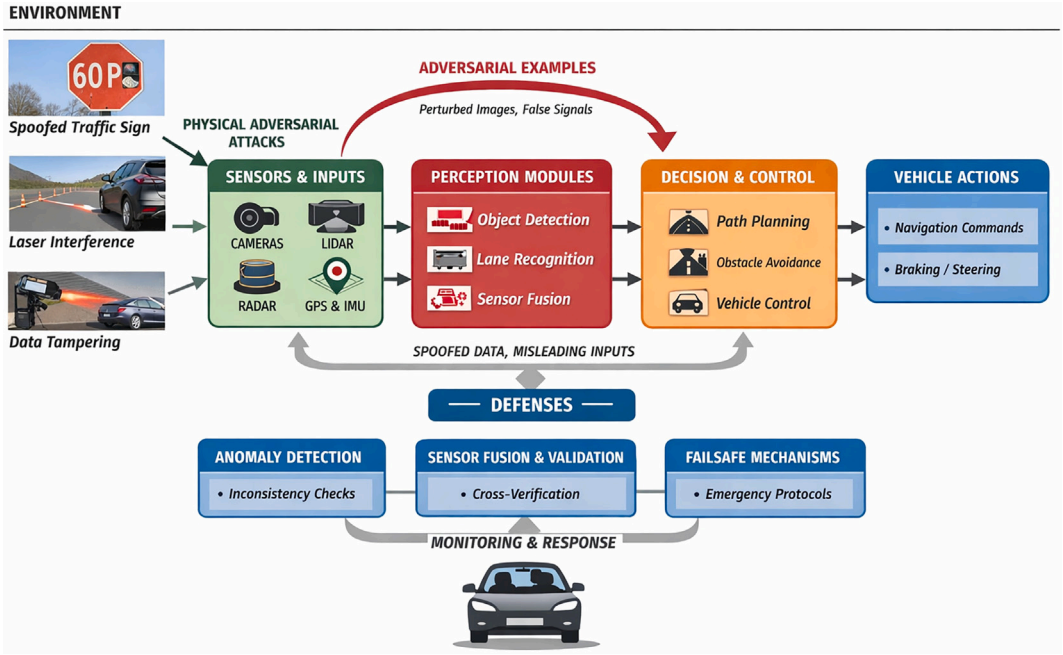

Fig. 6. Architecture-level overview of adversarial attack surfaces and defense mechanisms in autonomous vehicle systems. The diagram illustrates how physical-world and digital adversarial perturbations (e.g., spoofed traffic signs, sensor interference, and data manipulation) propagate through the sensing pipeline (camera, LiDAR, radar, GPS/IMU) into perception modules such as object detection, lane recognition, and sensor fusion. These perturbations may influence downstream decision-making and vehicle control actions, potentially causing unsafe behavior. The framework also highlights multi-layer defensive strategies including anomaly detection, cross-sensor validation, and fail-safe mechanisms, emphasizing the importance of defense-in-depth across sensing, perception, and control stages.

[167–172] further emphasize the growing need for adversarially robust perception pipelines, secure V2X communication architectures, and multi-layered defense frameworks integrating anomaly detection, adversarial training, and sensor redundancy.

Overall, autonomous vehicle research demonstrates a transition from isolated perception attacks toward system-level adversarial analysis encompassing sensor fusion, trajectory planning, and cyber-physical resilience. The ATLAS-aligned taxonomy highlights that realistic threat models increasingly combine physical-world perturbations, transferability-based attacks, and cross-modality exploitation, underscoring the importance of holistic defense strategies capable of preserving safety under adversarial conditions (Table 9).

6.6. Speech recognition

Adversarial attacks against Automatic Speech Recognition (ASR) systems represent a rapidly evolving threat landscape, particularly as speech interfaces become integral components of smart devices, autonomous systems, healthcare assistants, and security-critical infrastructures. Unlike traditional adversarial attacks in vision systems, adversarial audio attacks must operate under psychoacoustic constraints to remain imperceptible to human listeners while still manipulating model predictions. These attacks typically introduce carefully crafted perturbations into audio signals that exploit vulnerabilities in deep neural networks, enabling malicious command injection, misclassification, or unauthorized system actions.

The security implications are significant because ASR systems often operate in real-time environments with minimal human oversight. Attackers can embed hidden commands within benign audio or leverage environmental transmission channels to bypass detection. Fig. 7 illustrates a representative adversarial attack scenario where imperceptible perturbations induce incorrect transcription in a deployed ASR pipeline.

ATLAS: white-box adversarial generation and model manipulation. White-box adversarial attacks assume full knowledge of the target ASR model architecture and parameters, enabling precise gradient-based optimization of adversarial perturbations. Zong et al. [173] demonstrated universal adversarial perturbations targeting Connectionist Temporal Classification (CTC)-based ASR models, showing that targeted phrases can be embedded into audio signals while remaining imperceptible. Similarly, Xu et al. and Kwon et al. [174] explored targeted attacks that selectively manipulate transcription outputs or specific classifiers. Multi-target optimization approaches proposed by Ko et al. [175] further highlight how adversarial signals can simultaneously degrade multiple ASR models. These attacks align with ATLAS tactics such as *Defense Evasion* and *Model Manipulation*, as adversaries exploit model gradients to craft effective perturbations.

ATLAS: black-box, query-based, and transferability attacks. In realistic deployment scenarios, attackers rarely possess full model knowledge, motivating black-box attack strategies. Zheng et al. [176] introduced Occam and NI-Occam attacks that exploit minimal feedback from commercial speech APIs, demonstrating high success rates despite limited information. Query-efficient attacks such as the Monte Carlo Gradient Sign Attack (MGSA) [177] and Temporal Natural Evolution Strategies (T-NES) [178] reduce computational overhead while maintaining attack effectiveness. Zero-query and transferability-based attacks, including ensemble-based methods proposed by Fang et al. [179] and AdvDDoS [180], further demonstrate the feasibility of generating adversarial audio without interacting with the target model. These approaches map to ATLAS tactics including *Discovery*, *Defense Evasion*, and *Model Extraction*.

ATLAS: physical-world and realistic audio attacks. Beyond digital perturbations, recent research explores adversarial attacks that operate in real-world acoustic environments. Laser-based signal injection techniques such as LaserAdv [181] demonstrate long-range adversarial attacks capable of manipulating ASR systems through physical transmission channels. Psychoacoustic attacks such as IMPGA [182] exploit human auditory masking properties to maintain imperceptibility while

Table 9
ATLAS-aligned adversarial attack taxonomy for autonomous vehicle AI systems, including perception evasion, sensor manipulation, transferability-based black-box threats, and robustness-oriented defenses.

Title	Year	White-box	Black-box	Other	Targeted	Untargeted	Defense	ATLAS Tactic(s)	ATLAS Technique Anchor
Perception Evasion and Physical-World Attacks									
[148]	2020		x		x			Defense Evasion, Impact	Physical localized perturbations (license plates)
[158]	2021		x		x			Defense Evasion, Impact	Physical object-based LiDAR evasion
[157]	2020		x		x		x	Defense Evasion, Detection	LiDAR point-cloud spoofing; CARLO anomaly detection
White-box Attacks on AV Perception / Planning Pipelines									
[149]	2020	x				x	x	Defense Evasion, Model Manipulation	Iterative gradient-based segmentation attack + adv training
[150]	2020	x			x	x		Persistence, Defense Evasion	Poisoning + evasion via PSO on traffic signs
[151]	2021	x			x			Defense Evasion, Impact	Trajectory perturbation to fool LiDAR perception
[152]	2020	x				x		Defense Evasion, Impact	Universal perturbations against object detection
[154]	2021	x	x		x			Defense Evasion, Impact	Multi-task perception attack (segmentation/distance)
Black-box Transferability, Query-Efficient, and Boundary Attacks									
[155]	2020		x		x			Discovery, Defense Evasion	Query-efficient black-box (M-SimBA)
[156]	2021		x			x		Defense Evasion, Impact	Transferable universal perturbations across detectors
[159]	2021		x		x			Discovery, Defense Evasion	Stealthy boundary black-box attack (SBBA)
[160]	2021		x		x			Discovery, Defense Evasion	Context-aware black-box attack on spatio-temporal signals
[161]	2022		x		x			Defense Evasion, Impact	Frustum attack on camera–LiDAR fusion
[162]	2023		x			x	x	Discovery, Defense Evasion	Multi-sensor fusion disruption + defense analysis
Cryptographic / System Security Components in AV Ecosystems									
[153]	2019			x			x	Credential Access, Defense Evasion	LUT shuffling against reverse engineering
[169]	2023			x			x	Discovery, Mitigation	Reference architecture + security controls for CAV
Defenses, Robustness, and Detection Mechanisms									
[163]	2020	x					x	Mitigation, Detection	Adversarial training + distillation for driving models
[164]	2021			x			x	Detection, Mitigation	Cross-modal inconsistency detection (LIFE)
[165]	2022			x			x	Mitigation	Input transformation / denoising defense (traffic signs)
[166]	2023	x	x			x	x	Mitigation, Detection	Autoencoder + compressive memory defense
[171]	2022			x			x	Mitigation	Scene-analysis defenses; denoising and transformations
Surveys and Systematic Reviews									
[167]	2021			x			x	Discovery, Mitigation	AV cyberattack taxonomy (control/ADS/V2X)
[168]	2021			x			x	Discovery, Mitigation	Survey of adversarial attacks on ADS + defense directions
[170]	2023			x			x	Discovery, Mitigation	Review of AV cybersecurity challenges and defenses
[172]	2023			x			x	Discovery, Mitigation	Survey of robustness for object detection in AVs

achieving high attack success rates. Robust universal perturbations proposed by Qin et al. [183] integrate environmental noise modeling to ensure effectiveness under real-world conditions. These attacks correspond to ATLAS categories related to *Impact*, *Defense– Evasion*, and realistic adversarial execution.

ATLAS: defensive countermeasures and robustness-oriented techniques. To counter adversarial audio threats, several defense strategies have been proposed. Preprocessing-based defenses, such as speech coding and acoustic filtering techniques [184,185], aim to remove adversarial noise prior to transcription. Memory-based fingerprinting methods introduced by Guo et al. [186] analyze repeated query patterns to detect adversarial intent without retraining the underlying model. Adversarial training approaches, including mixPGD [187], enhance robustness by incorporating adversarial examples during training. Hybrid denoising architectures combining preprocessing and model adaptation [188] further improve

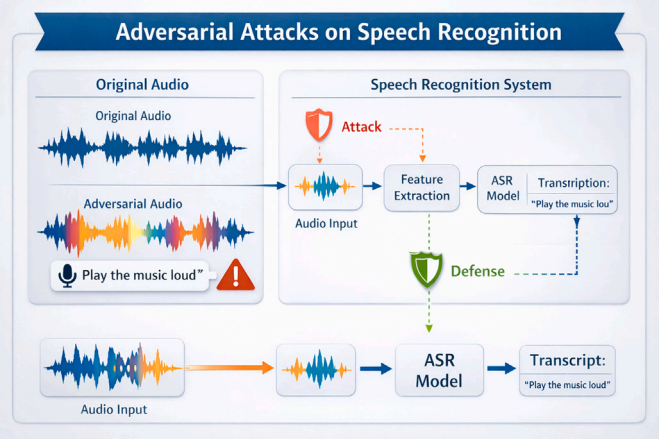

Fig. 7. Adversarial speech recognition pipeline illustrating how small perturbations to an audio waveform can propagate through feature extraction and the ASR model to induce a targeted transcription, and where defenses can be applied to detect or mitigate the attack before producing the final transcript.

resilience against both white-box and black-box attacks. These defenses align with ATLAS tactics such as *Mitigation*, *Detection*, and *Persistence*.

Overall, the ASR adversarial landscape reveals an escalating arms race between attack sophistication and defensive robustness. Table 10 summarizes the literature according to attacker knowledge, targeting objectives, defensive intent, and ATLAS-aligned techniques. While significant progress has been made in developing robust defenses, the diversity of adversarial threat models—including digital, query-based, and physical-world attacks—demonstrates that comprehensive protection requires multi-layered defense strategies and standardized evaluation frameworks tailored to speech recognition systems.

6.7. Natural language processing (NLP)

Adversarial attacks in Natural Language Processing (NLP) exploit structural vulnerabilities in language models by introducing subtle textual perturbations that alter model predictions while preserving semantic meaning for human readers. Unlike continuous domains such as computer vision, NLP attacks must operate within discrete linguistic constraints, making imperceptibility closely tied to grammar, syntax, and semantic consistency. Small changes such as synonym substitutions, character-level modifications, paraphrasing, or contextual alterations can significantly degrade performance in tasks such as sentiment analysis, machine translation, fake news detection, and named entity recognition. As NLP models increasingly support decision-making in sensitive domains—including legal analysis, healthcare documentation, and automated moderation—the development of robust adversarial defenses has become critical.

ATLAS: white-box and gradient-guided adversarial generation. White-box adversarial attacks leverage access to model parameters or gradients to identify influential tokens and optimize perturbations. Universal adversarial texts introduced by Li et al. [192] demonstrate that short trigger phrases can systematically manipulate predictions across different models. Differentiable attack strategies proposed by Fursov et al. [193] exploit pre-trained language models to generate adversarial samples with strong transferability while preserving fluency. SeqAttack [194] further shows that token classification tasks such as Named Entity Recognition remain vulnerable to character- and word-level manipulations. These methods align with ATLAS tactics such as *Model Manipulation* and *Defense Evasion*, where attackers exploit internal model behavior to produce targeted misclassifications.

ATLAS: black-box, query-efficient, and transferability-based attacks. Most real-world NLP attacks operate under restricted access assumptions, motivating black-box approaches that rely on model querying or surrogate optimization. Population-based word substitution attacks proposed by Alzantot et al. [195] demonstrate high success rates against sentiment analysis and textual entailment models while maintaining semantic similarity. Reinforcement-learning–based optimization introduced by Zang et al. [196] improves query efficiency by learning optimal perturbation strategies. Decision-based hard-label attacks such as those proposed by Maheshwary et al. [197] further reduce the required feedback from target models. Recent methods including BFS2Adv [198], attention-based genetic optimization [199], and LLM-driven adversarial generation [200] highlight growing sophistication in query-efficient attack strategies. These approaches correspond to ATLAS tactics such as *Discovery*, *Model Extraction*, and *Defense Evasion*.

Table 10
ATLAS-aligned taxonomy of adversarial attacks and defenses in speech recognition (ASR) systems, including universal perturbations, query-based attacks, physical-world attacks, and robustness mechanisms.

Title	Year	White-box	Black-box	Other	Targeted	Untargeted	Defense	ATLAS Tactic(s)	ATLAS Technique Anchor
Adversarial Attacks on ASR Systems									
[173]	2021	x			x			Defense Evasion	UAP attack on CTC-based ASR
[189]	2022	x	x	x	x	x		Discovery, Defense Evasion	Survey of ASR adversarial taxonomy
[176]	2021		x		x			Discovery, Defense Evasion	Occam / NI-Occam black-box attack
[175]	2023	x			x			Defense Evasion	Multi-target optimization attack
[190]	2021	x	x	x	x	x		Discovery	SoK analysis of ASR vulnerabilities
[191]	2024	x				x		Defense Evasion	Transferable CommanderUAP
[174]	2019	x			x			Defense Evasion	Selective adversarial targeting (DeepSpeech)
[183]	2023		x			x		Defense Evasion, Impact	Robust real-world UAP with noise modeling
[179]	2024		x		x			Discovery, Defense Evasion	Zero-query ensemble attack
[180]	2023		x		x			Impact, Defense Evasion	AdvDDoS zero-query UAP
[178]	2023		x			x		Discovery	Temporal NES query-efficient attack
[181]	2024		x		x			Impact, Defense Evasion	Laser-based physical ASR attack
[177]	2023		x			x		Discovery	Monte Carlo Gradient Sign Attack
[182]	2023		x			x		Defense Evasion	Imperceptible GA (psychoacoustic attack)
Defense Methods for ASR Systems									
[184]	2018			x			x	Mitigation	Speech coding + preprocessing defense
[186]	2023			x			x	Detection, Mitigation	Memory-based fingerprinting defense
[185]	2020			x			x	Mitigation	Acoustic-decoy filtering
[188]	2022			x			x	Mitigation	Denoiser + adversarial fine-tuning
[187]	2023	x	x				x	Mitigation, Persistence	MixPGD adversarial training

Table 11
Systematic taxonomy of adversarial attacks and defense strategies in NLP aligned with MITRE ATLAS tactics and techniques.

