Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.084672

REVIEW

Adversarial Threats and Defence Mechanisms in Artificial Intelligence of Things Systems: A Systematic Review

Ali Hassan[1], **Syed Rizwan Hassan**[2,*] **and Ammar Rafiq**[3]

[1]Department of Electrical Engineering, HITEC University, Taxila, Pakistan
[2]Department of Computer Engineering, Gachon University, Seongnam-Si, Republic of Korea
[3]Department of Computer Science, National University of Computer and Emerging Sciences, Islamabad, Chiniot-Faisalabad Campus, Chiniot, Pakistan

*Corresponding Author: Syed Rizwan Hassan. Email: syed9919@gachon.ac.kr

Received: 27 April 2026; Accepted: 04 June 2026; Published: 13 August 2026

ABSTRACT: Artificial Intelligence of Things (AIoT) systems have emerged through the rapid integration of artificial intelligence (AI) and the Internet of Things (IoT), enabling intelligent sensing, distributed learning, and real-time decision-making across diverse application domains. However, this convergence also introduces a significantly expanded adversarial attack surface spanning sensing devices, communication networks, learning pipelines, and actuation environments. This paper presents a comprehensive systematic review of adversarial threats and defence mechanisms in AIoT systems using a novel 3D-AIoT-TT (Three-Dimensional AIoT Threat Taxonomy) framework. The proposed taxonomy jointly models three fundamental dimensions: (i) AI pipeline stages, (ii) IoT architectural layers, and (iii) adversarial knowledge levels. By integrating these dimensions into a unified analytical framework, the taxonomy enables systematic characterization, classification, and evaluation of adversarial behaviours across heterogeneous AIoT environments. Using this framework, the study analyzes a broad spectrum of adversarial threats, including poisoning attacks, evasion attacks, backdoor attacks, model extraction attacks, and federated learning-based attacks. In addition, existing defence mechanisms such as adversarial training, hardware-assisted protection, federated defence strategies, anomaly detection, and certified robustness techniques are critically examined with particular emphasis on their practicality in resource-constrained edge AIoT deployments. A key contribution of this work is the identification and analysis of emergent cross-layer adversarial threats, where vulnerabilities arise through interactions among sensing, communication, learning, and actuation components. Unlike traditional isolated security models, these threats propagate across multiple AIoT layers and exploit systemic interdependencies within intelligent infrastructures. Overall, this study establishes a structured and unified understanding of adversarial AIoT security, highlights current research limitations and underexplored areas, and emphasizes the need for scalable, cross-layer, and deployment-aware defence mechanisms for next-generation intelligent systems.

KEYWORDS: Artificial Intelligence of Things (AIoT); adversarial machine learning (AML); AI threats; AIoT security; federated defence

1 Introduction

1.1 Background and Context

The Internet of Things (IoT), initially developed through interconnected sensing platforms and embedded communication technologies, has rapidly evolved into a large-scale global ecosystem comprising more than 15 billion active devices by 2024 [1,2]. This ecosystem includes heterogeneous components such as low-power sensors, embedded microcontrollers, intelligent actuators, edge gateways, and cloud-connected

devices operating across consumer, industrial, healthcare, transportation, and smart infrastructure environments [3]. At the same time, significant advances in lightweight deep learning architectures, neural network optimization techniques, and specialized AI hardware accelerators have enabled intelligence to move from centralized cloud infrastructures toward edge environments [4,5]. This paradigm, commonly referred to as the Artificial Intelligence of Things (AIoT), enables distributed systems to perform localized inference, adaptive control, and context-aware decision-making in real time while operating under stringent latency, bandwidth, energy, and connectivity constraints [6].

The integration of AI capabilities within IoT infrastructures fundamentally transforms the security landscape. Traditional IoT security research has primarily focused on communication-layer vulnerabilities, authentication protocols, access control, and firmware integrity [7,8]. In contrast, adversarial machine learning (AML) research has mainly investigated model-centric attacks in cloud-based and computationally resource-rich environments [9]. However, AIoT systems operate at the intersection of physical sensing environments, embedded computing platforms, communication networks, and intelligent learning models, thereby introducing a substantially broader and more complex adversarial attack surface. In such environments, machine learning models directly interact with noisy and potentially adversarial physical inputs while simultaneously operating under strict memory, computational, and energy limitations [6,10].

Fig. 1 illustrates the layered integration of sensing devices, embedded intelligence, communication infrastructure, and application-layer decision-making within AIoT environments. The figure further demonstrates how adversarial perturbations introduced at the physical or sensing layer can propagate across communication, processing, and decision-making stages, ultimately affecting system-level behaviour. Such cross-layer propagation may result in misclassification, unsafe actuation, privacy leakage, service disruption, or broader system compromise. These interconnected vulnerabilities highlight the limitations of conventional single-layer security models when applied to intelligent AIoT ecosystems.

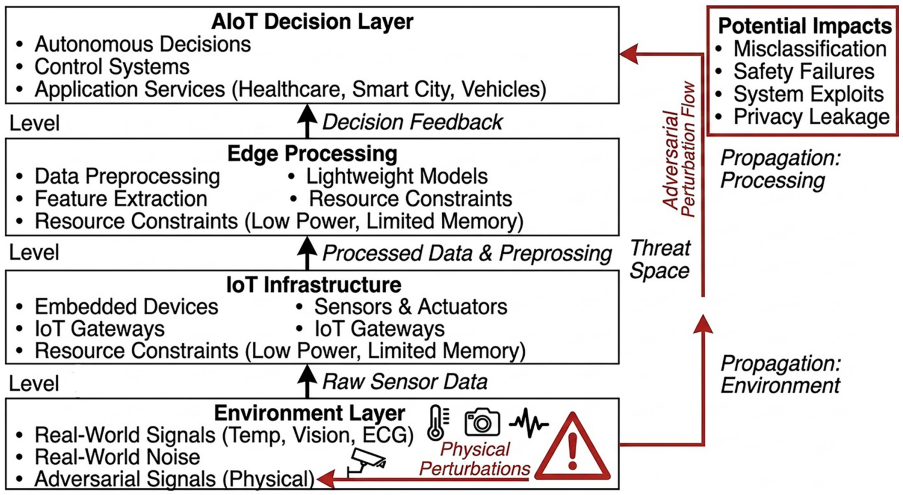

Figure 1: The depiction of the AIoT threat landscape.

Despite the increasing deployment of AIoT technologies in critical infrastructures and real-world intelligent systems, existing research remains fragmented [11]. Most IoT security surveys treat AI components as black-box modules without systematically analyzing adversarial learning threats, whereas many AML-focused reviews overlook the deployment constraints and heterogeneous operational conditions inherent to AIoT systems [12]. Consequently, there remains a lack of a unified analytical framework capable of systematically characterizing adversarial threats, defence mechanisms, and cross-layer interactions within AIoT environments. This gap limits the ability of researchers, system designers, and practitioners to accurately assess security risks, deploy effective defence strategies, and balance robustness with resource efficiency in practical intelligent systems [13,14].

Table 1 presents a comparative analysis between this study and representative state-of-the-art works selected from closely related domains, including AIoT security, adversarial machine learning, intelligent IoT architectures, cyber-physical systems, federated learning, and edge-AI environments. The studies were selected based on their relevance to AI-enabled IoT systems, distributed intelligence, security perspectives, and adversarial robustness considerations. The main comparison parameters include key idea, research gaps, and adversarial machine learning aspects, IoT layer coverage, AI pipeline stage focus, adversarial knowledge model, physical attacks, defence strategies, AIoT focus, scope, cross-layer threat analysis, and taxonomy support. In contrast to existing literature, which primarily emphasizes system optimization, application-specific architectures, or isolated security mechanisms, the revised comparison systematically maps the reviewed studies onto important dimensions of the proposed 3D-AIoT-TT framework. This analysis highlights the limited exploration of compound adversarial behaviors, cross-layer threat propagation, and physical-world attack scenarios in current AIoT research, thereby demonstrating the necessity for a unified system-level adversarial threat modeling framework.

As presented in Table 1, the majority of current research focuses on individual aspects of AIoT, such as application-specific deployments, intelligent system design, or performance optimization, while paying minimal attention to security problems.