Title	Year	White-box	Black-box	Other	Targeted	Untargeted	Defense	ATLAS Tactic(s)	ATLAS Technique Anchor
Black-box and Query-Efficient Attacks									
[195]	2018		x		x			Defense Evasion	Adversarial Example Generation
[196]	2020		x		x			Model Manipulation	Reinforcement Learning Attack
[197]	2021		x			x		Discovery	Decision-based Attack
[193]	2022	x			x			Model Manipulation	Differentiable Text Attack
[199]	2022		x			x		Discovery	Attention-based Optimization
[198]	2024		x			x		Defense Evasion	Search-based Perturbation
Semantic-preserving and Linguistic Attacks									
[201]	2021			x	x			Impact	Linguistic Expansion Attack
[192]	2021	x	x		x			Model Manipulation	Universal Adversarial Text
[202]	2019		x		x		x	Impact	Fact Manipulation
[200]	2024		x		x			Defense Evasion	LLM-driven Adversarial Generation
Task-specific and Structured Attacks									
[194]	2021	x			x			Model Manipulation	Token-level Perturbation
[204]	2023		x			x		Discovery	Code Structure Attack
Defense Mechanisms									
[205]	2020						x	Mitigation	Randomized Smoothing
[218]	2021			x			x	Mitigation	Robust Training
[206]	2021						x	Mitigation	Synonym Encoding Defense
[207]	2022	x	x			x	x	Mitigation	Adversarial Training
[208]	2023					x	x	Mitigation	Certified Robustness
Frameworks, Surveys and Evaluation									
[28]	2020	x	x		x	x	x	Discovery	Evaluation Framework
[209]	2020			x			x	Reconnaissance	Survey
[210]	2020			x			x	Reconnaissance	Survey
[211]	2021			x			x	Reconnaissance	Survey
[215]	2022			x			x	Reconnaissance	Taxonomy
[213]	2023			x			x	Reconnaissance	Survey
[214]	2024			x			x	Reconnaissance	LLM Taxonomy

ATLAS: linguistically constrained and semantic preserving attacks. A defining characteristic of NLP adversarial research is the need to preserve grammatical correctness and semantic plausibility. AdvExpander [201] introduces linguistic expansion techniques that generate adversarial examples beyond simple word substitutions. Fake-news manipulation studies [202] demonstrate how fact-tampering attacks exploit linguistics-only models, motivating the integration of external knowledge sources. Token-level taxonomies such as Roth et al. [203] standardize evaluation across different perturbation strategies. Additionally, frameworks targeting programming-language models, such as CodeAttack [204], reveal vulnerabilities extending beyond natural language to structured code representations. These techniques align with ATLAS behaviors emphasizing stealthy manipulation and targeted influence while maintaining operational realism.

ATLAS: defensive countermeasures and robustness-oriented techniques. Defense strategies in NLP remain fragmented and task-specific. Randomized smoothing approaches such as the Dirichlet Neighborhood Ensemble (DNE) [205] improve robustness against substitution-based attacks by sampling embedding representations. Synonym Encoding Methods (SEM) [206] mitigate semantic attacks by mapping equivalent word clusters into unified representations. Robust training strategies including ASCC and RIFT [207] address catastrophic forgetting during adversarial training. Certified defenses using randomized smoothing [208] provide probabilistic guarantees against perturbations. Evaluation frameworks such as TextAttack [28] facilitate standardized benchmarking and reveal limitations in existing attack methodologies. Surveys and systematic reviews [209–217] further highlight the absence of unified evaluation protocols and the need for domain-aware defenses.

Overall, adversarial NLP research demonstrates a clear progression from simple word-level perturbations toward semantically aware, query-efficient, and large language model–driven attacks. Table 11 organizes the literature according to attacker knowledge assumptions, targeting strategies, and defensive mechanisms aligned with the unified threat modeling framework introduced earlier. The predominance of black-box and semantic-preserving attacks reflects realistic deployment constraints, while the diversity of defensive techniques underscores the ongoing challenge of achieving comprehensive robustness across NLP tasks. These findings emphasize the need for standardized evaluation frameworks and cross-domain threat modeling approaches capable of capturing both linguistic constraints and adversarial behavior patterns.

6.8. Finance

As financial institutions increasingly leverage AI for tasks such as fraud detection, algorithmic trading, and risk assessment, they are also becoming vulnerable to adversarial techniques that can manipulate models with subtle, often imperceptible modifications to input data. There are many key studies that explore the implications of adversarial attacks within the financial domain. By examining methods, vulnerabilities, and potential defenses, the evolving landscape of adversarial threats and the need for robust security measures in finance-driven AI applications are highlighted.

ATLAS: evasion and imperceptible manipulation in structured finance data. In Ballet et al. [219] the authors discussed the notion of adversarial examples in the tabular domain and they proposed a formalization based on the imperceptibility of attacks in the tabular domain leading to an approach to generate imperceptible adversarial examples. Their experiments showed that imperceptible adversarial examples can be generated with a high fooling rate. Schreyer et al. [220] presented an adversarial attack against Computer Assisted Audit Techniques (CAATs) using DNNs. To detect potential misstatements and fraud, international audit standards demand that auditors directly assess journal entries using CAATs. The authors introduced a real-world threat model designed to camouflage accounting anomalies such as fraudulent journal entries, they showed that adversarial autoencoder neural networks are capable

of learning a human interpretable model of journal entries that disentangles the entries latent generative factors and, finally, they demonstrated how such a model can be maliciously misused by a perpetrator to generate robust adversarial journal entries that mislead CAATs. In the domain of corporate mergers and acquisitions it is crucial to have a highly robust and accurate model and to be able to generate useful explanations for a user regarding automated system decisions [221]. To address these issues, this paper proposed a novel methodology for producing plausible counterfactual explanations, whilst exploring the regularization benefits of adversarial training on language models in the domain of FinTech.

ATLAS: black-box attacks and security evaluation for fraud detection pipelines. Kumar et al. [222] utilized two large, publicly available datasets for credit card fraud detection to benchmark the performance of various ML models. They also compared the effectiveness of different black-box attack strategies on the best-performing model. The authors introduced a novel gradient-free black-box attack method called Evolution-based Specialized Perturbations for Attacks (ESPA), which required significantly fewer queries than traditional black-box attack techniques. Their results demonstrated the efficiency and effectiveness of ESPA in generating adversarial samples. Xiao et al. [223] propose a black-box attack-based security evaluation framework for CCFD models. Under this framework, the semisupervised learning technique and transfer-based black-box attack were combined to construct two versions of a semisupervised transfer black-box attack algorithm. In addition they introduced a new nonlinear optimization model to generate the adversarial examples against CCFD models and a security evaluation index to quantitatively evaluate the security of them. Lee et al. experimented with detecting anomalous data in financial datasets [224]. Their experiment involved two stages and focused on classification performance and detecting noisy data, respectively. In the first stage, it was revealed that the accuracy of the classification on the noisy dataset could drop almost 15% compared to the ordinary dataset. In the second stage, local outlier factors and isolation forest algorithms were applied to detect the noisy data and they achieved detection rates of 95.156% and 84.1%, respectively. Finally, Tsai et al. [225] discussed Effective Adversarial Examples Identification of Credit Card Transactions. They used neuron activation status distribution and DNNs as detection tools. Their experiments employed three methods to generate adversarial examples, showcasing the effectiveness of the proposed detection approach.

ATLAS: reinforcement learning and sequential decision manipulation. Chen Y.-Y. and Chen [226] investigated adversarial attacks on reinforcement learning (RL)-based trading agents. They proposed an enhancement to the ensemble of identical independent evaluators (EIIE) method, called Enhanced EIIE, which incorporates information about the best bids and asks. This enhanced method demonstrated improved portfolio performance compared to the original EIIE agent. Enhanced EIIE was further employed in an adversarial agent to determine the optimal timing and magnitude of attacks, introducing perturbations to the trading process. Experimental results indicated that the proposed adversarial attack mechanism was 30% more effective in reducing the accumulated portfolio value compared to conventional attack strategies like the FGSM and iterative FGSM. Liu et al. [227] proposed an adaptive attack strategy called LCB-H, designed for black-box settings and applicable to most reinforcement learning agents. The authors proved that LCB-H could manipulate any efficient RL agent—whose dynamic regret grows sublinearly with the total number of steps—to frequently select actions dictated by the attacker's policy, incurring only a sublinear cost. [228] demonstrated that a "gray-box" approach for attacking a Deep RL-based trading agent is possible by trading in the same stock market, with no extra access to the trading agent. In their proposed approach, an adversary agent uses a hybrid DNN as its policy consisting of convolutional layers and fully-connected layers.

ATLAS: time-series, streaming inputs, and sentiment-driven decision support. Gallagher et al. [229] examined the FGSM attack, a novel Single Value attack and the Label Flip attack on a trending architecture, namely a 1-Dimensional CNN model used for time series classification. Their results showed that the architecture was susceptible to these attacks and that the classifier accuracy was significantly impacted. Nehemya et al. [230] explored a realistic scenario where attackers exploit adversarial learning techniques to influence algorithmic trading systems by manipulating input data streams in real time. The attacker devised a universal adversarial perturbation that was independent of the target model and unaffected by the timing of its application. This perturbation was designed to be imperceptible when added to the data stream. The proposed attack was validated on real-world market data, showcasing its potential impact. Xie et al. [231] experimented with a variety of adversarial attack configurations to fool three stock prediction victim models. They addressed the task of adversarial generation by solving combinatorial optimization problems with semantics and budget constraints. Their results showed that the proposed attack method could achieve consistent success rates and cause significant monetary loss in trading simulations by simply concatenating a perturbed but semantically similar tweet. [232] explored the susceptibility of financial sentiment analysis to adversarial attacks that manipulate financial texts. Lunghi et al. addressed the domain of fraud detection as a critical defense mechanism for the data economy, which poses several challenges for ML [233]. They described how attacks against fraud detection systems differ from other applications of adversarial ML, and proposed a number of directions to bridge this gap. Melo et al. in [234] proposed a new form of adversarial training where attacks are propagated between the two spaces in the training loop. Subsequently they tested their method empirically on a real world dataset in the domain of credit card fraud detection.

ATLAS: graph-based and networked financial systems (DeFi and GNN threats). Zhou et al. [235] introduced a common reference frame to systematically evaluate and compare DeFi incidents, including both attacks and accidents. They investigated academic papers, audit reports, and real-world incidents and their data revealed several gaps between academia and the practitioners' community. Yang, et al. [236] presented effective adversarial attack methods tailored to BWGNN. Using node injection as an adversarial attack method, they constructed surrogate models that closely resemble the structure of BWGNN, significantly enhancing the attack performance. Additionally, by incorporating dropout layers after the input layer of the surrogate model, they further enhanced the attack effectiveness.

ATLAS: defensive countermeasures and robustness-oriented techniques. Maung et al. [237] introduced a voting ensemble defense mechanism for black-box attacks, utilizing block-wise transformed images secured with secret keys. While key-based defenses have proven effective against gradient-based (white-box) attacks, they typically fail to counter gradient-free (black-box) attacks without relying on secret keys. To address this limitation, their proposed approach involved training multiple models on images transformed with varying secret keys and block sizes, followed by applying a voting ensemble across these models. Experimental evaluations demonstrated that their defense achieved a clean accuracy of 95.56% and reduced the attack success rate to below 9% under adversarial attacks with a noise distance of 8/255 on the CIFAR-10 dataset. Wang et al. in [238] proposed a defense method where they avoid improving the model's robustness and realize the defense against adversarial attacks based on denoising and reconstruction. Their method is a two-step defense framework. The first step denoises the input adversarial example, then reconstructs the sample to be close to the original clean sample and helps the target model output the original label. Zhang, et al. proposed an adaptively scaled adversarial training (ASAT) in time series analysis, by rescaling data at different time slots with adaptive scales [239]. Their experimental results showed that the proposed ASAT can improve both the generalization ability and

the adversarial robustness of neural networks compared to the baselines. In [240] Zhu et al. studied the impacts of deep learning technology in the field of payment security. As an important branch of deep learning, GANs play an important role in financial payment security, but they also face many challenges. In their paper, they explored the application status of deep learning in financial payment security, with a focus on the application and challenges of GANs models in payment security. They detailed the application of GANs in payment fraud detection, identity verification, anti-fraud, and user behavior analysis, and analyzed their advantages and limitations in solving payment security challenges. They also discussed security threats to GAN models and how to address these challenges to ensure the security and stability of financial payment systems. [241] analyzed how integrating AI, blockchain technology, and ML bolsters neobank defenses against current and future threats. Amerirad, et al. reviewed adversarial AI and discussed its implications for the insurance sector [242]. They presented a taxonomy of adversarial attacks and an original, fully fledged example of claims falsification in health insurance, as well as some remedies that are consistent with the current regulatory framework. Huang, et al. [243] discussed the impactful role of ChatGPT in the finance and banking sector.

ATLAS: indirect manipulation and forecast-stream interception. Finally, Cramer et al. [244] proposed a black-box attack on DSM and EPF based on an adversarial surrogate model that intercepts and modifies the data flow of load forecasts and forces the DSM to result in financial losses. Notably, adversaries can design the data modifications without knowledge of the EPF model or the DSM optimization model.

Table 12 summarizes the reviewed finance-domain literature according to the taxonomy introduced in Section 4, organizing works by application-driven categories that reflect how financial AI is deployed in practice. The first group captures attacks and vulnerabilities in core financial services (fraud detection, auditing, sentiment-driven decision support, and insurance), where adversaries often operate in black-box settings and aim to evade detection or manipulate downstream decisions. A second category focuses on reinforcement learning and sequential decision systems, highlighting that trading agents introduce temporally coupled vulnerabilities where small, well-timed perturbations can produce outsized effects on cumulative reward and portfolio value. The tabular and structured-data category reflects the dominant format of financial records and decision pipelines, emphasizing imperceptibility constraints and plausibility requirements that differentiate finance from perceptual domains. Time-series and forecasting studies capture adversarial risks in signal-based pipelines (e.g., forecasting and demand-side management), where attacks can be executed either directly on models or indirectly by intercepting and modifying forecast streams. Finally, graph-based and networked systems highlight emerging threats in decentralized and interconnected financial infrastructures, including DeFi incident analysis and graph-learning attacks. Across categories, defensive work remains comparatively fragmented, spanning ensemble voting, denoising/reconstruction, certified robustness and training-based approaches, indicating that robust financial AI requires both domain-aware threat modeling and defenses tailored to the structure and operational constraints of financial data and decision systems.

Table 12
Systematic taxonomy of adversarial attacks and defenses in financial AI systems aligned with MITRE ATLAS.