Even when security is considered, it usually only addresses high-level strategies like authentication or privacy protection, ignoring the risks posed by AML. Furthermore, while some studies take into account layered or closed-loop architectures, they do not account for the propagation of adversarial perturbations across the layers of sensing, processing, communication, and control. Interestingly, none of the examined works clearly represent cross-layer adversarial threats or offer a hierarchical taxonomy. The proposed work, on the other hand, offers a more thorough and security-aware perspective by addressing these limitations by providing a multi-dimensional taxonomy and methodically examining emergent cross-layer adversarial risks in AIoT systems.

Table 1: Comparison of this study with the closely related state-of-the-art.

Key Parameter	[6]	[9]	[14]	[15]	[16]	[17]	[18]	[19]	**This Study**
Study Type	AIoT security survey	Cognitive IoT reasoning model	AIoT-as-a-Service (Neural Architecture Search) (NAS + RL)	Behavioral security model	AI + IoT survey	AIoT architecture framework	Domain AIoT application (aquaculture)	Digital Twin + GenAI system	Systematic review + taxonomy framework
AIoT Focus	Security & privacy	Intelligent reasoning	Automated AIoT optimization	Human behavior modeling	General AIoT applications	Data-centric AIoT systems	Domain-specific AIoT	Intelligent simulation systems	Adversarial AIoT security
Adversarial ML (AML)	FL, Edge AI, XAI	Cognitive AI concepts	NAS, RL optimization	Adaptive behavior models	General AI models	AI decision systems	Predictive AI models	GenAI + Digital Twins	Evasion, poisoning, backdoor, model extraction, FL attacks
IoT Layer Coverage	Network, Processing	Processing, Application	Processing, Network	Application, Sensing	General IoT	Full IoT stack	Sensing, Application	Full stack	Perception, Edge, Communication, Cloud
AI Pipeline Stage	Training + Inference	Not specified	Training stage	Inference stage	General workflow	Data management	Sensing + inference	Multi-stage loop	Full AI lifecycle (training → deployment → inference)
Adversarial Knowledge Model	Partial	Not defined	Not defined	Not defined	Not defined	Not defined	Not defined	Partial	White-box, Grey-box, Black-box, Physical-oracle
Physical-Layer Attacks	Not addressed	Not addressed	Not addressed	Not addressed	Not addressed	Not addressed	Not addressed	Not addressed	Sensor spoofing, physical evasion, real-world manipulation
Cross-Layer Threat Modeling	✗	✗	✗	✗	✗	Partial	✗	Partial	✓ (explicit cross-layer modeling)
Defense Mechanisms	FL, Blockchain, Cryptography	None	None	Behavioral nudging	General suggestions	Privacy tools	None	High-level robustness	Adversarial training, secure FL, runtime defense
Scope	Security review	Conceptual model	Optimization system	Behavioral focus	Broad survey	Architecture design	Domain application	System architecture	Adversarial AIoT security framework
Taxonomy Provided	✗	✗	✗	✗	✗	✗	✗	✗	✓ (3D-AIoT-TT)
Gap Coverage	Limited security view	No security focus	No adversarial focus	Partial behavioral	Weak security analysis	Domain-specific only	Application-specific	Architecture-only	Full cross-layer adversarial coverage

1.2 Motivation

Numerous converging technological and regulatory trends demand a dedicated study at this level.

- First, the rapid adoption of Tiny Machine Learning (TinyML) and on-device inference has pushed highly complex AI models onto ultra-constrained hardware platforms, typically operating with memory budgets as low as 256 KB. Many existing adversarial defensive techniques, which typically have a computing overhead of 2–10×, become unfeasible under such limitations, necessitating the development of completely new, resource-aware defence strategies [19,20].
- Second, the adversarial threat models are considerably altered by the close connection between the physical world and the AI models in AIoT systems. Unlike usual conditions where the attacks are injected digitally, the adversaries can make use of the physical channels by initiating a carefully designed optical, auditory, or electromagnetic (EM) disturbance, which can modify the inputs of the sensor before the learning systems can manage them [20,21].
- Third, the distributed and intrinsic collaborative nature of the AIoT deployments, which often utilize a federated or edge learning, reveals a novel class of weakness, i.e., the Byzantine behaviours and data poisoning attacks that seem to be different from those in centralized ML frameworks. These attack paths get worse due to the unreliable dependability of heterogeneous connectivity [21–23].
- Lastly, for the AI systems that are deployed in critical environments, the new regulatory authorities like the European Union's AI Act and the NIST (National Institute of Standards and Technology's) AI Risk Management Framework are starting to demand security, robustness, and transparency assurances [13,23].

For the adversarial threats in the AIoT ecosystem, these progressions reflect and motivate how vital it is to have an application-aware knowledge and understanding.

1.3 Search Methodology

To ensure analytical rigor, reproducibility, and methodological transparency, this study followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020) guidelines [24]. The literature search was conducted across several major scientific digital libraries, including IEEE Xplore, Web of Science, Scopus, SpringerLink, MDPI, Wiley Online Library, ScienceDirect, Association for Computing Machinery(ACM), Digital Library, and Google Scholar. The search process was performed during March–April 2026 and covered English-language publications published between January 2020 and March 2026.

A structured keyword-based Boolean search strategy was designed to systematically retrieve studies related to adversarial machine learning, AI-enabled IoT systems, and AIoT security and defense mechanisms. The general Boolean search formulation used throughout the review process is presented below:

("Adversarial Machine Learning" OR "Adversarial Attack" OR "Poisoning Attack" OR "Evasion Attack*" OR "Adversarial AI")

AND

("Artificial Intelligence of Things" OR "AIoT" OR "Intelligent IoT" OR "Edge AI" OR "Internet of Things")

AND

("Security" OR "Defense" OR "Robustness" OR "Trustworthy AI" OR "Attack Detection")

To improve reproducibility and database-level transparency, database-specific query adaptations were also employed. For example, IEEE Xplore queries were performed using the "All Metadata" field, while Scopus and Web of Science searches utilized TITLE-ABSTRACT-KEYWORD and TS (Topic Search) field

formulations, respectively. The search process primarily relied on keyword-based retrieval and relevance filtering across titles, abstracts, and indexed metadata fields.

To improve reproducibility and database-level transparency, database-specific query adaptations were also employed. For example, IEEE Xplore queries were performed using the "All Metadata" field, while Scopus and Web of Science searches utilized TITLE-ABSTRACT-KEYWORD and TS (Topic Search) field formulations, respectively. Similar syntax adaptations were applied to ACM Digital Library, ScienceDirect, SpringerLink, Wiley Online Library, MDPI, and Google Scholar to ensure consistent retrieval across platforms. Representative database-specific search query formulations are provided in Appendix A. The search process primarily relied on keyword-based retrieval and relevance filtering across titles, abstracts, and indexed metadata fields.

All retrieved records were exported to EndNote for automatic duplicate detection, followed by manual verification to ensure accurate removal of repeated entries across multiple databases. The screening process followed a multi-stage PRISMA workflow consisting of title and abstract screening, full-text eligibility assessment, and final quality-based inclusion. Initial relevance screening was independently rechecked to improve selection consistency, while disagreements regarding study eligibility were resolved through iterative discussion and consensus-based reassessment. Because not all reviewed studies explicitly define taxonomy dimensions, taxonomy mapping was performed at a conceptual aggregation level rather than via per-paper enforcement.

Studies unrelated to AIoT adversarial security, non-English publications, editorials, tutorial summaries, short papers (<4 pages), and non-peer-reviewed grey literature were excluded during the eligibility assessment stage. Following eligibility screening, a structured 10-point quality assessment framework was applied to evaluate the methodological rigor, novelty, experimental validation, reporting clarity, and domain relevance of the candidate studies. Only studies achieving a minimum quality score of 6/10 were retained in the final synthesis. The complete PRISMA workflow is summarized in Table 2 and illustrated in Fig. 2. Furthermore, the PRISMA 2020 checklist has been included in the Supplementary Materials to further strengthen methodological transparency, reproducibility, and reporting consistency. In addition, the detailed 10-point quality assessment rubric used during study selection is provided in Appendix B.

Table 2: The flow of the PRISMA process.

PRISMA Stage	Digital Libraries/Actions	Inclusion/Exclusion Criteria	Outcome
Identification	IEEE Xplore, Web of Science, SpringerLink, Scopus, MDPI, Wiley Online Library, ScienceDirect, ACM Digital Library, Google Scholar	Publications from January 2020–March 2026; English-language only	~1100 records identified
Deduplication	Automatic and manual duplicate removal executed via EndNote.	Removal of repeated entries across databases	~820 unique records retained
Title & Abstract Screening	Initial relevance screening based on AIoT adversarial security scope	Non-AIoT, non-security, and irrelevant studies were excluded	412 records retained

(Continued)

Table 2 (continued)

PRISMA Stage	Digital Libraries/Actions	Inclusion/Exclusion Criteria	Outcome
Full-Text Eligibility	Detailed evaluation of methodological relevance and technical contribution	Editorials, short papers (<4 pages), tutorial summaries, and non-research articles excluded	178 full-text articles assessed
Quality Assessment & Inclusion	10-point quality scoring framework applied	Studies scoring <6/10 and non-peer-reviewed grey literature excluded	115 studies included

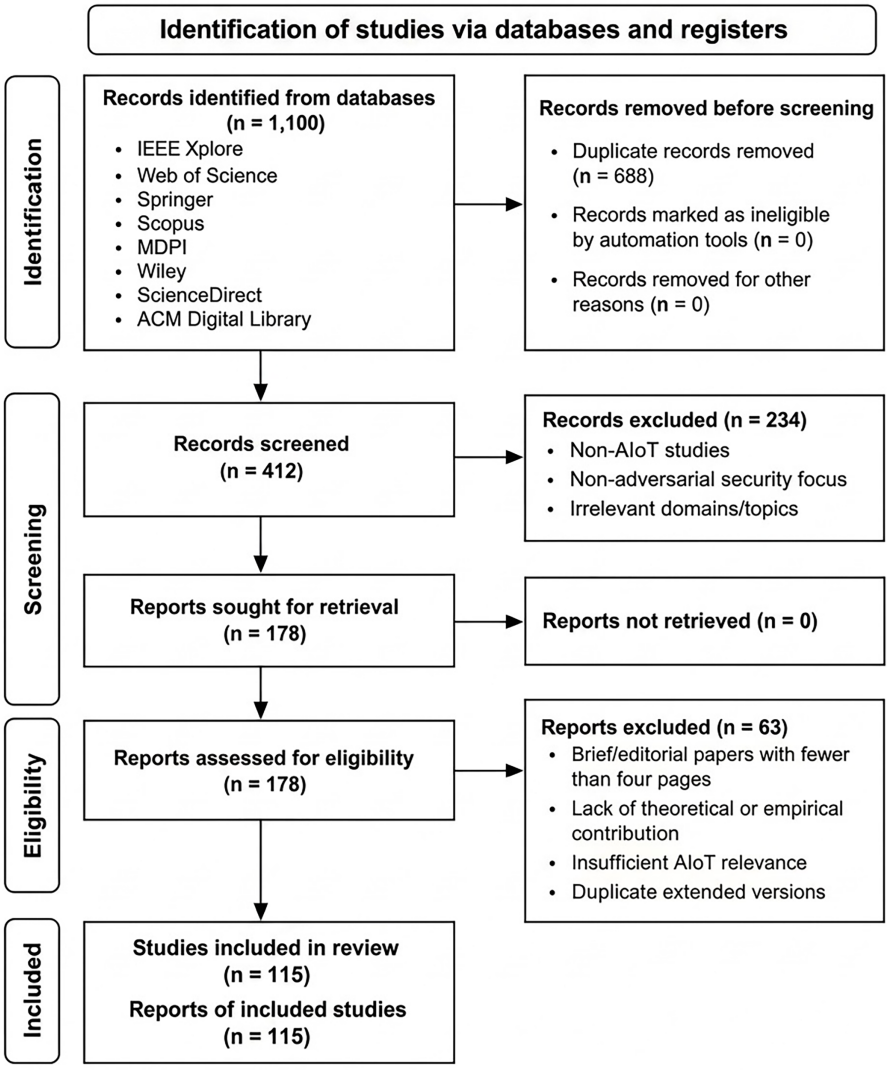

Figure 2: PRISMA workflow diagram.

Three research questions (*RQs*) are organized in this study that are centered on this carefully chosen dataset and are meant to thoroughly examine the AIoT adversarial landscape:

- ***RQ1:*** Threats Classification: In AIoT systems, what makes the entire landscape of adversarial AI threats, and how do they vary structurally from threats in conventional IoT or AI or domains when studied across AI pipeline stages, IoT architectural layers, and adversarial knowledge models?
- ***RQ2:*** Defence Mechanisms Effectiveness in AIoT: In AIoT systems, what are the pros and cons of existing adversarial defence mechanisms, and for the deployment in places with limited resources and heterogeneity, how well are they suited?
- ***RQ3:*** Emergent Cross-Layer Threats: From the integration of AI and IoT components, what new adversarial patterns particularly arise, and which systemic flaws are exploited by these cross-layer threats?

1.4 Contributions

The following are the key contributions of this study:

- **PRISMA-based Development and Evaluation:** According to the PRISMA 2020 method, a precisely chosen quantity of 115 research articles is developed, offering a comprehensive and systematically clear description of adversarial AI research in AIoT situations.
- **3D-AIoT-TT (Three-Dimensional Adversarial IoT Threat Taxonomy):** This study proposes a systematic 3D-AIoT-TT framework for AIoT systems that classifies adversarial threats based on (i) AI pipeline stages, (ii) IoT architectural layers, and (iii) adversarial knowledge levels. The proposed taxonomy enables structured cross-study comparison and facilitates the identification of unexplored threat regions within the AIoT security landscape.
- **Comprehensive Analysis of Adversarial Attacks and Defences:** The main types of adversarial attacks are provided in this study, including evasion, poisoning, backdoor, federated learning, and model extraction attacks, along with an organized analysis of defence mechanisms such as adversarial training, detection-based techniques, federated defences, and hardware-assisted security approaches.
- **Overview of Cross-Layer Adversarial Threats:** This study goes beyond conventional AI and IoT security perspectives to identify and systematically investigate new cross-layer adversarial threats arising from the interaction of sensing, learning, communication, and actuation layers. This highlights systemic weaknesses that traditional single-layer security approaches are unable to detect.

1.5 Paper Organization

This is how the rest of the paper is organized. Background information on adversarial machine learning foundations and AIoT architectures is given in Section 2. The proposed 3D-AIoT-TT taxonomy is presented in Section 3. Within this framework, adversarial attack techniques and defence mechanisms are methodically reviewed in Sections 4 and 5, respectively. Emergent cross-layer threats are examined in Section 6, and by summarizing the findings in connection to the research questions (RQ1–RQ3), Section 7 provides an integrated assessment of the findings. Section 8 finally concludes the paper.

2 Background and Preliminaries

The basic concepts required for understanding the fusion of IoT systems with the use of AI are presented in this section. It highlights architectural components of AIoT and emphasizes the major enabling technologies that support distributed, intelligent decision-making. It also creates the framework for examining adversarial threats and security issues in AIoT environments.

2.1 AIoT Architectural Overview

The architecture of AIoT systems is dispersed and hierarchical by nature, with sensing, computing, and intelligence spread over several layers. Real-time analytics, low-latency decision-making, and scale model training are made possible by this architectural paradigm, but its distributed trust boundaries and heterogeneous components also present difficult security issues [25]. The perception, edge, fog, and cloud layers of an AIoT architecture can be thought of as a four-tiered framework with different functional roles and computing capacities [26]. As depicted in Fig. 3, intelligence is dispersed among various layers rather than being restricted to centralized infrastructures, allowing for both global coordination and localized inference.