Title	Year	White-box	Black-box	Other	Targeted	Untargeted	Defense	ATLAS Tactic(s)	ATLAS Technique Anchor
Finance, Banking, Insurance, and Fraud Detection									
[220]	2019		x		x			Defense Evasion	Adversarial Example Generation
[222]	2021		x		x			Discovery	Query-based Optimization
[230]	2021		x		x			Impact	Universal Perturbation
[224]	2022		x		x		x	Mitigation	Anomaly Detection
[231]	2022		x		x			Defense Evasion	Text-driven Manipulation
[242]	2023			x			x	Reconnaissance	Taxonomy / Risk Analysis
[241]	2023			x			x	Mitigation	Secure System Design
[243]	2023			x				Reconnaissance	Emerging Threat Analysis
[232]	2023		x		x			Defense Evasion	NLP-based Manipulation
[233]	2023		x		x			Reconnaissance	Threat Modeling
[234]	2023		x		x		x	Mitigation	Adversarial Training
[223]	2023		x		x			Discovery	Transfer-based Black-box Attack
[225]	2024			x			x	Mitigation	Activation-based Detection
Reinforcement Learning and Sequential Decision Systems									
[226]	2021		x		x			Impact	Policy Manipulation
[227]	2021		x		x			Impact	Reward/Policy Shaping
[228]	2023		x		x			Discovery	Gray-box Interaction Attack
Tabular Data and Structured Domains									
[219]	2019		x		x			Defense Evasion	Imperceptible Perturbation
[221]	2020		x		x		x	Mitigation	Counterfactual Explanations
Time-Series, Forecasting, and Signal-Based Systems									
[229]	2022		x		x			Defense Evasion	Time-series Perturbation
[239]	2023			x	x		x	Mitigation	Robust Training (Scaling)
[244]	2024		x		x			Impact	Data Stream Interception
Graph-Based and Networked Systems									
[235]	2023			x				Reconnaissance	Incident Taxonomy
[236]	2024		x		x			Defense Evasion	Graph Node Injection
Defense Mechanisms									
[237]	2021			x		x	x	Mitigation	Ensemble Voting
[238]	2021			x			x	Mitigation	Denoising / Reconstruction
[240]	2024		x		x		x	Mitigation	GAN-based Robustness

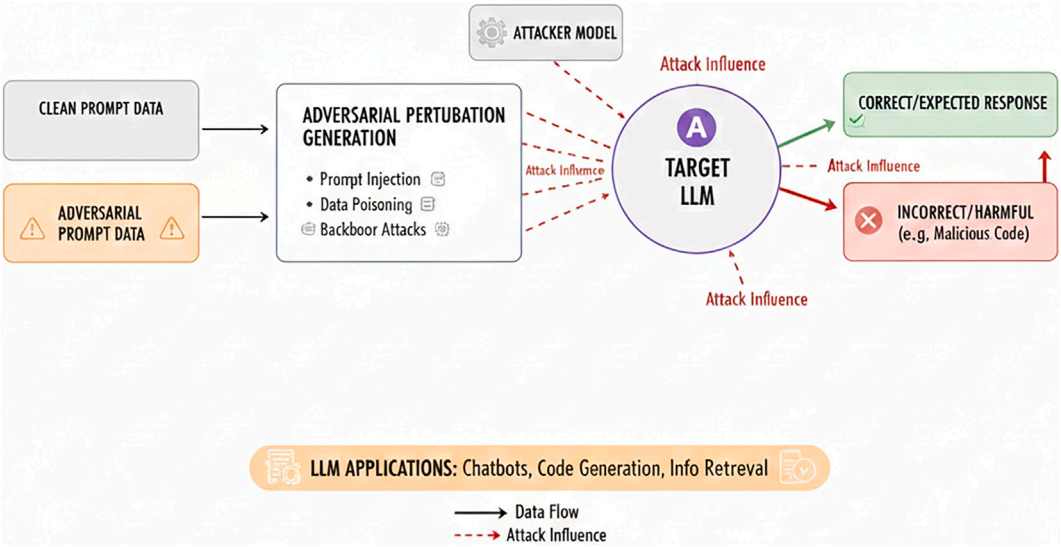

Fig. 8. Threat model for adversarial attacks on large language models (LLMs), showing how an attacker crafts adversarial prompts via perturbation mechanisms (e.g., prompt injection, data poisoning, backdoors) to influence a target LLM and steer outputs from correct/expected responses toward incorrect or harmful behavior (e.g., malicious code) across common LLM applications (chatbots, code generation, information retrieval).

6.9. Large language models

With the rapid advancement and widespread deployment of AI, adversarial attacks have been extensively studied across domains such as IoT, finance, healthcare, industrial control systems, and intrusion detection, as presented until now. These attacks exploit weaknesses in ML models by introducing carefully crafted perturbations that manipulate predictions, expose vulnerabilities, or compromise system reliability as mentioned above. While early research predominantly focused on traditional classifiers and domain-specific models, the rise of LLMs has introduced a new and significantly broader adversarial landscape. Due to their reliance on natural language instructions, contextual reasoning, and large-scale pretraining, LLMs exhibit unique failure modes that are fundamentally different from classical adversarial examples (Fig. 8). As a result, adversarial prompting, jailbreak attacks, training-time poisoning, and retrieval-based manipulation have emerged as critical new threat vectors, making the study of adversarial attacks on LLMs an essential extension of existing robustness research.

Adversarial prompting and jailbreak attacks (Inference-time manipulation). A principal line of research has focused on *adversarial prompting* and *jailbreak attacks*. Yang et al. [245] demonstrated that subtle adversarial phrasing of medical instructions can significantly degrade the reliability of clinical LLMs, resulting in unsafe or misleading diagnostic recommendations. Shu et al. [246] introduced *AttackEval*, one of the first standardized evaluation frameworks for jailbreak attacks, and showed that even highly aligned LLMs can be compromised by structured multi-step prompts. Their work also established a quantitative scoring methodology for measuring jailbreak severity across different safety categories. Zhu et al. [247] further showed that paraphrase-based adversarial prompts, contradiction embeddings, and role-inversion prompts can cause up to a 40% drop in reasoning accuracy across a variety of generative tasks. Beyond these studies, Mehrabi et al. [248] conducted one of the earliest holistic safety assessments of LLMs, revealing that adversarial prompts not only bypass safety filters but also induce harmful outputs that cross categories such as bias, misinformation, and unsafe advice. Qin et al. [249] expanded this work by proposing an automatic adversarial prompt generation method using reinforcement learning to explore the space of harmful instructions, consistently finding prompts that evade safety mechanisms across multiple LLM families.

Linguistic transformations and semantic deception (Robustness failures in reasoning). LLM vulnerabilities are also exposed by *semantic-preserving perturbations* that manipulate model decisions without altering human-perceived meaning. Morris et al. [28] demonstrated how adversarial linguistic transformations—synonym substitutions, grammatical modifications, and token-level perturbations—can mislead transformer models; these NLP attack principles later became foundational for analyzing semantic weaknesses that persist in LLM behavior. Recently, Li et al. [250] introduced *semantic illusion attacks*, where adversaries embed logically deceptive structures into prompts to mislead the model's reasoning chain, showing that the attack surface is not limited to surface-form perturbations but extends to higher-level logical manipulations.

Training-time poisoning and backdoor attacks (Persistent model compromise). A significant body of published work has investigated *poisoning and backdoor attacks* in LLMs. Alber et al. [251] showed that even small-scale data poisoning in medical corpora can induce systematic diagnostic failures. Yang et al. [245] further demonstrated that adversarial fine-tuning can shift model behavior toward harmful outputs without noticeably affecting standard benchmark performance. Zhang et al. [252] introduced *instruction backdoors*, showing that poisoning instruction–response pairs allows attackers to embed hidden trigger phrases that elicit controlled harmful outputs. Extending this idea, Chen et al. [253] proposed *multi-turn backdoors*, where the malicious trigger is spread across several conversational turns, making detection significantly more difficult. He et al. [254] demonstrated that backdoors can also be used for data exfiltration, enabling an attacker to retrieve sensitive training data via triggered prompts.

Retrieval-augmented generation (RAG) poisoning and indirect prompt injection. Poisoning attacks have expanded into *retrieval-augmented generation (RAG)* systems, where external documents or knowledge graphs influence LLM outputs. Zhao et al. [255] showed that injecting adversarial facts into RAG knowledge bases results in confident but incorrect responses. Similarly, Lermen et al. [256] demonstrated that poisoning document databases used by RAG systems can significantly distort

Table 13
Systematic taxonomy of adversarial threats and defenses in large language models, aligned with consistent survey columns and extended with MITRE ATLAS mappings (tactics and technique anchors).

Title	Year	White-box	Black-box	Other	Targeted	Untargeted	Defense	ATLAS Tactic(s)	ATLAS Technique Anchor
Adversarial Prompting and Jailbreak Attacks									
[245]	2025		x		x			Initial Access	Prompt Injection (AML.T0051)
[246]	2025		x	x	x		x	Initial Access	Prompt Injection (AML.T0051)
[247]	2024		x			x		Initial Access	Prompt Injection (AML.T0051)
[248]	2023		x	x	x		x	Initial Access	Prompt Injection (AML.T0051)
[249]	2024		x		x			Initial Access	Prompt Injection (AML.T0051)
[250]	2024		x		x			Defense Evasion	Prompt Injection (AML.T0051)
[259]	2023		x		x			Initial Access	Prompt Injection (AML.T0051)
Linguistic and Semantic Adversarial Attacks									
[28]	2020		x	x	x	x	x	Defense Evasion	Adversarial Perturbation (ATLAS)
Training-Time Poisoning and Backdoor Attacks									
[251]	2025	x			x			ML Attack Staging	Poison Training Data (AML.T0020)
[252]	2024	x			x			ML Attack Staging	Backdoor (ATLAS)
[253]	2024	x			x			ML Attack Staging	Backdoor (ATLAS)
[254]	2024	x	x		x			Exfiltration	Exfiltration via ML Inference API (ATLAS)
Poisoning Attacks in Retrieval-Augmented Generation (RAG)									
[255]	2025		x		x			Collection	RAG Database Retrieval (ATLAS)
[256]	2024		x		x			Collection	RAG Database Retrieval (ATLAS)
Defensive Mechanisms and Mitigations									
[257]	2025		x				x	Defense Evasion	Prompt Injection (AML.T0051)
[258]	2025		x				x	ML Attack Staging	Backdoor (ATLAS)
[260]	2024		x				x	Initial Access	Prompt Injection (AML.T0051)

fact-based reasoning and downstream classification tasks. These results reinforce that LLM security must account not only for the base model, but also for upstream retrieval and the integrity of external corpora that shape the final response.

Defensive countermeasures and evaluation frameworks. Defensive research is emerging but remains limited in scope. Ergün and Onan [257] proposed a supervised classifier for detecting adversarial prompts, achieving strong performance across adversarial categories. Yi et al. [258] introduced BEAT, a black-box defense for detecting backdoor activations via refusal-behavior analysis. Zou et al. [259] examined universal jailbreak attacks and provided partial mitigations using structured decoding constraints. Wang et al. [260] proposed an alignment-smoothing mechanism that reduces jailbreak susceptibility by modifying internal refusal logits. Despite these efforts, the literature consistently concludes that LLMs remain highly vulnerable to adversarial manipulation and that current defenses provide only partial protection.

Table 13 organizes the existing literature on adversarial attacks against large language models according to a structured taxonomy that reflects the evolving LLM threat landscape. The surveyed works show that most attacks operate in black-box settings and primarily exploit prompt-level manipulation, jailbreak strategies, and semantic deception, underscoring the fragility of alignment and safety mechanisms under carefully crafted natural language inputs. Beyond inference-time attacks, several studies demonstrate that training-time poisoning and instruction backdoors can embed persistent vulnerabilities that remain undetected under standard evaluation protocols. More recently, retrieval-augmented generation (RAG) systems have introduced additional attack surfaces, where poisoning external knowledge sources can systematically corrupt downstream reasoning and factual reliability. Defensive efforts remain comparatively limited and fragmented, focusing mainly on adversarial prompt detection, backdoor activation monitoring, and alignment-smoothing techniques, which currently provide only partial mitigation. Overall, this categorization highlights that adversarial threats to LLMs span the entire model lifecycle—from data collection and training to deployment and user interaction—emphasizing the urgent need for standardized evaluation frameworks, robust defenses, and domain-aware threat models as LLMs are increasingly deployed in safety-critical applications.

6.10. Cyber-physical versus pure IT adversarial threat models

While adversarial attacks appear across a wide range of application domains, the threat landscape differs fundamentally between cyber-physical systems (CPS) and purely digital IT environments. Domains such as smart grids, industrial control systems, and autonomous vehicles exhibit tightly coupled interactions between machine learning components, physical processes, sensing infrastructure, and control logic. In contrast, adversarial attacks in pure IT systems—such as NLP models, speech recognition platforms, financial analytics pipelines, and large language models—primarily operate within software-defined environments where constraints are informational rather than physical.

From the perspective of the unified adversarial threat modeling framework introduced earlier, these differences manifest across multiple dimensions. First, the *adversary model* is shaped by system observability and operational constraints. In pure IT systems, attackers frequently exploit high query accessibility, semantic manipulation, or surrogate-based transferability. Conversely, CPS environments impose physical feasibility requirements: adversarial perturbations must respect system dynamics, sensor noise bounds, protocol constraints, or control-loop stability. For example, in smart grid and ICS scenarios, malicious modifications to measurements or control signals must remain consistent with underlying physical laws to avoid triggering fail-safe mechanisms.

Second, the *attack stage* and lifecycle dynamics diverge significantly. Pure IT adversarial attacks often concentrate on inference-time evasion or prompt manipulation, whereas CPS adversaries frequently exploit multi-stage attack chains that combine reconnaissance, system probing, and gradual manipulation of state estimation or control processes. These attack chains align closely with MITRE ATLAS tactics such as model probing, evasion, and impact-oriented manipulation, highlighting how adversarial threats evolve differently when embedded within operational technology ecosystems.

Third, the *perturbation strategy* reflects modality-specific constraints. In vision or language models, perturbations can be optimized directly in

Table 14
Comparison of adversarial threat characteristics in cyber-physical versus pure IT machine learning systems.

Dimension	Cyber-Physical Systems (Smart Grid, ICS, AV)	Pure IT Systems (NLP, LLM, Finance, ASR)
Operational Context	Sensor-driven control loops, physical processes	Software-defined data pipelines
Adversary Constraints	Physics-aware, protocol-constrained perturbations	Semantic or feature-space manipulation
Attack Lifecycle	Multi-stage (probing → manipulation → impact)	Often inference-time or prompt-level attacks
Perturbation Design	Stealthy, physically feasible signal modifications	Gradient-based or semantic perturbations
Primary Impact	System instability, unsafe behavior, control disruption	Misclassification, data corruption, reasoning errors
Defense Requirements	System-level resilience, safety monitoring, redundancy	Robust training, detection, alignment mechanisms
MITRE ATLAS Alignment	Evasion, model manipulation, impact tactics	Prompt manipulation, model extraction, evasion

high-dimensional feature spaces. In CPS environments, however, adversarial inputs must propagate through sensing pipelines, communication protocols, and control feedback loops, creating additional constraints on perturbation magnitude, timing, and persistence. This leads to attack strategies that emphasize stealth, gradual drift, or physics-consistent manipulation rather than purely gradient-based optimization.

Fourth, the *intended effect* differs in operational scope. Pure IT attacks frequently target prediction accuracy or model outputs (e.g., misclassification, hallucination, or semantic manipulation), whereas CPS attacks aim to induce real-world consequences such as destabilizing control processes, degrading situational awareness, or causing unsafe system behavior, as illustrated Fig. 9. As a result, defense strategies must extend beyond model robustness to include system-level resilience, redundancy, and safety verification.