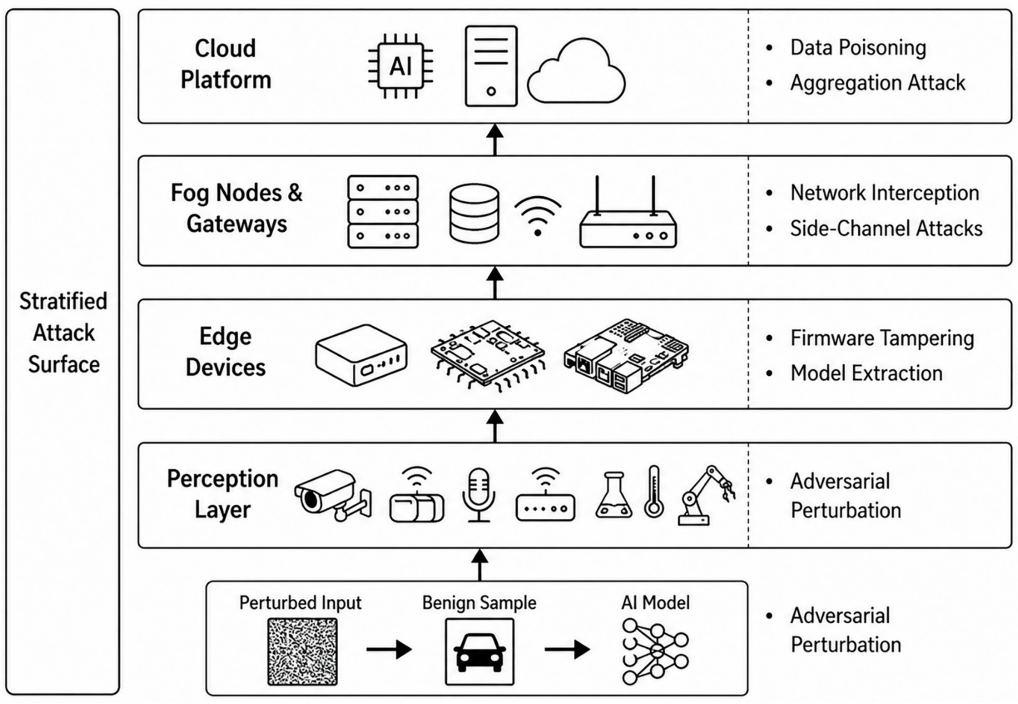

Figure 3: Mapping intelligence and attack vectors across the AIoT continuum, from perception to cloud.

The perception layer, which is the lowest level, consists of a variety of sensing and actuation components, including radio frequency identification (RFID) tags, cameras, microphones, inertial measurement units (IMUs), and chemical sensors [27]. Through direct interaction with the physical world, these devices create raw data, which is then fed into the AI model. As for their inappropriate computational security and physical exposure, this layer is primarily susceptible to adversarial perturbation, where well-constructed inputs can modify sensor readings and mislead downstream inference systems.

The edge layer above this is comprised of single-board computing systems, microcontrollers, and embedded processors, which run lightweight ML models that normally use the TinyML paradigm [26]. This layer allows on-device inference by reducing communication latency and preserving bandwidth. However, it also introduces vulnerabilities like side-channel attacks, firmware tampering, and model extraction because adversaries can exploit direct access to the device software stack or hardware.

As a midway processing tier, the fog layer comprises edge servers and gateways that collect data from various edge devices [28]. More coordinating and complex tasks are offered by this layer, like distributed inference and data preparation. This layer serves as a communication hub, so it is prone to network-level

attacks like eavesdropping, traffic manipulation, and the injection of malicious data streams, which can distribute compromised information across the entire system.

The cloud layer at the top level offers central computational resources for vast data storage, orchestration, and model training [29] as it is vital when combining model updates from various edge nodes in distributed or FL contexts. Even with its resistance, this layer can be targeted by Byzantine attacks and data poisoning, in which an attacker modifies the training data or model updates to influence the overall model performance.

Each inter-layer border on the tire surface created by the hierarchical distribution of intelligence among these layers denotes a possible site for adversary involvement [30]. As depicted in Fig. 2, at any layer, attacks can initiate and propagate across the system to raise their impact. For instance, false updates at the cloud layer can impact all linked devices, and questionable inputs at the perception layer can disrupt edge inference. Additionally, the adversaries can target not just data and communication pathways, but also the learning mechanisms themselves, the deployment of AI models across several tiers greatly increases the complexity of protecting AIoT systems.

2.2 Core Concepts of Adversarial Machine Learning (AML)

AML research examines how ML models are susceptible to intended inputs and manipulations that can change the behavior of the model. Initial studies in this area reflect that deliberate misclassification or significantly reduced model performance can result from well-planned perturbations, which are frequently undetectable to human observers [7]. Although these findings are initially studied in centralized and static ML environments, their ramifications become more substantial in AIoT systems, where the models operate under resource-constrained, distributed, and dynamic conditions [31].

ML models that process continuous streaming data instead of static datasets are commonly implemented across edge and embedded devices in AIoT contexts. Additional limitations brought forth by this change include constrained computational power, inconsistent connectivity, and close ties to the real world [32]. As a result, the threat environment becomes more intricate and multidimensional as the traditional AML framework needs to be expanded to take into consideration real-time operation, physical reliability of attacks, and system-level interactions. The stage of the ML lifecycle that adversarial attacks target is usually used to classify them. Three main categories of adversarial attacks can be distinguished, as shown in Fig. 4: privacy or model-centric attacks, inference-time attacks (evasion), and training-time attacks (poisoning).

The goal of training-time attacks, also known as poisoning attacks [33], is to undermine the learning process by altering the model parameters or training dataset. Attackers can insert malicious samples, change labels, or include backdoors or secret triggers that are inactive during normal operations but become active under particular circumstances. In AIoT systems, these risks raise the possibility of malware data injection and aggregation level manipulation, especially in distributed and FL environments where various edge devices contribute to model training [34]. Evasion attacks or inference-time attacks [35], likewise, arise in the deployment phase. Without changing the model itself, these attacks provide suspicious predictions by producing adversarial inputs that exploit the model's decision boundaries. These attacks in the AIoT context can be physically possible disruptions like altered sensor inputs, distorted visual patterns, or ambient manipulations that can trick perception systems in the practical world [6].

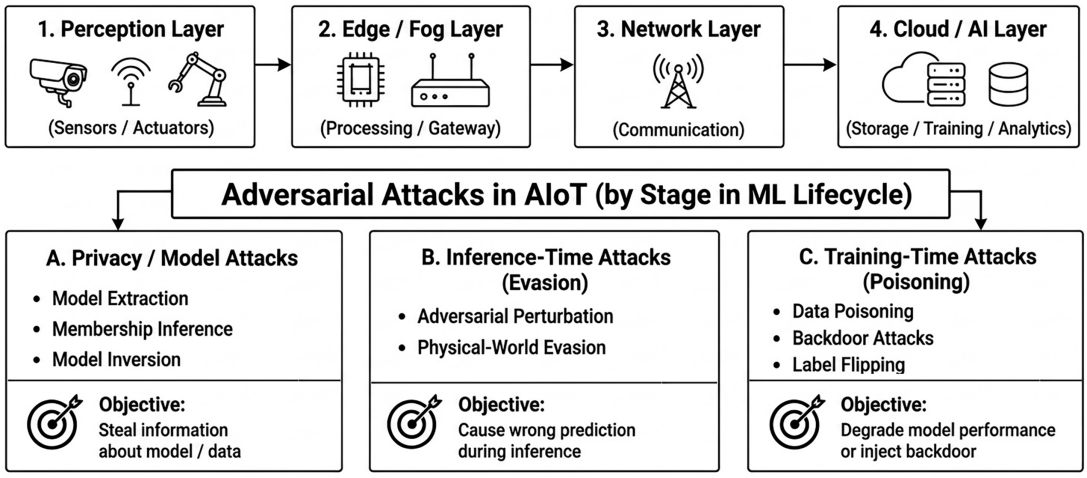

Figure 4: Classification of AML attacks in the AIoT systems.

Privacy and model-centric attack is the third category of attack, that target the integrity and confidentiality of ML models [36]. They comprise model extraction attacks, in which the adversaries try to duplicate private models by querying them, and membership inference attacks, which seek to determine if particular data points are included in the training dataset [37]. Duvh attacks provide significant dangers to users' intellectual property and privacy in AIoT systems, where models are frequently put on accessible edge devices or exposed through Application Programming Interfaces (APIs).