Table 14 summarizes the principal differences between cyber-physical and pure IT adversarial threat models. The comparison highlights how adversarial robustness cannot be treated as a purely algorithmic property; instead, it must be interpreted within the broader socio-technical context in which ML systems operate.

By explicitly distinguishing CPS from pure IT adversarial threat models, this analysis strengthens cross-domain understanding and clarifies why defenses developed for traditional ML tasks may not transfer directly to safety-critical cyber-physical environments. The comparison also reinforces the importance of structured threat frameworks such as MITRE ATLAS for capturing operational context, adversary capabilities, and lifecycle-dependent attack behavior across heterogeneous AI deployment scenarios.

Recent studies have also explored adversarial vulnerabilities from a frequency-domain perspective, showing that deep neural networks may exhibit sensitivity to specific spectral components. Instead of generating perturbations directly in the spatial domain, these approaches manipulate frequency representations to exploit vulnerable spectral regions of models. For example, recent works have proposed attacks that leverage non-overlapping vulnerable frequency regions across multiple models, mixed-frequency input transformations that combine high-frequency components from different images, and multimodal attacks that enhance perturbations through frequency-domain guidance [18]. These findings suggest that frequency-aware adversarial attacks represent an emerging research direction, particularly for improving transferability in realistic black-box settings.

7. Discussion

This survey highlights the landscape of adversarial attacks across multiple domains and the evolving approaches required to safeguard ML models against such threats. Beyond cataloging attack families, the evidence across domains consistently shows that (i) attacker knowledge and operational constraints shape what is feasible in practice, (ii) robustness is inseparable from deployment context (sensors, protocols, workflows), and (iii) defenses must be evaluated under realistic assumptions such as query limits, physical constraints, distribution shifts, and safety requirements. The following section discusses key takeaways, lessons learned, and future directions for advancing adversarial defense mechanisms.

7.1. Main remarks

Domain-specific threat surfaces and constraints. Adversarial attacks are context-sensitive, and each domain—such as IoT, healthcare, cybersecurity/IDS, autonomous vehicles, finance, and LLM-centric applications—reveals distinct vulnerability patterns driven by sensing modalities, data structure, and operational constraints. In IoT environments, black-box evasion and poisoning are particularly realistic due to limited model visibility, heterogeneous device stacks, and resource constraints that restrict heavyweight defenses. In healthcare, both white-box and black-box attacks can be high-impact: small perturbations in imaging or subtle

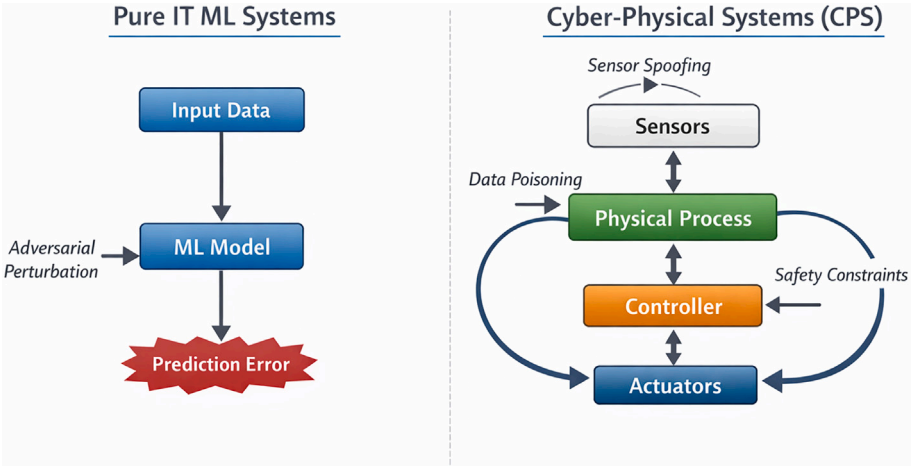

Fig. 9. Conceptual comparison between adversarial threat models in cyber-physical systems and pure IT environments. CPS attacks must satisfy physical constraints and influence control loops, while IT-domain attacks primarily manipulate data representations or semantic structures.

manipulations of clinical records may induce clinically meaningful errors, motivating robust validation and fail-safe operation. In ICS and cyber-physical environments, feasibility is governed by protocol compliance and physical constraints; adversaries may prioritize stealth and persistence, while defenders must ensure detection and control stability under real-time requirements.

Defense design must match the deployment objective. Although certain defense primitives reappear across domains (e.g., adversarial training, input filtering, detection, and ensemble checks), the survey indicates that effective protection is rarely achieved by a single universal method. Instead, defenses should be selected according to what the system must optimize: (i) *availability* (e.g., resilience to disruption and false positives in IDS/ICS), (ii) *integrity* (e.g., preventing targeted manipulation in finance and healthcare), and (iii) *safety* (e.g., perception reliability in autonomous systems). Consequently, industrial/ICS pipelines often favor anomaly detection and physics- or protocol-aware validation, financial tabular pipelines emphasize plausibility constraints and robust feature handling, and ASR pipelines frequently rely on input transformations and denoising that preserve intelligibility.

Feature-aware adversarial training. Recent advances in adversarial defense research emphasize the importance of learning robust semantic features rather than relying solely on perturbation-based robustness. One recent example is Feature-Focusing Adversarial Training (F2AT), which introduces a training strategy that disentangles adversarial examples into natural and perturbed patterns using bit-plane slicing. By separating semantically meaningful information from perturbation-related artifacts, F2AT encourages models to focus on core features associated with natural patterns while reducing the influence of adversarial noise. This feature-aware learning process improves the trade-off between clean accuracy and adversarial robustness and highlights a promising direction for designing defenses that maintain model performance while increasing resilience against adversarial manipulation [261].

Attacker knowledge is increasingly varied, not binary. Across domains, the practical boundary between white-box and black-box settings is blurred by surrogate models, transferability, partial leakage (e.g., confidence scores or explanations), and repeated querying. This motivates evaluation protocols that explicitly incorporate *query budgets*, *surrogate diversity*, and *adaptive attackers*. In particular, cybersecurity and LLM deployments commonly exhibit black-box interfaces, but attackers can still achieve strong outcomes via transfer attacks, prompt manipulation, or poisoning of upstream data and retrieval sources.

Cross-domain synthesis through MITRE ATLAS. A key outcome of this survey is that the MITRE ATLAS framework provides a practical unifying perspective for interpreting adversarial threats across heterogeneous domains. Although implementation details differ significantly between IoT, healthcare, cyber-physical systems, NLP pipelines, and LLM deployments, many observed attacks align with recurring ATLAS tactics such as *inference-time evasion*, *model probing and extraction*, *training-time poisoning*, and *capability misuse via input manipulation*. Mapping domain-specific studies onto ATLAS highlights that adversarial strategies are less domain-isolated than often assumed; instead, similar operational objectives reappear under different sensing modalities and deployment contexts. This cross-domain alignment is further clarified by the comparative analysis between cyber-physical and pure IT adversarial threat models introduced in Section 6.10. That analysis demonstrates that while ATLAS tactics remain consistent across domains, their operational manifestations differ significantly depending on whether attacks must satisfy physical constraints and control-loop dynamics (e.g., smart grids, ICS, autonomous systems) or operate purely within software-defined environments such as NLP, finance, and LLM applications.

Robustness–performance trade-offs remain a core barrier. A recurring lesson is the trade-off between robustness and operational overhead. Strong defenses can introduce latency, extra computation, calibration drift, or accuracy degradation on clean inputs—especially problematic in resource-limited IoT/edge deployments and real-time CPS environments. This suggests a practical shift toward *selective prediction*, *risk-aware abstention*, and *lightweight detection* as complements to heavy retraining-based approaches. Future work should prioritize defenses that are deployable: bounded overhead, stable calibration, and measurable gains under realistic threat models.

Implications beyond accuracy: trust, compliance, and safety. Adversarial vulnerabilities are not only technical failure modes; they affect user trust, safety assurance, auditability, and regulatory compliance. In healthcare and finance, failures can translate into direct harm, privacy violations, and legal exposure, requiring defenses that support traceability, human oversight, and conservative failure behavior (e.g., abstain/escalate policies). In LLM settings, risks extend across the lifecycle—from prompt-time manipulation to training-time poisoning and RAG corpus integrity—highlighting that model robustness must include data governance and secure deployment practices.

Table 15
Holistic comparison of adversarial attacks and defenses across domains, synthesizing predominant threat settings, defense priorities, operational constraints, emerging research directions, and dominant ATLAS tactics.

Domain	Predominant Threat Setting	Defense Focus	Attack Goal	Key Challenges	Future Work	Dominant ATLAS Tactic(s)
Internet of Things	Black-box evasion/poisoning	Lightweight detection	Targeted evasion	Resource limits, latency	Edge-friendly robustness	Inference-time evasion, Model probing
Healthcare	White-box & black-box	Validation + robust training	Targeted manipulation	Safety-critical decisions	Federated robustness	Targeted evasion, Data poisoning
Cybersecurity / IDS	Black-box adaptive attacks	Robust detection	Evasion/disruption	Concept drift	Multi-modal telemetry	Evasion, Model extraction
Industrial Control Systems	Protocol-constrained attacks	Physics-aware defense	Stealth manipulation	Legacy systems	Secure sensing/control	Stealthy evasion, Signal manipulation
Autonomous Vehicles	Physical-world attacks	Sensor fusion validation	Safety degradation	Multi-sensor coupling	Fail-operational perception	Perception evasion
Speech Recognition	Query-efficient attacks	Input transformation	Command injection	Imperceptibility constraints	Hybrid defenses	Input manipulation
Natural Language Processing	Black-box semantic attacks	Robust training	Label flipping	Discrete input space	Certified robustness	Semantic manipulation
Finance	Black-box tabular attacks	Anomaly detection	Decision manipulation	Economic incentives	Secure pipelines	Data poisoning
Large Language Models	Prompt injection/jailbreak	Alignment monitoring	Policy bypass	Lifecycle vulnerabilities	Secure RAG/guardrails	Prompt injection, Capability abuse

Table 15 provides a comparative overview of adversarial threats across domains, emphasizing that security posture emerges from the interaction between attacker capability, deployment context, and operational constraints. A consistent observation is that domains with direct economic or safety incentives attract targeted manipulation, while large-scale deployed systems face adaptive black-box evasion. Framing adversarial threats through ATLAS-aligned tactics emphasizes that robustness must be evaluated across the entire AI lifecycle, including training data, model interfaces, deployment pipelines, and human–AI interaction layers.

7.2. Lessons learned

This survey of adversarial attacks across multiple domains reveals several critical insights regarding both adversarial behavior and the design of effective defensive strategies. A central observation is the recurrence of fundamental attack families across heterogeneous application areas. Techniques such as inference-time evasion, data poisoning, model extraction, and manipulation of input representations appear consistently in domains including IoT, healthcare, autonomous systems, cybersecurity, finance, and large language models. The persistence of these shared attack patterns indicates that many adversarial vulnerabilities originate from structural properties of machine learning systems rather than domain-specific implementation details, highlighting the importance of foundational security principles that can generalize across deployment contexts.

At the same time, adversarial behavior exhibits strong domain dependence shaped by operational constraints, sensing modalities, and system architecture. A particularly important distinction emerges between cyber-physical systems (CPS) and purely digital IT environments. In CPS domains such as smart grids, industrial control systems, autonomous vehicles, and certain IoT infrastructures, adversarial actions must satisfy physical feasibility constraints, temporal dependencies, and control-loop stability requirements. Consequently, attacks often exploit sensor manipulation, environmental perturbations, or protocol-aware strategies while remaining bounded by real-world dynamics. In contrast, adversarial attacks in purely digital domains—such as natural language processing, financial decision systems, or LLM interfaces—primarily operate within semantic or informational spaces, where adversaries leverage linguistic manipulation, query-based probing, or data-level poisoning without physical-world constraints. Recognizing this distinction is essential for accurate threat modeling and explains why similar adversarial objectives manifest differently across domains despite sharing underlying tactical patterns.

Another key lesson concerns the increasing sophistication and contextual awareness of adversarial strategies. Modern attacks frequently incorporate domain knowledge, adaptive querying, and transferability techniques to bypass traditional defenses. For example, healthcare-focused attacks exploit subtle medical imaging characteristics, while attacks against autonomous systems leverage environmental cues or perception pipeline weaknesses. This evolution indicates that defenses must move beyond static protection mechanisms toward adaptive, risk-aware approaches capable of responding to evolving adversarial behavior.

The survey also emphasizes the persistent trade-off between robustness and operational performance. While robust training, ensemble methods, or detection frameworks can significantly improve resilience, they often introduce additional computational overhead, latency, or calibration challenges. These trade-offs are particularly pronounced in resource-constrained environments such as edge devices or real-time cyber-physical systems, where safety and availability requirements impose strict operational limits. Consequently, practical defense strategies increasingly combine lightweight detection, selective prediction, and uncertainty-aware abstention to balance robustness with deployability.

Finally, adversarial vulnerabilities extend beyond technical accuracy metrics, affecting trustworthiness, regulatory compliance, and system safety. In high-stakes domains such as healthcare, finance, and critical infrastructure, adversarial failures may translate into safety risks, privacy breaches, or economic consequences. Effective adversarial defense must therefore integrate technical robustness with governance considerations, including transparency, auditability, and lifecycle security practices. Together, these lessons reinforce the need for cross-domain threat modeling frameworks—such as MITRE ATLAS—that support structured evaluation and consistent interpretation of adversarial risks across diverse AI applications.

7.3. Future directions

A central direction emerging from this survey is the development of cross-domain, foundational defense strategies that transcend individual application areas. Many adversarial behaviors — including inference-time evasion, training-time poisoning, model probing, and input manipulation — recur across domains despite differences in sensing modalities or deployment constraints. Mapping these behaviors onto structured frameworks such as MITRE ATLAS suggests that future

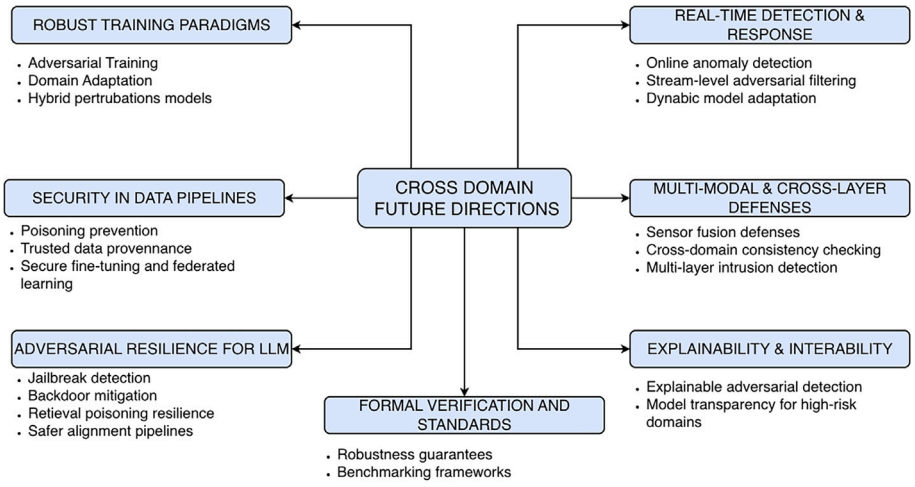

Fig. 10. Conceptual framework outlining key future research directions in adversarial machine learning, spanning robustness, detection, explainability, and cross-domain defenses.

defenses should target underlying adversarial tactics rather than isolated attack implementations. **As illustrated in Fig. 10, foundational techniques lie at the center of future cross-domain research directions, serving as a unifying layer that connects robust training, detection, explainability, and data security.** Such defense layers should be modular and adaptable, enabling domain-specific customization while maintaining consistent protection against shared adversarial objectives.