In AIoT systems, the formation of various threat classes is impacted by several distinct variables. While hardware heterogeneity improves model vulnerability and performance unpredictability, resource constraints restrict the viability of computationally expensive complex countermeasures [38]. Furthermore, the tight link with the physical world allows for hybrid cyber-physical attack vectors that are not present in classic ML systems, and unpredictable network connectivity may impede coordinated defence efforts [39]. Fig. 2 highlights that these adversarial risks are linked across the ML pipeline, with each attack type focusing on a different stage of the model lifecycle. Additional constraints brought about by the AIoT context, like scarce resources, a variety of hardware platforms, and real-world interactions, affect the viability of attacks and the creation of effective defensive strategies.

This thorough comprehension of AML principles serves as the basis of creating an organized taxonomy of risk unique to the AIoT, which is discussed in the following sections.

2.3 Threat Model Formalization

In AIoT systems, to systematically analyze the adversarial behavior, a multi-dimensional threat modelling approach is used in this study, which incorporates the intricate nature of interactions during the ML lifespan and distributed system architecture. This formalization offers the analytical foundation for the proposed taxonomy and enables a consistent portrayal of adversarial threats across different AIoT implementations. As shown in Fig. 5, the proposed threat model views an adversarial instance as a composition of multiple independent dimensions, each of which characterizes a unique aspect of the attack surface. The framework is first defined along its primary axes that represent the structural characteristics of adversarial encounters:

- **Pipeline stage (*s*):** The first dimension illustrates the point in the ML lifecycle where the threat has begun. This includes stages like data collection, preparing, model training, or model distribution, and concluding [29]. Attack stages exhibit distinct operational behaviours like poisoning attempts, mainly targeting the training phase, while evasion attacks occur during inferences.
- **IoT architectural layer (*l*):** In an AIoT system, the adversary's targeted layer is indicated by the second layer, and it includes perception, edge, communication, and cloud/aggregation tiers, each of which has different computational capabilities and exposure levels [40]. This dimension is vital for accurately simulating attack propagation in dispersed scenarios, as vulnerabilities vary significantly between various layers, as was discussed in Section 2.1.
- **Adversarial knowledge level (*k*):** The extent to which the adversary can access the internal operations of a targeted system is measured by this third dimension. This dimension is altered to reflect actual AIoT situations in which adversaries might gain partial access via modified hacked edge devices or side-channel observations [7]. It follows the conventional classification of white-box, gray-box, and black-box settings. To overcome this, the model incorporates two additional factors that reflect physical world deployment situations.
- **The Adversarial Goal (*g*):** it describes the purpose of the attack, which may include backdoor installation, model extraction, denial-of-service (DoS), or targeted or untargeted misclassification [41,42]. These objectives show the various motives of adversaries in AIoT ecosystems by capturing both functional disruption and information-oriented threats.
- **Constraint profile (*c*):** Replicating deployment boundaries in the real-world, defined as a tuple, i.e., $c = (\Delta, E, T)$, where Δ indicates perturbation magnitude, E reflects energy constraints, and T captures delay requirements. Consequently, an adversarial AIoT instance can be expressed as:

$$A = (s, l, k, g, c) \tag{1}$$

By integrating lifecycle, architectural, knowledge-based, and operational viewpoints into a single framework, this formulation offers a thorough and expandable depiction of adversarial behavior. These elements are interrelated and together define the features of an aggressive AIoT threat.

Compared to typical threat models that focus on discrete areas of the attack surface, this multi-dimensional method enables cross-layer and cross-stage analysis, making it simpler to uncover complex attack patterns and connections. It also makes it easier to compare different attack methods consistently and provides a formal basis for developing context-specific and constraint-aware response systems. The multi-layered taxonomy of adversarial threats described in Section 3, where these dimensions are methodically mapped to classify and examine new attack avenues in AIoT systems, is directly supported by this formal threat model.

The proposed multi-dimensional threat model, AML paradigms, and the integrated conceptual framework of AIoT architecture are summarized in Table 3, emphasizing their interdependencies and applicability to AIoT security analysis.

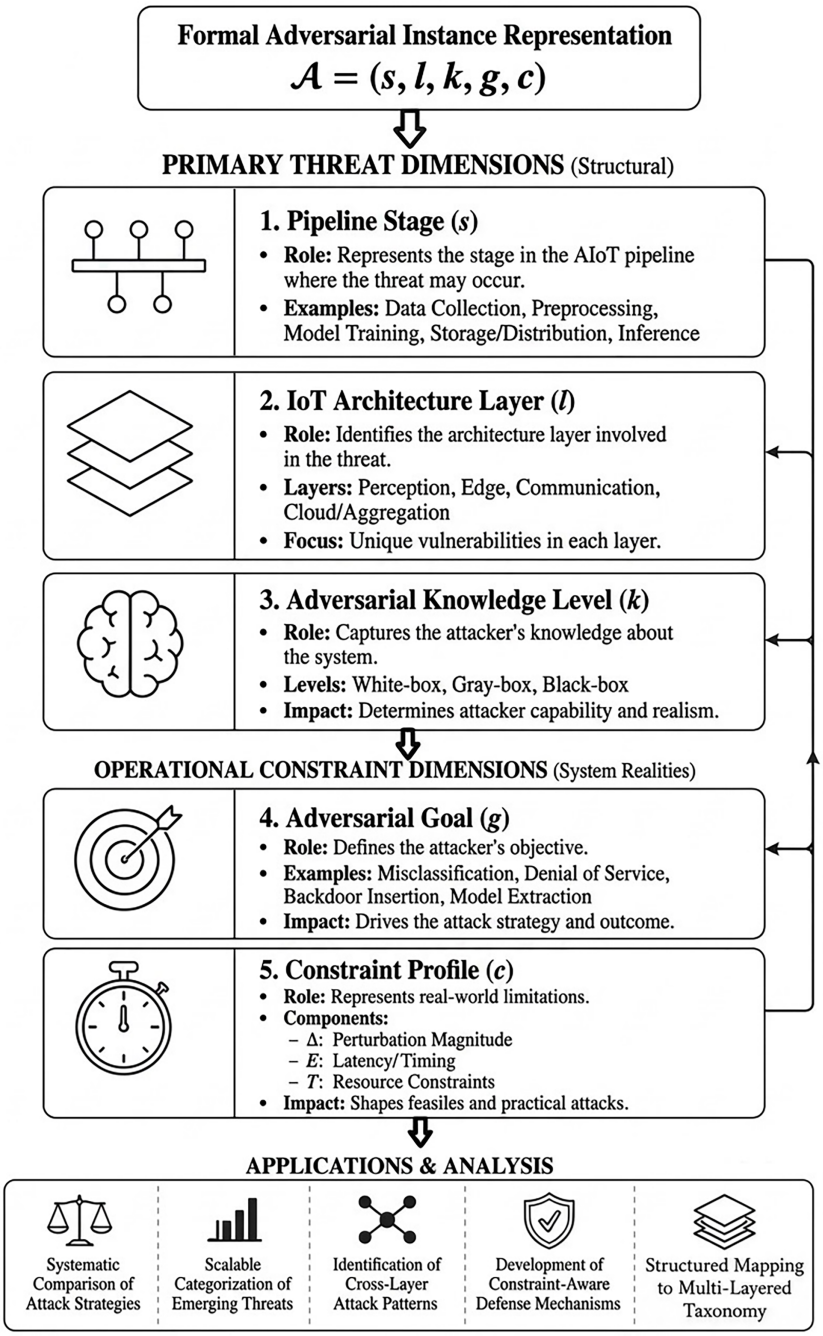

Figure 5: A simplified threat model framework of AIoT.

3 Proposed Taxonomy: 3D-AIoT-TT

To address the complex and multi-layered nature of adversarial threats in AIoT systems, this study proposes a unified and structured framework termed the 3D-AIoT-TT (Three-Dimensional AIoT Threat Taxonomy). The taxonomy provides a systematic representation of adversarial behaviors by modeling the AIoT threat landscape as a three-dimensional space defined by: (i) AI pipeline stage, (ii) IoT architectural

layer, and (iii) adversarial knowledge level. By capturing the interaction between sensing, learning, communication, and control components, the framework enables structured analysis, comparison, and defence design across heterogeneous AIoT environments.