Another critical research direction involves adaptive and real-time defense mechanisms. As adversarial strategies increasingly exploit contextual knowledge and dynamic system behavior, static defenses are insufficient. Future work should emphasize proactive protection through online monitoring, continuous risk assessment, and runtime model adaptation. **This aligns with the *Real-Time Detection & Response* branch in Fig. 10, highlighting the need for online anomaly detection, dynamic model updates, and response-aware architectures.** Integrating adaptive defenses is particularly important for safety-critical domains such as autonomous vehicles, cyber-physical systems, and financial decision pipelines, where delayed detection may lead to irreversible consequences.

Resource-efficient robustness remains a major open challenge, especially in edge environments such as IoT and distributed sensing systems. Traditional adversarial defenses often introduce computational overhead that conflicts with real-world deployment constraints. Consequently, future research should prioritize lightweight methods including efficient adversarial training, compressed or distilled defensive models, and low-latency anomaly detection pipelines. **These directions resonate with the *Security in Data Pipelines* and *Robust Training Paradigms* components of Fig. 10, emphasizing trustworthy data processing and optimized training procedures compatible with constrained environments.**

In parallel with defense development, designing novel adversarial attack methodologies remains essential for advancing robustness research. New attack paradigms reveal hidden model vulnerabilities and drive progress in defense mechanisms. Future work should explore multi-stage attacks, hybrid threat models combining physical and digital manipulation, and lifecycle-aware adversarial strategies aligned with ATLAS threat categories. **This is reflected in Fig. 10 through interconnected research branches demonstrating how advances in attack development directly inform defense innovation, verification strategies, and secure deployment practices.** Proactively studying emerging adversarial techniques will enable researchers to anticipate evolving threat patterns and design more resilient systems.

As LLMs and multimodal AI systems continue to expand across application domains, their unique adversarial risks require specialized investigation. Unlike traditional adversarial examples, LLM vulnerabilities include prompt injection, jailbreak attacks, semantic manipulation, training-time backdoors, and retrieval-augmented generation (RAG) poisoning. Future research should focus on lifecycle security approaches encompassing data governance, alignment robustness, retrieval integrity, and safe interaction protocols. **The *Adversarial Resilience for LLMs* branch in Fig. 10 highlights these priorities, including jailbreak detection, secure retrieval pipelines, and alignment-aware training strategies.**

Ethical, regulatory, and governance considerations will also play a decisive role in shaping future adversarial defense research. As adversarial vulnerabilities increasingly impact healthcare, finance, infrastructure, and public safety systems, robustness must extend beyond technical accuracy to include transparency, accountability, and compliance. Future work should explore standardized evaluation benchmarks, explainable defense mechanisms, and certification-oriented robustness frameworks aligned with emerging regulatory landscapes. **This corresponds to the *Explainability & Interoperability* and *Formal Verification & Standards* domains in Fig. 10, which emphasize trustworthy AI deployment and auditable decision-making.**

Finally, improving the understanding of attack transferability and defense generalization remains a key open problem. Investigating why certain adversarial strategies generalize across models and domains — and how defenses can achieve similar generalization — is essential for building resilient AI systems. Structured threat modeling using ATLAS-style lifecycle categorization provides a promising direction for unifying evaluation across domains. **Fig. 10 reinforces this cross-domain perspective by linking multiple research pillars through a central hub, illustrating how advancements in one area can strengthen resilience across others.** Advancing knowledge in these areas will be instrumental in building ML systems capable of withstanding evolving adversarial threats in realistic deployment environments.

8. Conclusions

This survey emphasizes the crucial and escalating problem of adversarial attacks across a variety of ML applications, including IoT, healthcare, finance, and autonomous systems. By exploring prevalent attack methods such as evasion and data poisoning, as well as defenses specific to each domain, the study highlights the necessity for adaptable security measures. Many adversarial attacks take advantage of fundamental weaknesses that are common across models, necessitating core defenses that can be adapted to the unique requirements of each application. Although existing defenses like adversarial training and anomaly detection provide certain protections, they often fall short in terms of the flexibility required to combat rapidly evolving threats. For the future, creating cross-domain frameworks and lightweight defenses suitable for constrained environments will be crucial to developing robust systems. This survey acts as a foundational resource for enhancing adversarial defenses, ultimately aiding the secure and ethical implementation of ML systems in high-stakes and resource-sensitive fields.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

[1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv preprint arXiv:1312.6199, 2013.
[2] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, A.K. Jain, Adversarial attacks and defenses in images, graphs and text: a review, Int. J. Autom. Comput. 17 (2020) 151–178.
[3] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, D. Mukhopadhyay, A survey on adversarial attacks and defences, CAAI Trans. Intell. Technol. 6 (1) (2021) 25–45.
[4] N. Akhtar, A. Mian, N. Kardan, M. Shah, Advances in adversarial attacks and defenses in computer vision: a survey, IEEE Access 9 (2021) 155161–155196.
[5] Y. Li, M. Cheng, C.-J. Hsieh, T.C. Lee, A review of adversarial attack and defense for classification methods, Am. Stat. 76 (4) (2022) 329–345.
[6] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, D. Mukhopadhyay, Adversarial attacks and defences: A survey, arXiv preprint arXiv:1810.00069, 2018.
[7] N. Akhtar, A. Mian, Threat of adversarial attacks on deep learning in computer vision: a survey, IEEE Access 6 (2018) 14410–14430.
[8] Z. Kong, J. Xue, Y. Wang, L. Huang, Z. Niu, F. Li, A survey on adversarial attack in the age of artificial intelligence, Wirel. Commun. Mob. Comput. 2021 (1) (2021) 4907754.
[9] T. Long, Q. Gao, L. Xu, Z. Zhou, A survey on adversarial attacks in computer vision: taxonomy, visualization and future directions, Comput. Secur. 121 (2022) 102847.
[10] S.Y. Khamaiseh, D. Bagagem, A. Al-Alaj, M. Mancino, H.W. Alomari, Adversarial deep learning: a survey on adversarial attacks and defense mechanisms on image classification, IEEE Access 10 (2022) 102266–102291.
[11] E. Shayegani, M.A.A. Mamun al, Y. Fu, P. Zaree, Y. Dong, N. Abu-Ghazaleh, Survey of vulnerabilities in large language models revealed by adversarial attacks, arXiv preprint arXiv:2310.10844, 2023.
[12] S. Yan, J. Ren, W. Wang, L. Sun, W. Zhang, Q. Yu, A survey of adversarial attack and defense methods for malware classification in cyber security, IEEE Commun. Surv. Tutorials 25 (1) (2022) 467–496.
[13] D. Wang, W. Yao, T. Jiang, G. Tang, X. Chen, A survey on physical adversarial attack in computer vision, arXiv preprint arXiv:2209.14262, 2022.

[14] K. He, D.D. Kim, M.R. Asghar, Adversarial machine learning for network intrusion detection systems: a comprehensive survey, IEEE Commun. Surv. Tutorials 25 (1) (2023) 538–566.

[15] Z. Zhang, M. Liu, M. Sun, R. Deng, P. Cheng, D. Niyato, M.-Y. Chow, J. Chen, Vulnerability of machine learning approaches applied in iot-based smart grid: a review, IEEE Internet Things J. 11 (11) (2024) 18951–18975, https://doi.org/10.1109/JIOT.2024.3349381

[16] Y. Qian, S. He, C. Zhao, J. Sha, W. Wang, B. Wang, Lea2: a lightweight ensemble adversarial attack via non-overlapping vulnerable frequency regions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4510–4521.

[17] Y. Qian, K. Chen, B. Wang, Z. Gu, S. Ji, W. Wang, Y. Zhang, Enhancing transferability of adversarial examples through mixed-frequency inputs, IEEE Trans. Inf. Forensics Secur. 19 (2024) 7633–7645, https://doi.org/10.1109/TIFS.2024.3430508

[18] Y. Qian, Q. Yu, Q. Bao, S. Ji, W. Wang, B. Wang, Z. Gu, Z. Lei, A multimodal adversarial attack method via frequency domain enhancement and fine-grained cross-modal guidance, IEEE Trans. Dependable Secure Comput. 22 (6) (2025) 7877–7889, https://doi.org/10.1109/TDSC.2025.3601232

[19] L. Ye, S.M. Hamidi, Thundernna: a white box adversarial attack, arXiv preprint arXiv:2111.12305, 2021.

[20] S. Agnihotri, S. Jung, M. Keuper, Cospgd: a unified white-box adversarial attack for pixel-wise prediction tasks, arXiv preprint arXiv:2302.02213, 2023.

[21] M.A. Ayub, W.A. Johnson, D.A. Talbert, A. Siraj, Model evasion attack on intrusion detection systems using adversarial machine learning, in: 2020 54th Annual Conference on Information Sciences and Systems (CISS), IEEE, 2020, pp. 1–6.

[22] G. Apruzzese, M. Conti, Y. Yuan, Spacephish: the evasion-space of adversarial attacks against phishing website detectors using machine learning, in: Proceedings of the 38th Annual Computer Security Applications Conference, 2022, pp. 171–185.

[23] S. Wu, J. Xue, Y. Wang, Z. Kong, Black-box evasion attack method based on confidence score of benign samples, Electronics 12 (11) (2023) 2346.

[24] H. Qiu, L.L. Custode, G. Iacca, Black-box adversarial attacks using evolution strategies, in: Proceedings of the Genetic and Evolutionary Computation Conference Companion, 2021, pp. 1827–1833.

[25] X. Wei, Y. Guo, B. Li, Black-box adversarial attacks by manipulating image attributes, Inf. Sci. 550 (2021) 285–296.

[26] G. Apruzzese, M. Andreolini, L. Ferretti, M. Marchetti, M. Colajanni, Modeling realistic adversarial attacks against network intrusion detection systems, Digit. Threats Res. Pract. 3 (3) (2022) 1–19.

[27] E. Alshahrani, D. Alghazzawi, R. Alotaibi, O. Rabie, Adversarial attacks against supervised machine learning based network intrusion detection systems, PLoS One 17 (10) (2022) e0275971.

[28] J. Morris, E. Lifland, J.Y. Yoo, J. Grigsby, D. Jin, Y. Qi, Textattack: a framework for adversarial attacks, data augmentation, and adversarial training in NLP, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020, pp. 119–126.

[29] S. Ghamizi, M. Cordy, M. Papadakis, Y. Le Traon, Evasion attack steganography: turning vulnerability of machine learning to adversarial attacks into a real-world application, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 31–40.

[30] D.C. Asimopoulos, P. Radoglou-Grammatikis, P. Fouliras, K. Panitsidis, G. Efstathopoulos, T. Lagkas, V. Argyriou, I. Kotsiuba, P. Sarigiannidis, Surrogate-guided adversarial attacks: enabling white-box methods in black-box scenarios, in: 2025 IEEE International Conference on Cyber Security and Resilience (CSR), IEEE, 2025, pp. 950–956.

[31] H. Qiu, T. Dong, T. Zhang, J. Lu, G. Memmi, M. Qiu, Adversarial attacks against network intrusion detection in IOT systems, IEEE Internet Things J. 8 (13) (2020) 10327–10335.

[32] X. Zhou, W. Liang, W. Li, K. Yan, S. Shimizu, I. Kevin, K. Wang, Hierarchical adversarial attacks against graph-neural-network-based IOT network intrusion detection system, IEEE Internet Things J. 9 (12) (2021) 9310–9319.

[33] P. Papadopoulos, O. Thornewill von Essen, N. Pitropakis, C. Chrysoulas, A. Mylonas, W.J. Buchanan, Launching adversarial attacks against network intrusion detection systems for IOT, J. Cybersecur. Priv. 1 (2) (2021) 252–273.

[34] H. Du, Q. Wen, S. Zhang, M. Gao, A new provably secure certificateless signature scheme for internet of things, Ad Hoc Netw. 100 (2020) 102074.

[35] Z. Bao, Y. Lin, S. Zhang, Z. Li, S. Mao, Threat of adversarial attacks on dl-based IOT device identification, IEEE Internet Things J. 9 (11) (2021) 9012–9024.

[36] J. Tian, B. Wang, R. Guo, Z. Wang, K. Cao, X. Wang, Adversarial attacks and defenses for deep-learning-based unmanned aerial vehicles, IEEE Internet Things J. 9 (22) (2021) 22399–22409.

[37] A. Raja, L. Njilla, J. Yuan, Adversarial attacks and defenses toward ai-assisted UAV infrastructure inspection, IEEE Internet Things J. 9 (23) (2022) 23379–23389.

[38] X. Ding, S. Zhang, M. Song, X. Ding, F. Li, Toward invisible adversarial examples against Dnn-based privacy leakage for internet of things, IEEE Internet Things J. 8 (2) (2020) 802–812.

[39] R. Yang, J. Ma, J. Zhang, S. Kumari, S. Kumar, J.J. Rodrigues, Practical feature inference attack in vertical federated learning during prediction in artificial internet of things, IEEE Internet Things J. 11 (1) (2023) 5–16.

[40] Z. Chen, A. Fu, Y. Zhang, Z. Liu, F. Zeng, R.H. Deng, Secure collaborative deep learning against GAN attacks in the internet of things, IEEE Internet Things J. 8 (7) (2020) 5839–5849.

[41] M.A. Ferrag, O. Friha, L. Maglaras, H. Janicke, L. Shu, Federated deep learning for cyber security in the internet of things: concepts, applications, and experimental analysis, IEEE Access 9 (2021) 138509–138542.

[42] G. Li, J. Wu, S. Li, W. Yang, C. Li, Multitentacle federated learning over software-defined industrial internet of things against adaptive poisoning attacks, IEEE Trans. Ind. Inform. 19 (2) (2022) 1260–1269.

[43] C. Dunn, N. Moustafa, B. Turnbull, Robustness evaluations of sustainable machine learning models against data poisoning attacks in the internet of things, Sustainability 12 (16) (2020) 6434.

[44] M. Liu, H. Zhang, Z. Liu, N. Zhao, Attacking spectrum sensing with adversarial deep learning in cognitive radio-enabled internet of things, IEEE Trans. Rel. 72 (2) (2022) 431–444.

[45] Y. Qian, Y. Guo, Q. Shao, J. Wang, B. Wang, Z. Gu, X. Ling, C. Wu, Ei-mtd: moving target defense for edge intelligence against adversarial attacks, ACM Trans. Priv. Secur. 25 (3) (2022) 1–24.

[46] X. Fu, N. Zhou, L. Jiao, H. Li, J. Zhang, The robust deep learning–based schemes for intrusion detection in internet of things environments, Ann. Telecommun. 76 (5) (2021) 273–285.

[47] H. Jiang, J. Lin, H. Kang, Fgmd: a robust detector against adversarial attacks in the IOT network, Futur. Gener. Comput. Syst. 132 (2022) 194–210.

[48] M.M. Rashid, J. Kamruzzaman, M.M. Hassan, T. Imam, S. Wibowo, S. Gordon, G. Fortino, Adversarial training for deep learning-based cyberattack detection in iot-based smart city applications, Comput. Secur. 120 (2022) 102783.

[49] E. Anthi, L. Williams, M. Rhode, P. Burnap, A. Wedgbury, Adversarial attacks on machine learning cybersecurity defences in industrial control systems, J. Inf. Secur. Appl. 58 (2021) 102717.

[50] L. Nie, Y. Wu, X. Wang, L. Guo, G. Wang, X. Gao, S. Li, Intrusion detection for secure social internet of things based on collaborative edge computing: a generative adversarial network-based approach, IEEE Trans. Comput. Soc. Syst. 9 (1) (2021) 134–145.

[51] Y. Wu, L. Nie, S. Wang, Z. Ning, S. Li, Intelligent intrusion detection for Internet of Things security: a deep convolutional generative adversarial network-enabled approach, IEEE Internet Things J. 10 (4) (2021) 3094–3106.

[52] I. Idrissi, M. Azizi, O. Moussaoui, An unsupervised generative adversarial network based-host intrusion detection system for Internet of things devices, Indones. J. Electr. Eng. Comput. Sci. 25 (2) (2022) 1140–1150.

[53] M.M. Hassan, M.R. Hassan, S. Huda, V.H.C. De Albuquerque, A robust deep-learning-enabled trust-boundary protection for adversarial industrial IOT environment, IEEE Internet Things J. 8 (12) (2020) 9611–9621.

[54] C. Qian, W. Yu, C. Lu, D. Griffith, N. Golmie, Toward generative adversarial networks for the industrial internet of things, IEEE Internet Things J. 9 (19) (2022) 19147–19159.

[55] M.A. Ferrag, D. Hamouda, M. Debbah, L. Maglaras, A. Lakas, Generative adversarial networks-driven cyber threat intelligence detection framework for securing internet of things, in: 2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT), IEEE, 2023, pp. 196–200.

[56] H. Benaddi, M. Jouhari, K. Ibrahimi, A. Benslimane, E.M. Amhoud, Adversarial attacks against IOT networks using conditional GAN based learning, in: GLOBECOM 2022-2022 IEEE Global Communications Conference, IEEE, 2022, pp. 2788–2793.

[57] S. Nayak, N. Ahmed, S. Misra, Deep learning-based reliable routing attack detection mechanism for industrial internet of things, Ad Hoc Netw. 123 (2021) 102661.

[58] R. Yumlembam, B. Issac, S.M. Jacob, L. Yang, Iot-based Android malware detection using graph neural network with adversarial defense, IEEE Internet Things J. 10 (10) (2022) 8432–8444.

[59] R. Chaganti, V. Ravi, T.D. Pham, Deep learning based cross architecture internet of things malware detection and classification, Comput. Secur. 120 (2022) 102779.

[60] R.K. Shrivastava, S.P. Singh, M.K. Hasan, S. Islam, S. Abdullah, A.H.M. Aman, et al., Securing Internet of Things devices against code tampering attacks using return oriented programming, Comput. Commun. 193 (2022) 38–46.

[61] X. Hao, W. Ren, R. Xiong, T. Zhu, K.-K.R. Choo, Asymmetric cryptographic functions based on generative adversarial neural networks for internet of things, Futur. Gener. Comput. Syst. 124 (2021) 243–253.

[62] F. Hu, W. Zhou, K. Liao, H. Li, D. Tong, Toward federated Learning models resistant to adversarial attacks, IEEE Internet Things J. 10 (19) (2023) 16917–16930.

[63] X. Liu, W. Yu, F. Liang, D. Griffith, N. Golmie, On deep reinforcement learning security for industrial internet of things, Comput. Commun. 168 (2021) 20–32.

[64] A. Singh, B. Sikdar, Adversarial attack and defence strategies for deep-learning-based IOT device classification techniques, IEEE Internet Things J. 9 (4) (2021) 2602–2613.

[65] X.-H. Nguyen, K.-H. Le, Robust detection of unknown Dos/ddos attacks in IOT networks using a hybrid learning model, Internet of Things 23 (2023) 100851.

[66] M.A. Ferrag, O. Friha, B. Kantarci, N. Tihanyi, L. Cordeiro, M. Debbah, D. Hamouda, M. Al-Hawawreh, K.-K.R. Choo, Edge learning for 6G-enabled internet of things: a comprehensive survey of vulnerabilities, datasets, and defenses, IEEE Commun. Surv. Tutor. 25 (4) (2023) 2654–2713.

[67] S. Zeadally, J.T. Isaac, Z. Baig, Security attacks and solutions in electronic health (e-health) systems, J. Med. Syst. 40 (2016) 1–12.

[68] T. Kanwal, A. Anjum, S.U. Malik, A. Khan, M.A. Khan, Privacy preservation of electronic health records with adversarial attacks identification in hybrid cloud, Comput. Stand. Interfaces 78 (2021) 103522.

[69] A.P. Kalapaaking, I. Khalil, X. Yi, Blockchain-based federated learning with smpc model verification against poisoning attack for healthcare systems, IEEE Trans. Emerg. Top. Comput. 12 (1) (2023) 269–280.

[70] I. Siniosoglou, P. Sarigiannidis, V. Argyriou, T. Lagkas, S.K. Goudos, M. Poveda, Federated intrusion detection in ng-Iot healthcare systems: an adversarial approach, in: ICC 2021-IEEE International Conference on Communications, IEEE, 2021, pp. 1–6.

[71] S. Ravikumar, S. Tasneem, N. Sakib, K.A. Islam, Securing AI of healthcare: a selective review on identifying and preventing adversarial attacks, in: 2024 IEEE Opportunity Research Scholars Symposium (ORSS), IEEE, 2024, pp. 75–78.

[72] H. Kim, D.C. Jung, B.W. Choi, Exploiting the vulnerability of deep learning-based artificial intelligence models in medical imaging: adversarial attacks, J. Korean Soc. Radiol. 80 (2) (2019) 259–273.

[73] F. Hussain, R. Ksantini, M. Hammad, A review of malicious altering healthcare imagery using artificial intelligence, in: 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), IEEE, 2021, pp. 646–651.

[74] S. Kaviani, K.J. Han, I. Sohn, Adversarial attacks and defenses on AI in medical imaging Informatics: a survey, Expert Syst. Appl. 198 (2022) 116815.

[75] J. Dong, J. Chen, X. Xie, J. Lai, H. Chen, Survey on adversarial attack and defense for medical image analysis: methods and challenges, ACM Comput. Surv. 57 (3) (2024) 1–38.

[76] V. Sorin, S. Soffer, B.S. Glicksberg, Y. Barash, E. Konen, E. Klang, Adversarial attacks in radiology–a systematic review, Eur. J. Radiol. (2023) 111085.

[77] H. Hirano, A. Minagi, K. Takemoto, Universal adversarial attacks on deep neural networks for medical image classification, BMC Med. Imaging 21 (2021) 1–13.

[78] T. Sipola, T. Kokkonen, One-pixel attacks against medical imaging: a conceptual framework, in: Trends and Applications in Information Systems and Technologies: Volume 1 9, Springer, 2021, pp. 197–203.

[79] M.-J. Tsai, P.-Y. Lin, M.-E. Lee, Adversarial attacks on medical image classification, Cancers 15 (17) (2023) 4228.

[80] S. Pal, S. Rahman, M. Beheshti, A. Habib, Z. Jadidi, C. Karmakar, The impact of simultaneous adversarial attacks on robustness of medical image analysis, IEEE Access 12 (2024) 66478–66494.

[81] A. Rahman, M.S. Hossain, N.A. Alrajeh, F. Alsolami, Adversarial examples—security threats to Covid-19 deep learning systems in medical IOT devices, IEEE Internet Things J. 8 (12) (2020) 9603–9610.

[82] K.D. Apostolidis, G.A. Papakostas, Digital watermarking as an adversarial attack on medical image analysis with deep learning, J. Imaging. 8 (6) (2022) 155.

[83] S. Lal, S.U. Rehman, J.H. Shah, T. Meraj, H.T. Rauf, R. Damaševičius, M.A. Mohammed, K.H. Abdulkareem, Adversarial attack and defence through adversarial training and feature fusion for diabetic retinopathy recognition, Sensors 21 (11) (2021) 3922.

[84] O. Daanouni, B. Cherradi, A. Tmiri, Nsl-Mha-CNN: a novel CNN architecture for robust diabetic retinopathy prediction against adversarial attacks, IEEE Access 10 (2022) 103987–103999.

[85] D. Bharath Kumar, N. Kumar, S.D. Dunston, V.M.A. Rajam, Analysis of the impact of white box adversarial attacks in Resnet while classifying retinal fundus images, in: International Conference on Computational Intelligence in Data Science, Springer, 2022, pp. 162–175.

[86] Q. Xue, M.C. Chuah, New attacks on RNN based healthcare learning system and their detections, Smart Health 9 (2018) 144–157.

[87] J. Lam, P. Quan, J. Xu, J.V. Jeyakumar, M. Srivastava, Hard-label black-box adversarial attack on deep electrocardiogram classifier, in: Proceedings of the 1st ACM International Workshop on Security and Safety for Intelligent Cyber-Physical Systems, 2020, pp. 6–12.

[88] J. Shao, S. Geng, Z. Fu, W. Xu, T. Liu, S. Hong, Cardiodefense: defending against adversarial attack in ECG classification with adversarial distillation training, Biomed. Signal Process. Control 91 (2024) 105922.

[89] A. Albattah, M.A. Rassam, Detection of adversarial attacks against the hybrid convolutional long short-term memory deep learning technique for healthcare monitoring applications, Appl. Sci. 13 (11) (2023) 6807.

[90] M. Mozaffari-Kermani, S. Sur-Kolay, A. Raghunathan, N.K. Jha, Systematic poisoning attacks on and defenses for machine learning in healthcare, IEEE J. Biomed. Health Inform. 19 (6) (2014) 1893–1905.

[91] M. Sun, F. Tang, J. Yi, F. Wang, J. Zhou, Identify susceptible locations in medical records via adversarial attacks on deep predictive models, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 793–801.

[92] M. Ye, J. Luo, G. Zheng, C. Xiao, H. Xiao, T. Wang, F. Ma, Medattacker: exploring black-box adversarial attacks on risk prediction models in healthcare, in: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2022, pp. 1777–1780.

[93] S. Selvaganapathy, S. Sadasivam, N. Raj, Safexai: explainable AI to detect adversarial attacks in electronic medical records, in: Intelligent Data Engineering and Analytics: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021), Springer, 2022, pp. 501–509.

[94] A.I. Newaz, N.I. Haque, A.K. Sikder, M.A. Rahman, A.S. Uluagac, Adversarial attacks to machine learning-based smart healthcare systems, in: GLOBECOM 2020-2020 IEEE Global Communications Conference, IEEE, 2020, pp. 1–6.

[95] S.G. Selvaganapathy, S. Sadasivam, Healthcare security: usage of generative models for malware adversarial attacks and defense, in: Communication and Intelligent Systems: Proceedings of ICCIS 2020, Springer, 2021, pp. 885–897.

[96] S. Gaglio, A. Giammanco, G. Lo Re, M. Morana, Adversarial machine learning in e-health: attacking a smart prescription system, in: International Conference of the Italian Association for Artificial Intelligence, Springer, 2021, pp. 490–502.

[97] R. Paul, M. Schabath, R. Gillies, L. Hall, D. Goldgof, Mitigating adversarial attacks on medical image understanding systems, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), IEEE, 2020, pp. 1517–1521.

[98] X. Shi, Y. Peng, Q. Chen, T. Keenan, A.T. Thavikulwat, S. Lee, Y. Tang, E.Y. Chew, R.M. Summers, Z. Lu, Robust convolutional neural networks against adversarial attacks on medical images, Pattern Recognition 132 (2022) 108923.

[99] K. Kansal, P.S. Krishna, P.B. Jain, R. Surya, P. Honnavalli, S. Eswaran, Defending against adversarial attacks on Covid-19 classifier: a denoiser-based approach, Heliyon 8 (10) (2022).

[100] X. Li, D. Zhu, Robust detection of adversarial attacks on medical images, in: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), IEEE, 2020, pp. 1154–1158.

[101] L. Alzubaidi, A.-D. Khamael, H.A.-H. Obeed, A. Saihood, M.A. Fadhel, S.A. Jebur, Y. Chen, A.S. Albahri, J. Santamaría, A. Gupta, et al., Meff–a model ensemble feature fusion approach for tackling adversarial attacks in medical imaging, Intell. Syst. With Appl. 22 (2024) 200355.

[102] M. Watson, N. Al Moubayed, Attack-agnostic adversarial detection on medical data using explainable machine learning, in: 2020 25th International Conference on Pattern Recognition (ICPR), IEEE, 2021, pp. 8180–8187.

[103] H. Venugopal, M.S. Christo, Strengthening healthcare cybersecurity by optimizing adversarial attack defences in deep learning models using GPU and parallel processing technologies, in: 2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT), vol. 1, IEEE, 2024, pp. 1266–1272.

[104] N. Ghaffari Laleh, D. Truhn, G.P. Veldhuizen, T. Han, M. van Treeck, R.D. Buelow, R. Langer, B. Dislich, P. Boor, V. Schulz, et al., Adversarial attacks and adversarial robustness in computational pathology, Nat. Commun. 13 (1) (2022) 5711.

[105] M. Skoglund, F. Warg, H. Hansson, S. Punnekkat, Black-box testing for security-informed safety of automated driving systems, in: 2021 IEEE 93rd Vehicular Technology Conference (VTC2021-Spring), IEEE, 2021, pp. 1–7.

[106] B.U.H. Sheikh, A. Zafar, Untargeted white-box adversarial attack to break into deep learning based Covid-19 monitoring face mask detection system, Multim. Tools Appl. 83 (8) (2024) 23873–23899.

[107] G. Bortsova, C. González-Gonzalo, S.C. Wetstein, F. Dubost, I. Katramados, L. Hogeweg, B. Liefers, B. van Ginneken, J.P. Pluim, M. Veta, et al., Adversarial attack vulnerability of medical image analysis systems: unexplored factors, Med. Image Anal. 73 (2021) 102141.

[108] N. Mangaokar, J. Pu, P. Bhattacharya, C.K. Reddy, B. Viswanath, Jekyll: attacking medical image diagnostics using deep generative models, in: 2020 IEEE European Symposium on Security and Privacy (EuroS&P), IEEE, 2020, pp. 139–157.

[109] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z.B. Celik, A. Swami, Practical black-box attacks against machine learning, in: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, 2017, pp. 506–519.

[110] S. Zhang, X. Xie, Y. Xu, A brute-force black-box method to attack machine learning-based systems in cybersecurity, IEEE Access 8 (2020) 128250–128263.

[111] K. Roshan, A. Zafar, Black-box adversarial transferability: an empirical study in cybersecurity perspective, Comput. Secur. 141 (2024) 103853.

[112] G. Apruzzese, M. Colajanni, L. Ferretti, M. Marchetti, Addressing adversarial attacks against security systems based on machine learning, in: 2019 11th International Conference on Cyber Conflict (CyCon), vol. 900, IEEE, 2019, pp. 1–18.

[113] G. Apruzzese, M. Colajanni, M. Marchetti, Evaluating the effectiveness of adversarial attacks against botnet detectors, in: 2019 IEEE 18th International Symposium on Network Computing and Applications (NCA), IEEE, 2019, pp. 1–8.

[114] O. Ibitoye, O. Shafiq, A. Matrawy, Analyzing adversarial attacks against deep learning for intrusion detection in IOT networks, in: 2019 IEEE Global Communications Conference (GLOBECOM), IEEE, 2019, pp. 1–6.

[115] Y. Zhu, L. Cui, Z. Ding, L. Li, Y. Liu, Z. Hao, Black box attack and network intrusion detection using machine learning for malicious traffic, Comput. Secur. 123 (2022) 102922.

[116] H.A. Alatwi, A. Aldweesh, Adversarial black-box attacks against network intrusion detection systems: a survey, in: 2021 IEEE World AI IoT Congress (AIIoT), IEEE, 2021, pp. 0034–0040.

[117] A. Kuppa, N.-A. Le-Khac, Black box attacks on explainable artificial intelligence (XAI) methods in cyber security, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–8.

[118] A. Kuppa, N.-A. Le-Khac, Adversarial XAI methods in cybersecurity, IEEE Trans. Inf. Forensics Secur. 16 (2021) 4924–4938.

[119] I. Rosenberg, A. Shabtai, Y. Elovici, L. Rokach, Adversarial machine learning attacks and defense methods in the cyber security domain, ACM Comput. Surv. 54 (5) (2021) 1–36.