Table 3: Integrated conceptual framework of AIoT architecture and adversarial threat modeling.

Component	Sub-Aspect	Description	Security Implications	Link to Threat Model Dimensions
AIoT Architecture (Section 2.1)	Perception Layer	Sensors and actuators interacting with the physical environment	Vulnerable to physical adversarial perturbations and spoofing	(l), (c)
	Edge Layer	Embedded devices performing local inference (TinyML)	Susceptible to model extraction, firmware tampering, and side-channel attacks	(l), (k), (c)
	Fog Layer	Intermediate gateways for aggregation and processing	Risk of data injection, traffic manipulation, and aggregation compromise	(l), (s)
	Cloud Layer	Centralized training and orchestration	Vulnerable to poisoning, Byzantine attacks, and large-scale manipulation	(l), (s), (g)
Adversarial ML Concepts (Section 2.2)	Training-Time Attacks	Poisoning of training data or model parameters	Degrades model integrity and introduces hidden backdoors	(s), (g), (k)
	Inference-Time Attacks	Adversarial inputs causing misclassification	Exploits decision boundaries during real-time deployment	(s), (c)
	Model/Privacy Attacks	Model extraction and membership inference	Leads to intellectual property theft and privacy leakage	(k), (g)
	AIoT Constraints	Resource, latency, and physical interaction limitations	Restricts both attack feasibility and defense design	(c)
Threat Model (Section 2.3)	Pipeline Stage	Attack point in ML lifecycle	Defines temporal attack entry	(s)
	Architecture Layer	Targeted system layer	Defines spatial attack surface	(l)
	Knowledge Level	Adversary system awareness	Influences attack sophistication	(k)
	Adversarial Goal	Intended outcome of the attack	Determines attack strategy	(g)
	Constraint Profile	Operational limitations ((Δ, E, T))	Governs feasibility in the AIoT context	(c)

It is important to clarify that the proposed taxonomy is strictly based on these three orthogonal dimensions. Additional aspects, such as adversarial goals and constraint profiles, are treated as supporting contextual descriptors rather than independent classification axes, used only to enrich the interpretation of attacks in practical scenarios.

3.1 Rationale and Design Principles

Existing AML taxonomies are predominantly single-axis or dual-axis systems, typically organized around attack objectives (targeted vs. untargeted) or adversarial knowledge (white/grey/black-box) [43]. While these models are effective for benchmarking isolated learning systems, they are insufficient for AIoT environments, which exhibit coupled cyber-physical interactions, distributed computation, and resource-constrained deployment settings.

Similarly, traditional IoT security taxonomies primarily focus on network and protocol-level threats, failing to capture vulnerabilities introduced by embedded intelligence and learning-based decision-making [44]. Consequently, the AIoT threat surface emerges as a composite interaction of physical, computational, and learning-driven attack vectors that cannot be adequately represented using existing paradigms.

To bridge this gap, the proposed 3D-AIoT-TT framework models adversarial threats across three interacting but orthogonal dimensions. Importantly, the taxonomy is NOT sequential; instead, it represents a unified multidimensional threat space where each axis interacts jointly to define an attack coordinate. The design of this taxonomy is inspired by the following five principles:

i. **Completeness:** For each known AIoT adversarial threat, the taxonomy can explicitly incorporate.
ii. **Orthogonality:** Each axis reduces the redundancy and overlap by capturing the distinct aspects of adversarial behavior.
iii. **Unambiguous Mapping:** In the taxonomy space, by mapping the occurrence of each attack to an individual coordinate, consistent classification can be possible.
iv. **Actionability:** By each taxonomy region, different system design issues and defence tactics are implied.
v. **Constraint Awareness (a new contribution):** The taxonomy allows for evaluation that goes beyond threat models, which are only theoretical, by indirectly encoding deployment restrictions like physical reliability, latency, and energy.

3.2 Taxonomy Axes

3.2.1 Axis 1: AI Pipeline Stage

The first dimension captures the lifecycle stage of the AI pipeline where an adversarial intervention occurs. In AIoT systems, the AI lifecycle is modeled as five interconnected stages: data collection, pre-processing, training, model distribution/storage, and inference. The data collection stage is the foundation of the pipeline, where real-world signals are captured via sensors. Pre-processing involves filtering, normalization, and feature extraction. Training involves model learning using centralized, federated, or transfer learning paradigms. Model distribution includes deployment via edge/cloud systems and over-the-Air (OTA) updates. Finally, inference performs real-time decision-making.

From a security perspective, different stages exhibit different exposure levels. In practical AIoT deployments, adversarial access is particularly feasible at sensing and inference boundaries due to the physical exposure of sensors, wireless interfaces, and resource-constrained edge devices. Such exposure has been extensively discussed in embedded neural network and edge-AI security literature, where attackers may gain physical proximity or side-channel access to deployed devices [45]. Early-stage attacks (especially data collection and training) tend to have persistent system-wide effects, while inference-stage attacks are often immediate but localized.

3.2.2 Axis 2: IoT Architectural Layer

The second dimension represents the system layer where adversarial interaction occurs: perception layer, edge layer, communication layer, and cloud/aggregation layer. The perception layer is the primary entry point for physically grounded attacks through sensors and actuators. The edge layer hosts embedded models for local inference. The communication layer enables data exchange and is vulnerable to interception and injection attacks. The cloud layer manages centralized training, orchestration, and storage.

A key feature of AIoT systems is cross-layer dependency, where disturbances at one layer propagate non-linearly to others. For example, small perturbations at the perception layer (e.g., sensor spoofing) can influence edge inference outputs, which may further affect cloud-level model updates in federated learning (FL) systems, leading to system-wide degradation.

3.2.3 Axis 3: Adversarial Knowledge Level

According to the adversary system, access and the degree of knowledge, the proposed taxonomy improves and integrates upon the conventional AML threats (white/grey/black-boxes) pattern by classifying attacks. For highly optimized and targeted attacks, the white box situations allow as the adversary has full access to the model parameters, architecture, and training process. For grey-box situations, the adversary has partial knowledge, like limited training data insights, approximation model structure, and side channel information. For a black-box scenario, the adversary only has output observation or query-based interactions to deduce the behavior of the system, as they do not have any inside knowledge of the model.

In addition, the study introduces the physical-oracle setting, which is particularly relevant in AIoT environments. In this setting, the adversary does not directly observe model outputs but instead infers system behavior through physical interactions such as sensor readings, actuator responses, environmental changes, or system feedback loops. For example, an attacker may vary environmental conditions (light, sound, vibration, or RF interference) and observe system reactions to infer decision boundaries. This paradigm is especially relevant in autonomous systems, industrial IoT, and healthcare devices, where outputs are embodied in physical actions rather than explicit digital responses.

3.3 Taxonomy as a Structural Threat Space

The proposed 3D-AIoT-TT framework represents adversarial threats as a unified three-dimensional space rather than a sequential classification pipeline. Each attack is mapped to a coordinate defined by (AI pipeline stage, IoT layer, and adversarial knowledge level).

Importantly, the framework captures complex interactions between axes rather than treating them independently. For instance, a sensor poisoning attack (data collection stage + perception layer + physical-oracle access) can propagate into corrupted model training outcomes, which later amplify inference-time misclassifications at the edge layer. Similarly, a black-box evasion attack at the inference stage, combined with edge-layer constraints (limited compute and energy), can significantly increase system vulnerability due to reduced defense capacity. This interaction-based modeling enables cross-layer reasoning, allowing the analysis of cascading adversarial effects across multiple system components.

3.4 Taxonomy Instantiation

Table 4 demonstrates the instantiation of the 3D-AIoT-TT framework across representative adversarial attack classes. Each attack is mapped to a unique coordinate in the 3D space, enabling systematic classification and comparison.

Table 4: The 3D-AIoT-TT taxonomy instantiation across ten representative attack classes.