[120] S. Zhou, C. Liu, D. Ye, T. Zhu, W. Zhou, P.S. Yu, Adversarial attacks and defenses in deep learning: from a perspective of cybersecurity, ACM Comput. Surv. 55 (8) (2022) 1–39.

[121] C. Yinka-Banjo, O.-A. Ugot, A review of generative adversarial networks and its application in cybersecurity, Artif. Intell. Rev. 53 (2020) 1721–1736.

[122] I.K. Dutta, B. Ghosh, A. Carlson, M. Totaro, M. Bayoumi, Generative adversarial networks in security: a survey, in: 2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), IEEE, 2020, pp. 0399–0405.

[123] M. Pawlicki, M. Choraś, R. Kozik, Defending network intrusion detection systems against adversarial evasion attacks, Futur. Gener. Comput. Syst. 110 (2020) 148–154.

[124] T.A. Khaleel, Developing robust machine learning models to defend against adversarial attacks in the field of cybersecurity, in: 2024 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), IEEE, 2024, pp. 1–7.

[125] K. Barik, S. Misra, Adversarial attack defense analysis: an empirical approach in cybersecurity perspective, Softw. Impacts 21 (2024) 100681.

[126] M. Khan, L. Ghafoor, Adversarial machine learning in the context of network security: challenges and solutions, J. Comput. Intell. Robot. 4 (1) (2024) 51–63.

[127] M. Aurangzeb, Y. Wang, S. Iqbal, A. Naveed, Z. Ahmed, M. Alenezi, M. Shouran, Enhancing cybersecurity in smart grids: deep black box adversarial attacks and quantum voting ensemble models for blockchain privacy-preserving storage, Energy Rep. 11 (2024) 2493–2515.
[128] L. Sun, M. Tan, Z. Zhou, A survey of practical adversarial example attacks, Cybersecurity 1 (1) (2018) 9.
[129] M.B. Line, A. Zand, G. Stringhini, R. Kemmerer, Targeted attacks against industrial control systems: is the power industry prepared? in: Proceedings of the 2nd Workshop on Smart Energy Grid Security, 2014, pp. 13–22.
[130] D.I. Urbina, J.A. Giraldo, A.A. Cardenas, N.O. Tippenhauer, J. Valente, M. Faisal, J. Ruths, R. Candell, H. Sandberg, Limiting the impact of stealthy attacks on industrial control systems, in: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 1092–1105.
[131] K. Paridari, N. O'Mahony, A.E.-D. Mady, R. Chabukswar, M. Boubekeur, H. Sandberg, A framework for attack-resilient industrial control systems: attack detection and controller reconfiguration, Proc. IEEE 106 (1) (2017) 113–128.
[132] C. Feng, T. Li, Z. Zhu, D. Chana, A deep learning-based framework for conducting stealthy attacks in industrial control systems, arXiv preprint arXiv:1709.06397, 2017.
[133] A. Erba, R. Taormina, S. Galelli, M. Pogliani, M. Carminati, S. Zanero, N.O. Tippenhauer, Real-time evasion attacks with physical constraints on deep learning-based anomaly detectors in industrial control systems, arXiv preprint arXiv:1907.07487, 2019.
[134] J. Chen, X. Gao, R. Deng, Y. He, C. Fang, P. Cheng, Generating adversarial examples against machine learning-based intrusion detector in industrial control systems, IEEE Trans. Dependable Secure Comput. 19 (3) (2020) 1810–1825.
[135] E. Sarkar, H. Benkraouda, M. Maniatakos, I came, i saw, i hacked: automated generation of process-independent attacks for industrial control systems, in: Proceedings of the 15th ACM Asia Conference on Computer and Communications Security, 2020, pp. 744–758.
[136] S.K. Alabugin, A.N. Sokolov, Applying of generative adversarial networks for anomaly detection in industrial control systems, in: 2020 Global Smart Industry Conference (GloSIC), IEEE, 2020, pp. 199–203.
[137] A. Erba, R. Taormina, S. Galelli, M. Pogliani, M. Carminati, S. Zanero, N.O. Tippenhauer, Constrained concealment attacks against reconstruction-based anomaly detectors in industrial control systems, in: Proceedings of the 36th Annual Computer Security Applications Conference, 2020, pp. 480–495.
[138] M. Kravchik, B. Biggio, A. Shabtai, Poisoning attacks on cyber attack detectors for industrial control systems, in: Proceedings of the 36th Annual ACM Symposium on Applied Computing, 2021, pp. 116–125.
[139] G.M. Makrakis, C. Kolias, G. Kambourakis, C. Rieger, J. Benjamin, Vulnerabilities and attacks against industrial control systems and critical infrastructures, arXiv preprint arXiv:2109.03945, 2021.
[140] Á.L.P. Gómez, L.F. Maimó, A.H. Celdrán, F.J.G. Clemente, F. Cleary, Crafting adversarial samples for anomaly detectors in industrial control systems, Proc. Comput. Sci. 184 (2021) 573–580.
[141] M.A. Umer, C.M. Ahmed, M.T. Jilani, A.P. Mathur, Attack rules: an adversarial approach to generate attacks for industrial control systems using machine learning, in: Proceedings of the 2th Workshop on CPS&IoT Security and Privacy, 2021, pp. 35–40.
[142] H. Figueroa, Y. Wang, G.C. Giakos, Adversarial attacks in industrial control cyber physical systems, in: 2022 IEEE International Conference on Imaging Systems and Techniques (IST), IEEE, 2022, pp. 1–6.
[143] M. Kravchik, L. Demetrio, B. Biggio, A. Shabtai, Practical evaluation of poisoning attacks on online anomaly detectors in industrial control systems, Comput. Secur. 122 (2022) 102901.
[144] L. Yao, S. Shao, S. Hariri, Resilient machine learning (rml) against adversarial attacks on industrial control systems, in: 2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA), IEEE, 2023, pp. 1–8.
[145] M. Duan, G. Xiao, K. Li, B. Xiao, A black-box attack algorithm targeting unlabeled industrial AI systems with contrastive learning, IEEE Trans. Ind. Inform. 20 (4) (2024) 6325–6335.
[146] V. Pozdnyakov, A. Kovalenko, I. Makarov, M. Drobyshevskiy, K. Lukyanov, Adversarial attacks and defenses in automated control systems: A comprehensive benchmark, arXiv preprint arXiv:2403.13502, 2024.
[147] Y. Liu, L. Xu, S. Yang, D. Zhao, X. Li, Adversarial sample attacks and defenses based on lstm-ed in industrial control systems, Comput. & Secur. 140 (2024) 103750.
[148] Y. Qian, D. Ma, B. Wang, J. Pan, J. Wang, Z. Gu, J. Chen, W. Zhou, J. Lei, Spot evasion attacks: adversarial examples for license plate recognition systems with convolutional neural networks, Comput. Secur. 95 (2020) 101826.
[149] X. Xu, J. Zhang, Y. Li, Y. Wang, Y. Yang, H.T. Shen, Adversarial attack against urban scene segmentation for autonomous vehicles, IEEE Trans. Ind. Informat. 17 (6) (2020) 4117–4126.
[150] W. Jiang, H. Li, S. Liu, X. Luo, R. Lu, Poisoning and evasion attacks against deep learning algorithms in autonomous vehicles, IEEE Trans. Veh. Technol. 69 (4) (2020) 4439–4449.
[151] Y. Li, C. Wen, F. Juefei-Xu, C. Feng, Fooling lidar perception via adversarial trajectory perturbation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7898–7907.
[152] Q. Zhang, Y. Zhao, Y. Wang, T. Baker, J. Zhang, J. Hu, Towards cross-task universal perturbation against black-box object detectors in autonomous driving, Comput. Netw. 180 (2020) 107388.
[153] J. Won, S.-H. Seo, E. Bertino, A secure shuffling mechanism for white-box attack-resistant unmanned vehicles, IEEE Trans. Mob. Comput. 19 (5) (2019) 1023–1039.
[154] I. Sobh, A. Hamed, V.R. Kumar, S. Yogamani, Adversarial attacks on multi-task visual perception for autonomous driving, arXiv preprint arXiv:2107.07449, 2021.
[155] K.N. Kumar, C. Vishnu, R. Mitra, C.K. Mohan, Black-box adversarial attacks in autonomous vehicle technology, in: 2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), IEEE, 2020, pp. 1–7.
[156] J. Zhang, Y. Lou, J. Wang, K. Wu, K. Lu, X. Jia, Evaluating adversarial attacks on driving safety in vision-based autonomous vehicles, IEEE Internet Things J. 9 (5) (2021) 3443–3456.
[157] J. Sun, Y. Cao, Q.A. Chen, Z.M. Mao, Towards robust {LiDAR-based} perception in autonomous driving: general black-box adversarial sensor attack and countermeasures, in: 29th USENIX Security Symposium (USENIX Security 20), 2020, pp. 877–894.
[158] Y. Zhu, C. Miao, F. Hajiaghajani, M. Huai, L. Su, C. Qiao, Adversarial attacks against lidar semantic segmentation in autonomous driving, in: Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems, 2021, pp. 329–342.
[159] A. Sarker, H. Shen, T. Sen, A suspicion-free black-box adversarial attack for deep driving maneuver classification models, in: 2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS), IEEE, 2021, pp. 786–796.
[160] A. Sarker, H. Shen, T. Sen, A context-aware black-box adversarial attack for deep driving maneuver classification models, in: 2021 18th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), IEEE, 2021, pp. 1–9.
[161] R.S. Hallyburton, Y. Liu, Y. Cao, Z.M. Mao, M. Pajic, Security analysis of{Camera-LiDAR} fusion against {Black-Box} attacks on autonomous vehicles, in: 31st USENIX Security Symposium (USENIX Security 22), 2022, pp. 1903–1920.
[162] S.B. Jakobsen, K.S. Knudsen, B. Andersen, Analysis of sensor attacks against autonomous vehicles, in: 8th International Conference on Internet of Things, Big Data and Security, SCITEPRESS Digital Library, 2023, pp. 131–139.
[163] Y. Deng, X. Zheng, T. Zhang, C. Chen, G. Lou, M. Kim, An analysis of adversarial attacks and defenses on autonomous driving models, in: 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom), IEEE, 2020, pp. 1–10.
[164] J. Liu, J.-M. Park, "seeing is not always believing": detecting perception error attacks against autonomous vehicles, IEEE Trans. Dependable Secure Comput. 18 (5) (2021) 2209–2223.
[165] K. Dhawale, P. Gupta, T.K. Jain, AI approach for autonomous vehicles to defend from adversarial attacks, in: Proceedings of International Conference on Intelligent Cyber-Physical Systems: ICPS 2021, Springer, 2022, pp. 207–221.
[166] K.H. Shibly, M.D. Hossain, H. Inoue, Y. Taenaka, Y. Kadobayashi, Towards autonomous driving model resistant to adversarial attack, Appl. Artif. Intell. 37 (1) (2023) 2193461.
[167] K. Kim, J.S. Kim, S. Jeong, J.-H. Park, H.K. Kim, Cybersecurity for autonomous vehicles: review of attacks and defense, Comput. Secur. 103 (2021) 102150.
[168] Y. Deng, T. Zhang, G. Lou, X. Zheng, J. Jin, Q.-L. Han, Deep learning-based autonomous driving systems: a survey of attacks and defenses, IEEE Trans. Ind. Informat. 17 (12) (2021) 7897–7912.
[169] S. Gupta, C. Maple, R. Passerone, An investigation of cyber-attacks and security mechanisms for connected and autonomous vehicles, IEEE Access 11 (2023) 90641–90669.
[170] M. Girdhar, J. Hong, J. Moore, Cybersecurity of autonomous vehicles: a systematic literature review of adversarial attacks and defense models, IEEE Open J. Veh. Technol. 4 (2023) 417–437.
[171] A. Kloukiniotis, A. Papandreou, A. Lalos, P. Kapsalas, D.-V. Nguyen, K. Moustakas, Countering adversarial attacks on autonomous vehicles using denoising techniques: a review, IEEE Open J. Intell. Transp. Syst. 3 (2022) 61–80.
[172] A. Amirkhani, M.P. Karimi, A. Banitalebi-Dehkordi, A survey on adversarial attacks and defenses for object detection and their applications in autonomous vehicles, The Visual Computer 39 (11) (2023) 5293–5307.
[173] W. Zong, Y.-W. Chow, W. Susilo, S. Rana, S. Venkatesh, Targeted universal adversarial perturbations for automatic speech recognition, in: Information Security: 24th International Conference, ISC 2021, Virtual Event, November 10–12, 2021, Proceedings 24, Springer, 2021, pp. 358–373.
[174] H. Kwon, Y. Kim, H. Yoon, D. Choi, Selective audio adversarial example in evasion attack on speech recognition system, IEEE Trans. Inf. Forensics Secur. 15 (2019) 526–538.
[175] K. Ko, S. Kim, H. Kwon, Multi-targeted audio adversarial example for use against speech recognition systems, Comput. Secur. 128 (2023) 103168.
[176] B. Zheng, P. Jiang, Q. Wang, Q. Li, C. Shen, C. Wang, Y. Ge, Q. Teng, S. Zhang, Black-box adversarial attacks on commercial speech platforms with minimal information, in: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021, pp. 86–107.
[177] S. Wang, Z. Zhang, G. Zhu, X. Zhang, Y. Zhou, J. Huang, Query-efficient adversarial attack with low perturbation against end-to-end speech recognition systems, IEEE Trans. Inf. Forensics Secur. 18 (2022) 351–364.
[178] C. Tong, X. Zheng, J. Li, X. Ma, L. Gao, Y. Xiang, Query-efficient black-box adversarial attacks on automatic speech recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 31 (2023) 3981–3992.
[179] Z. Fang, T. Wang, L. Zhao, S. Zhang, B. Li, Y. Ge, Q. Li, C. Shen, Q. Wang, Zero-query adversarial attack on black-box automatic speech recognition systems, in: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, 2024, pp. 630–644.
[180] Y. Ge, L. Zhao, Q. Wang, Y. Duan, M. Du, Advddos: zero-query adversarial attacks against commercial speech recognition systems, IEEE Trans. Inf. Forensics Secur. 18 (2023) 3647–3661.

[181] G. Zhang, X. Ma, H. Zhang, Z. Xiang, X. Ji, Y. Yang, X. Cheng, P. Hu, {LaserAdv}: laser adversarial attacks on speech recognition systems, in: 33rd USENIX Security Symposium (USENIX Security 24), 2024, pp. 3945–3961.
[182] L. Liang, B. Guo, Z. Lian, Q. Li, H. Jing, Impga: an effective and imperceptible black-box attack against automatic speech recognition systems, in: Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data, Springer, 2022, pp. 349–363.
[183] Z. Qin, X. Zhang, S. Li, A robust adversarial attack against speech recognition with Uap, High-Confid. Comput. 3 (1) (2023) 100098.
[184] K. Rajaratnam, B. Alshemali, J. Kalita, Speech coding and audio preprocessing for mitigating and detecting audio adversarial examples on automatic speech recognition, Mach. Learn. Comput. Vis. Nat. Lang. Process. (2018) 1.
[185] H. Kwon, H. Yoon, K.-W. Park, Acoustic-decoy: detection of adversarial examples through audio modification on speech recognition system, Neurocomputing 417 (2020) 357–370.
[186] F. Guo, Z. Sun, Y. Chen, L. Ju, Towards the universal defense for query-based audio adversarial attacks on speech recognition system, Cybersecurity 6 (1) (2023) 40.
[187] A. Huq, W. Zhang, X. Hu, Mixpgd: Hybrid adversarial training for speech recognition systems, arXiv preprint arXiv:2303.05758, 2023.
[188] S. Joshi, S. Kataria, Y. Shao, P. Zelasko, J. Villalba, S. Khudanpur, N. Dehak, Defense against adversarial attacks on hybrid speech recognition using joint adversarial fine-tuning with denoiser, arXiv preprint arXiv:2204.03851, 2022.
[189] X. Zhang, H. Tan, X. Huang, D. Zhang, K. Tang, Z. Gu, Adversarial attacks on asr systems: An overview, arXiv preprint arXiv:2208.02250, 2022.
[190] H. Abdullah, K. Warren, V. Bindschaedler, N. Papernot, P. Traynor, Sok: the faults in our asrs: an overview of attacks against automatic speech recognition and speaker identification systems, in: 2021 IEEE Symposium on Security and Privacy (SP), IEEE, 2021, pp. 730–747.
[191] Z. Sun, J. Zhao, F. Guo, Y. Chen, L. Ju, Commanderuap: a practical and transferable universal adversarial attacks on speech recognition models, Cybersecurity 7 (1) (2024) 38.
[192] X. Li, M. Liu, X. Ma, L. Gao, Exploring the vulnerability of natural language processing models via universal adversarial texts, in: Proceedings of the 19th Annual Workshop of the Australasian Language Technology Association, 2021, pp. 138–148.
[193] I. Fursov, A. Zaytsev, P. Burnyshev, E. Dmitrieva, N. Klyuchnikov, A. Kravchenko, E. Artemova, E. Komleva, E. Burnaev, A differentiable language model adversarial attack on text classifiers, IEEE Access 10 (2022) 17966–17976.
[194] W. Simoncini, G. Spanakis, Seqattack: on adversarial attacks for named entity recognition, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2021, pp. 308–318.
[195] M. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava, K.-W. Chang, Generating natural language adversarial examples, arXiv preprint arXiv:1804.07998, 2018.
[196] Y. Zang, B. Hou, F. Qi, Z. Liu, X. Meng, M. Sun, Learning to attack: Towards textual adversarial attacking in real-world situations, arXiv preprint arXiv:2009.09192, 2020.
[197] R. Maheshwary, S. Maheshwary, V. Pudi, Generating natural language attacks in a hard label black box setting, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 13525–13533.
[198] X. Han, Q. Li, H. Cao, L. Han, B. Wang, X. Bao, Y. Han, W. Wang, Bfs2adv: black-box adversarial attack towards hard-to-attack short texts, Comput. Secur. 141 (2024) 103817.
[199] S. Zhou, K. Li, G. Min, Attention-based genetic algorithm for adversarial attack in natural language processing, in: International Conference on Parallel Problem Solving from Nature, Springer, 2022, pp. 341–355.
[200] Z. Wang, W. Wang, Q. Chen, Q. Wang, A. Nguyen, Generating valid and natural adversarial examples with large language models, in: 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD), IEEE, 2024, pp. 1716–1721.
[201] Z. Shao, Z. Wu, M. Huang, Advexpander: generating natural language adversarial examples by expanding text, IEEE/ACM Trans. Audio Speech Lang. Process. 30 (2021) 1184–1196.
[202] Z. Zhou, H. Guan, M.M. Bhat, J. Hsu, Fake news detection via nlp is vulnerable to adversarial attacks, arXiv preprint arXiv:1901.09657, 2019.
[203] T. Roth, Y. Gao, A. Abuadbba, S. Nepal, W. Liu, Token-modification adversarial attacks for natural language processing: a survey, AI Commun. 37 (4) (2024) 655–676.
[204] A. Jha, C.K. Reddy, Codeattack: Code-based adversarial attacks for pre-trained programming language models, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, 2023, pp. 14892–14900.
[205] Y. Zhou, X. Zheng, C.-J. Hsieh, K.-W. Chang, X. Huang, Defense against adversarial attacks in nlp via dirichlet neighborhood ensemble, arXiv preprint arXiv:2006.11627, 2020.
[206] X. Wang, J. Hao, Y. Yang, K. He, Natural language adversarial defense through synonym encoding, in: Uncertainty in Artificial Intelligence, PMLR, 2021, pp. 823–833.
[207] X. Dong, Adversarial attacks and defenses in natural language processing, PhD thesis, Nanyang Technological University, 2022.
[208] J. Zeng, J. Xu, X. Zheng, X. Huang, Certified robustness to text adversarial attacks by randomized [mask], Comput. Linguist. 49 (2) (2023) 395–427.
[209] W.E. Zhang, Q.Z. Sheng, A. Alhazmi, C. Li, Adversarial attacks on deep-learning models in natural language processing: a survey, ACM Trans. Intell. Syst. Technol. 11 (3) (2020) 1–41.
[210] A. Huq, M. Pervin, et al., Adversarial attacks and defense on texts: A survey, arXiv preprint arXiv:2005.14108, 2020.
[211] Y. Zhang, K. Shao, J. Yang, H. Liu, Adversarial attacks and defenses on deep learning models in natural language processing, in: 2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), vol. 5, IEEE, 2021, pp. 1281–1285.
[212] N. Minh, R. Andini, Advanced adversarial attack techniques on natural language processing systems: methods, impacts, and defense mechanisms, Adv. Intell. Inf. Syst. 8 (4) (2023) 12–20.
[213] S. Goyal, S. Doddapaneni, M.M. Khapra, B. Ravindran, A survey of adversarial defenses and robustness in NLP, ACM Comput. Surv. 55 (14s) (2023) 1–39.
[214] J. Zou, S. Zhang, M. Qiu, Adversarial attacks on large language models, in: International Conference on Knowledge Science, Engineering and Management, Springer, 2024, pp. 85–96.
[215] S. Qiu, Q. Liu, S. Zhou, W. Huang, Adversarial attack and defense technologies in natural language processing: a survey, Neurocomputing 492 (2022) 278–307.
[216] I. Alsmadi, K. Ahmad, M. Nazzal, F. Alam, A. Al-Fuqaha, A. Khreishah, A. Algosaibi, Adversarial NLP for social network applications: attacks, defenses, and research directions, IEEE Trans. Comput. Soc. Syst. 10 (6) (2022) 3089–3108.
[217] M. Omar, S. Choi, D. Nyang, D. Mohaisen, Robust natural language processing: recent advances, challenges, and future directions, IEEE Access 10 (2022) 86038–86056.
[218] K.-W. Chang, H. He, R. Jia, S. Singh, Robustness and adversarial examples in natural language processing, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, 2021, pp. 22–26.
[219] V. Ballet, X. Renard, J. Aigrain, T. Laugel, P. Frossard, M. Detyniecki, Imperceptible adversarial attacks on tabular data, arXiv preprint arXiv:1911.03274, 2019.
[220] M. Schreyer, T. Sattarov, B. Reimer, D. Borth, Adversarial learning of deepfakes in accounting, arXiv preprint arXiv:1910.03810, 2019.
[221] L. Yang, E. Kenny, T.L.J. Ng, Y. Yang, B. Smyth, R. Dong, Generating plausible counterfactual explanations for deep transformers in financial text classification, in: Proceedings of the 28th International Conference on Computational Linguistics, 2020, pp. 6150–6160.
[222] N. Kumar, S. Vimal, K. Kayathwal, G. Dhama, Evolutionary adversarial attacks on payment systems, in: 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2021, pp. 813–818.
[223] J. Xiao, Y. Tian, Y. Jia, X. Jiang, L. Yu, S. Wang, Black-box attack-based security evaluation framework for credit card fraud detection models, Inf. J. Comput. 35 (5) (2023) 986–1001.
[224] O. Lee, H. Ha, H. Choi, H. Joo, M. Cheon, Alerting the impact of adversarial attacks and how to detect it effectively via machine learning approach: with financial and esg data, in: Communication and Intelligent Systems: Proceedings of ICCIS 2021, Springer, 2022, pp. 713–724.
[225] M.-Y. Tsai, H.-H. Cho, C.-M. Yu, Y.-C. Chang, H.-C. Chao, Effective adversarial examples identification of credit card transactions, IEEE Intell. Syst. 39 (4) (2024) 50–59.
[226] Y.-Y. Chen, C.-T. Chen, C.-Y. Sang, Y.-C. Yang, S.-H. Huang, Adversarial attacks against reinforcement learning-based portfolio management strategy, IEEE Access 9 (2021) 50667–50685.
[227] G. Liu, L. Lai, Provably efficient black-box action poisoning attacks against reinforcement learning, Adv. Neural Inf. Process. Syst. 34 (2021) 12400–12410.
[228] F. Ataiefard, H. Hemmati, Gray-box adversarial attack of deep reinforcement learning-based trading agents, in: 2023 International Conference on Machine Learning and Applications (ICMLA), IEEE, 2023, pp. 675–682.
[229] M. Gallagher, N. Pitropakis, C. Chrysoulas, P. Papadopoulos, A. Mylonas, S. Katsikas, Investigating machine learning attacks on financial time series models, Comput. & Secur. 123 (2022) 102933.
[230] E. Nehemya, Y. Mathov, A. Shabtai, Y. Elovici, Taking over the stock market: adversarial perturbations against algorithmic traders, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2021, pp. 221–236.
[231] Y. Xie, D. Wang, P.-Y. Chen, J. Xiong, S. Liu, S. Koyejo, A word is worth a thousand dollars: Adversarial attack on tweets fools stock predictions, arXiv preprint arXiv:2205.01094, 2022.
[232] M. Leippold, Sentiment spin: attacking financial sentiment with gpt-3, Financ. Res. Lett. 55 (2023) 103957.
[233] D. Lunghi, A. Simitsis, O. Caelen, G. Bontempi, Adversarial learning in real-world fraud detection: challenges and perspectives, in: Proceedings of the Second ACM Data Economy Workshop, 2023, pp. 27–33.
[234] T.L. Melo, J. Bravo, M.O. Sampaio, P. Romano, H. Ferreira, J.T. Ascensão, P. Bizarro, Adversarial training for tabular data with attack propagation, arXiv preprint arXiv:2307.15677, 2023.
[235] L. Zhou, X. Xiong, J. Ernstberger, S. Chaliasos, Z. Wang, Y. Wang, K. Qin, R. Wattenhofer, D. Song, A. Gervais, Sok: decentralized finance (defi) attacks, in: 2023 IEEE Symposium on Security and Privacy (SP), IEEE, 2023, pp. 2444–2461.
[236] S. Yang, H. Guo, X. Du, J. Yang, Z. Lu, An adversarial attack method against financial fraud detection model beta wavelet graph neural network via node injection, in: 2024 IEEE 11th International Conference on Cyber Security and Cloud Computing (CSCloud), IEEE, 2024, pp. 7–12.
[237] A. MaungMaung, H. Kiya, Ensemble of key-based models: defense against black-box adversarial attacks, in: 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), IEEE, 2021, pp. 95–98.
[238] J. Wang, Y. Hu, Y. Qi, Z. Peng, C. Zhou, Mitigating adversarial attacks based on denoising & reconstruction with finance authentication system case study, IEEE Trans. Comput. 73 (2) (2024) 314–326.
[239] Z. Zhang, W. Li, R. Bao, K. Harimoto, Y. Wu, X. Sun, Asat: adaptively scaled adversarial training in time series, Neurocomputing 522 (2023) 11–23.

[240] M. Zhu, M. Zhu, Z. Xu, L. Yu, Y. Zong, The application of deep learning in financial payment security and the challenge of generating adversarial network models, in: The 8th International Scientific and Practical Conference “Priority Areas of Research in the Scientific Activity of Teachers”(February 27–March 01, 2024) Zagreb, Croatia. International Science Group. 2024. 298 P, 2024, pp. 174.
[241] A.S. George, Securing the future of finance: how AI, blockchain, and machine learning safeguard emerging neobank technology against evolving cyber threats, Partners Univ. Innov. Res. Publ. 1 (1) (2023) 54–66.
[242] B. Amerirad, M. Cattaneo, R.S. Kenett, E. Luciano, Adversarial artificial intelligence in insurance: from an example to some potential remedies, Risks 11 (1) (2023) 20.
[243] K. Huang, X. Chen, Y. Yang, J. Ponnapalli, G. Huang, Chatgpt in finance and banking, in: Beyond AI: ChatGPT, Web3, and the Business Landscape of Tomorrow, Springer, 2023, pp. 187–218.
[244] E. Cramer, J. Gao, A black-box adversarial attack on demand side management, Comput. Chem. Eng. 186 (2024) 108681.
[245] D. Yang, H. Liu, X. Wang, Adversarial attacks on medical large language models: safety risks in clinical decision-making, J. Biomed. Inform. 152 (2025) 104567.
[246] T. Shu, R. Zhang, A. Gupta, Attackeval: a systematic evaluation framework for jailbreak attacks on large language models, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI Press, 2025, pp. 14123–14131.
[247] R. Zhu, W. Xing, H. Ma, Robustness benchmarking of large language models under adversarial prompt perturbations, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL, 2024, pp. 5120–5134.
[248] N. Mehrabi, F. Morstatter, A. Galstyan, A holistic framework for assessing safety of large language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL, 2023, pp. 11401–11418.
[249] Y. Qin, T. Zhang, M. Du, Automatic generation of adversarial prompts for large language models, in: Proceedings of the Web Conference (WWW), ACM, 2024, pp. 2845–2856.
[250] P. Li, G. Chen, Y. Fang, Semantic illusion attacks on language models, in: Proceedings of the 2024 Annual Meeting of the ACL, ACL, 2024, pp. 842–855.
[251] M. Alber, J. Schneider, F. Krämer, Vulnerability of medical large language models to targeted data poisoning, Nat. Med. 31 (2025) 112–123.
[252] Y. Zhang, T. Wang, P. Chen, Instruction backdoor attacks against customized large language models, in: Proceedings of the ACM Conference on Computer and Communications Security (CCS), ACM, 2024, pp. 843–857.
[253] X. Chen, J. Liu, K. Ren, Hidden multi-turn backdoor attacks on chat-based large language models, in: Proceedings of the IEEE Symposium on Security and Privacy (S&P), IEEE, 2024, pp. 1159–1176.
[254] Z. He, L. Huang, S. Gao, Backdoor-based data stealing attacks against large language models, in: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), IJCAI, 2024, pp. 3345–3353.
[255] Q. Zhao, C. Lin, J. Wu, Poisoning retrieval-augmented generation via adversarial knowledge injection, in: Proceedings of the SIAM International Conference on Data Mining (SDM), SIAM, 2025, pp. 1015–1026.
[256] P. Lermen, R. Schuster, Poisoning retrieval-augmented language models via adversarial document injection, Trans. Mach. Learn. Res. 12 (2024) 1–23.
[257] B. Ergün, A. Onan, Adversarial prompt detection using supervised classification for large language models, Expert Syst. Appl. 245 (2025) 123097.
[258] M. Yi, Y. Zhao, S. Huang, Beat: a black-box defense against backdoor unalignment in large language models, in: Proceedings of the ACM Web Conference (The WebConf), ACM, 2025, pp. 2184–2195.
[259] A. Zou, M. Liu, N. Carlini, Universal and transferable adversarial attacks on aligned language models, in: Proceedings of NeurIPS 2023, 2023, pp. 20123–20137.
[260] H. Wang, Y. Rao, W. Zhang, Mitigating jailbreak attacks via alignment smoothing in large language models, IEEE Trans. Artif. Intell. 5 (2024) 884–899.
[261] Y. Qian, C. Zhao, Z. Gu, B. Wang, S. Ji, W. Wang, Y. Zhang, $F^2$2at: feature-focusing adversarial training via disentanglement of natural and perturbed patterns, IEEE Trans. Knowl. Data Eng. 37 (9) (2025) 5201–5213, https://doi.org/10.1109/TKDE.2025.3580116