Attack Class	Pipeline Stage (s)	IoT Layer (l)	Knowledge Level (k)	Description	Key Impact	Typical Defense Strategy
Physical-Evasion-Black Box	Inference	Perception Layer	Black-box	Adversarial perturbations applied in the physical world to mislead sensors	Misclassification → unsafe actuation	Sensor filtering, multi-sensor validation
Digital-Evasion-Edge	Inference	Edge Layer	Black/Gray-box	Crafted digital inputs targeting edge-deployed models	Local prediction errors	Model robustness, runtime monitoring
Data Poisoning (Centralized)	Training	Cloud Layer	White/Gray-box	Injection of malicious samples during centralized training	Global model corruption	Data sanitization, anomaly detection
Federated Model Poisoning	Training	Cloud/Aggregation	Gray/Black-box	Malicious client updates in federated learning	Backdoor insertion, biased aggregation	Robust aggregation, client auditing
Sensor Manipulation Attack	Data Collection	Perception Layer	Physical-Oracle	Physical interference with the sensing process (noise, spoofing)	Corrupted input data	Sensor hardening, redundancy
Feature Manipulation Attack	Pre-processing	Edge Layer	Gray-box	Tampering with feature extraction or normalization	Distorted feature space	Secure preprocessing pipelines
OTA Trojan Attack	Distribution	Network/Edge	White/Gray-box	Malicious model injected via OTA update	Latent malicious behavior	Secure OTA, cryptographic attestation
Model Extraction Attack	Inference	Edge/Cloud	Black-box	Query-based extraction of model parameters	IP leakage, cloning	Query limiting, watermarking
Membership Inference Attack	Inference	Cloud Layer	Black/Gray-box	Inferring training data membership from outputs	Privacy leakage	Differential privacy, output perturbation
Cross-Device Poisoning	Training	Distributed Edge	Gray-box	Coordinated poisoning across multiple devices	Amplified global impact	Trust scoring, contribution auditing
Energy-Latency Sponge Attack	Inference	Edge Layer	Black-box	Inputs designed to maximize inference cost	Resource exhaustion (DoS)	Input complexity control, inference budgeting
Cross-Layer Cascade Attack	Multi-stage	Multi-layer	Any	Attack propagates across sensing, learning, and control layers	System-wide cascading failures	Cross-layer defense, system-level monitoring

This mapping highlights that attacks targeting the same AI pipeline stage may exhibit fundamentally different behaviors depending on their architectural layer and knowledge model. For example, a perception-layer evasion attack requires sensor-level defenses such as redundancy and filtering, whereas edge-layer attacks require runtime monitoring and adversarial robustness mechanisms.

The taxonomy also captures compound and cross-layer attacks, where adversaries exploit multiple stages and layers simultaneously. For example, a federated poisoning attack may originate at the training stage (pipeline axis), propagate through the cloud aggregation layer, and exploit grey-box knowledge to manipulate global model updates. Such multi-stage threats demonstrate cascading failure patterns that are not captured in traditional single-axis taxonomies. These observations highlight the necessity of integrated defense mechanisms that operate across multiple system dimensions rather than isolated security solutions.

It should be noted that the 115 studies included in this review comprise adversarial attack studies, defense-oriented works, AIoT security frameworks, FL research, and foundational AIoT surveys. Consequently, not every study corresponds to a specific adversarial instance that can be uniquely positioned within the proposed taxonomy. The taxonomy instantiation presented in Table 4 therefore focuses on representative attack and threat classes extracted from the reviewed literature, while the broader body of reviewed studies collectively informed the taxonomy design and dimensional structure.

Fig. 6 illustrates the proposed 3D-AIoT-TT framework as a unified multidimensional threat space. The three axes jointly define a coordinate system in which adversarial behaviors are positioned based on their pipeline stage, architectural layer, and knowledge level.

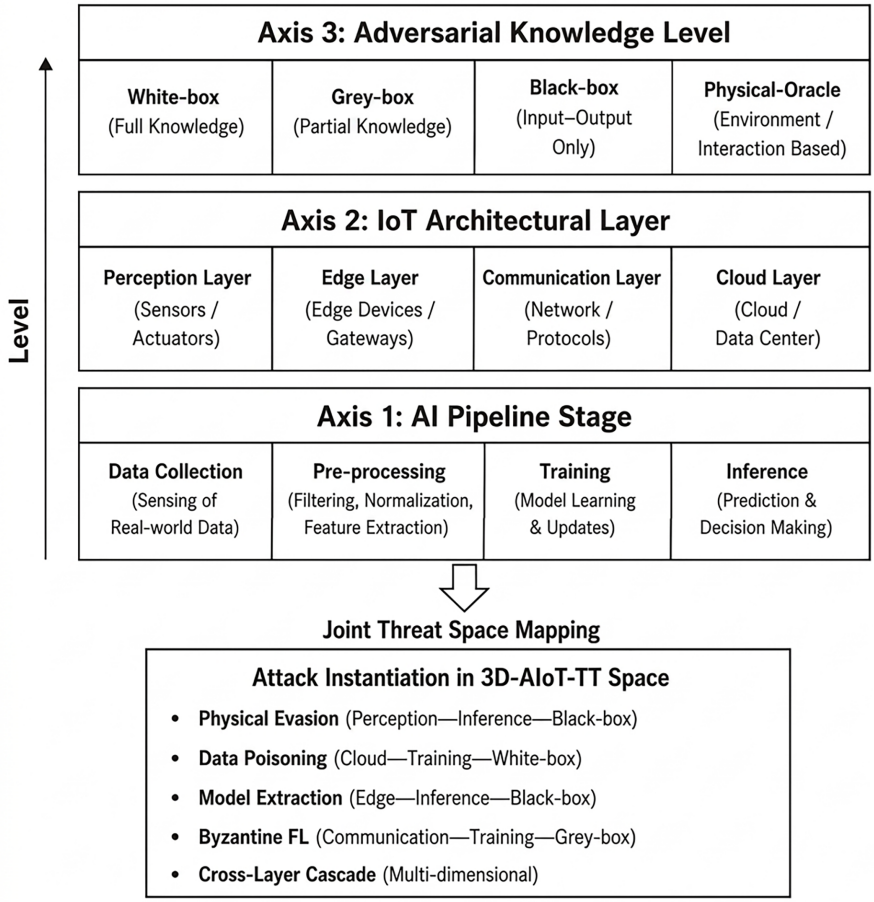

Figure 6: The 3D-AIoT-TT framework proposed model.

Unlike traditional linear taxonomies, this model captures both isolated and interacting adversarial behaviors, enabling a structured representation of cross-layer dependencies and compound threat scenarios in AIoT systems.

The overall summary of the key dimensions, their components, description, security implications, and related functions of the proposed taxonomy in illustrating adversarial threats in the AIoT systems, as presented in Table 5, serves to put more solidification of its structure and systematic scope.

Table 5: The overall summary of the proposed 3D-AIoT-TT taxonomy.

Dimension	Sub-Components	Description	Security Implications	Example Attack Types
AI Pipeline Stage (s)	Data Collection, Pre-processing, Training, Storage/Distribution, Inference	Defines the stage in the AI lifecycle where the adversary intervenes	Early-stage attacks propagate system-wide; inference attacks are localized but easier to execute	Data Poisoning, Backdoor Injection, Adversarial Examples
IoT Architectural Layer (l)	Perception, Edge, Communication, Cloud	Identifies the system layer where the attack is physically or logically applied	Enables distinction between physical and cyber-attack surfaces; impacts defense deployment.	Sensor Spoofing, Firmware Attack, Man-in-the-Middle (MITM), Byzantine Attack
Adversarial Knowledge Level (k)	White-box, Grey-box, Black-box, Physical-Oracle	Captures the adversary's knowledge and access capabilities	Determines attack feasibility, stealth, and required defense complexity	Model Extraction, Query-based Attacks, Side-channel Inference
Adversarial Goal (g)	Misclassification, Backdoor, Model Extraction, Denial-of-Service	Specifies the objective of the attack	Influences evaluation metrics and defense strategies	Targeted Attack, Trojan Attack, Resource Exhaustion
Constraint Profile (c)	Perturbation Budget (Δ), Energy (E), Latency (T)	Reflect real-world deployment constraints in AIoT systems	Limits both attack and defense feasibility under edge conditions	Sponge Examples, Low-power Attacks, Real-time Evasion

4 Adversarial AI Attacks in the AIoT: Analysis

This section presents a systematic inspection of AML attacks in the AIoT systems, highlighting their operational methods and effects across multiple layers.

4.1 Evasion Attacks

4.1.1 Digital Evasion at the Edge

Evasion attacks' main goal is to create an adversarial perturbation that creates an inaccurate model prediction, while remaining statistically indistinguishable or undetectable from valid inputs. In AIoT edge deployments, the adversary regular use limited query settings or black box making use of either model APIs or indirect input from system outputs [46].

For producing adversarial instances in continuous input spaces (L∞, L2, L0 norms), gradient-based techniques like Projected Gradient Descent (PGD) [47], Carlini & Wagner (C&W) [48], and their adaptive variations continue to be fundamental. However, because of model compression methods like quantization, pruning, and distillation, AIoT deployments have special limitations [30]. At inference time, these modifications result in non-smooth loss surfaces, which decrease direct transferability from full-precision surrogate models and encourage quantization-aware adversarial optimization techniques.

A growing number of research focuses on edge-runtime evasion, which involves injecting adversarial manipulation directly into sensor streams before preprocessing [49]. This includes driver-layer manipulation, Direct Memory Access (DMA) level interception, and compromised sensor firmware [50]. This successfully shifts the attack surface from the model boundary to the data acquisition layer by completely avoiding conventional input sanitization pipelines, in contrast to typical evasion.

Adaptive evasion under concept drift [51] is a recently developed approach in which attackers take advantage of continually learning AIoT systems by progressively changing input distributions, resulting in model degradation without setting off anomaly detectors [52].

4.1.2 Physical and Cyber-Physical Adversarial Attacks

Evasion attacks are extended into the real world by physical adversarial attacks, which necessitate resilience against environmental changes, including illumination, occlusion, viewpoint fluctuation, and sensor noise. Adversarial patches continue to be the utmost instantiation because of their robust transferability across localized structures and vision models [53]. In addition to static vision systems, AIoT creates multi-model physical attach surfaces like:

- **Optical attacks:** Reflects adversarial stickers, projected patterns, and disturbances that target camera-based perception systems in smart surveillance and robotics [54].
- **Acoustic attacks:** Target keyword spotting and voice recognition models in smart assistants by injecting structured and ultrasonic noise [55].
- **Wireless and RF Adversarial Interference:** In AIoT literature, it is a very novel and understudied attack vector that involves manipulation in radio signals to impact RF-based sensors, such as occupancy detection and gesture recognition [56].

Sensor spoofing and Light Detection and Ranging (LiDAR) attacks, in which signal replay or controlled laser injection creates phantom objects or restrains physical world detections in autonomous systems, are considered an important category [57]. These attacks directly spread to control-layer decisions in AIoT-enabled cyber-physical systems, making them adversarial dangers that are critical to safety. Cross-modal physical attacks [58], in which disruptions in one modality (such as audio or radio frequency) cause failures in another modality due to sensor fusion dependencies in multimodal AIoT systems, are a new trend.

Fig. 7 shows the general workflow of adversarial evasion attacks in the AIoT systems. It identifies three main attack vectors: physical/cyber-physical attacks, adaptive evasion under concept drift, and digital evasion at the edge. Adversarial perturbation creation using query-based or gradient-based algorithms and

sensor-level manipulation are the first steps in digital escape. By gradually modifying input distributions, adaptive evasion takes advantage of ongoing learning to get around anomaly detection systems. Concurrently, optical, acoustic, and radiofrequency interference are used by physical and cyber-physical attacks to target multimodal AIoT systems. Together, these attack routes cause the AIoT edge model to gradually deteriorate, which eventually compromises inference outputs, including incorrect classification and risky system behavior.

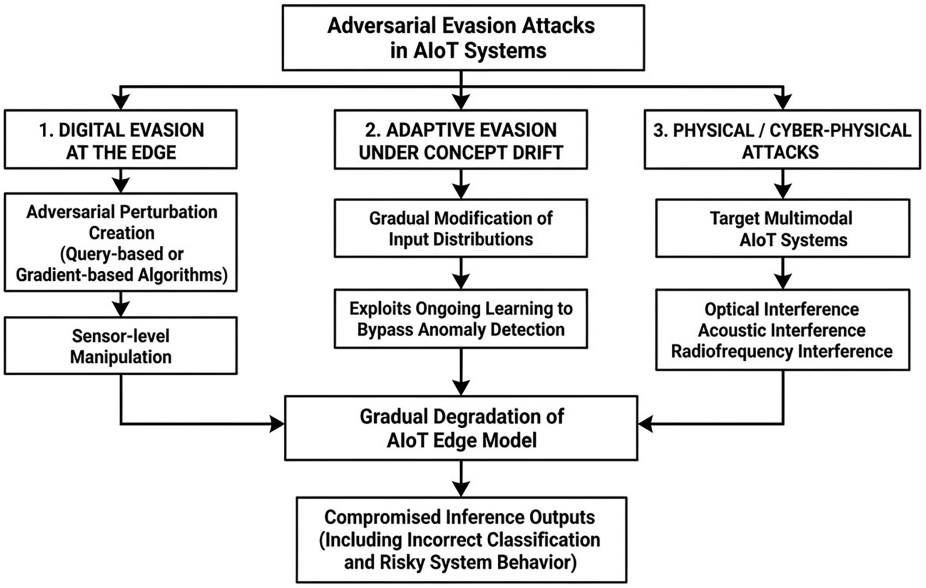

Figure 7: Adversarial Evasion attack flow in the AIoT systems.

4.2 Poisoning and Backdoor Attacks

4.2.1 Data Poisoning in the AIoT Pipelines

Attacks known as "data poisoning" alter training datasets to cause targeted misbehavior or model degradation [59]. Since training data is frequently continuously gathered from dispersed sensors in uncontrolled environments, the AIoT systems are especially vulnerable.

The existence of spatiotemporal correlation bias, which allows minor physical environment perturbations (such as temperature changes, lighting changes, or localized interference) to consistently bias long-term datasets without being detected, is a major problem in AIoT poisoning [60].

In AIoT pipelines, clean label poisoning has grown in significance, surpassing conventional label flipping threats [61]. In this scenario, positioned samples have feature space alterations that slightly distort decision boundaries while keeping accurate labelling. This is especially risky in automated processes where validation mainly verifies label integrity instead of feature distribution consistency. In online learning AIoT systems, data stream poisoning is a current danger in which attackers progressively alter model behavior in continuous learning deployments by injecting malicious samples over time [62].

4.2.2 Backdoor Attacks via Supply Chain Channels and OTA (Over-the-Air)

Hidden triggers are used in backdoor attack models, like specific input can cause malicious behavior while preserving regular accuracy [63]. In the AIoT ecosystem, the OTA update pipeline that is often used for model development across dispersed edge devices is the most vital vector.

In the deployed devices, Trojan models can be injected by unsafe firmware signing processes, compromised updated servers, or adversarial intermediates. Most of the post-deployment detection is challenging because several IoT devices are not able to perform complete cryptographic verification or weight-level auditing of model integrity due to limited computational ability [64].

In addition to OTA attacks, supply chain model poisoning [65], in which affected models acquired from a third-party source already cover backdoors, has become an important danger. Since these models are frequently adjusted on-device, hidden triggers may inadvertently be preserved or amplified. Detection becomes more difficult in a resource-constrained AIoT situation because of the latent backdoor persistence under compression and pruning, where even intensive model optimization is not able to eliminate the embedded trigger [66].

Fig. 8 shows the general process of backdoor and poisoning attacks in the AIoT systems. On the left, clean-label poisoning, long-term data bias, and feature-space distortions are introduced by carefully manipulating data gathered from dispersed sensors during the training phase. Model learning is degraded as a result of these contaminated inputs spreading throughout the training pipeline. On the right, backdoor attacks use OTA updates or third-party supply chain sources to inject Trojan models during the deployment process. These models have hidden triggers that, in some situations, cause harmful behavior while otherwise operating normally. In the AIoT systems, both attack vectors eventually converge at the deployed AI model, leading to targeted misclassification, corrupted model behavior, and concealed malicious functionality.

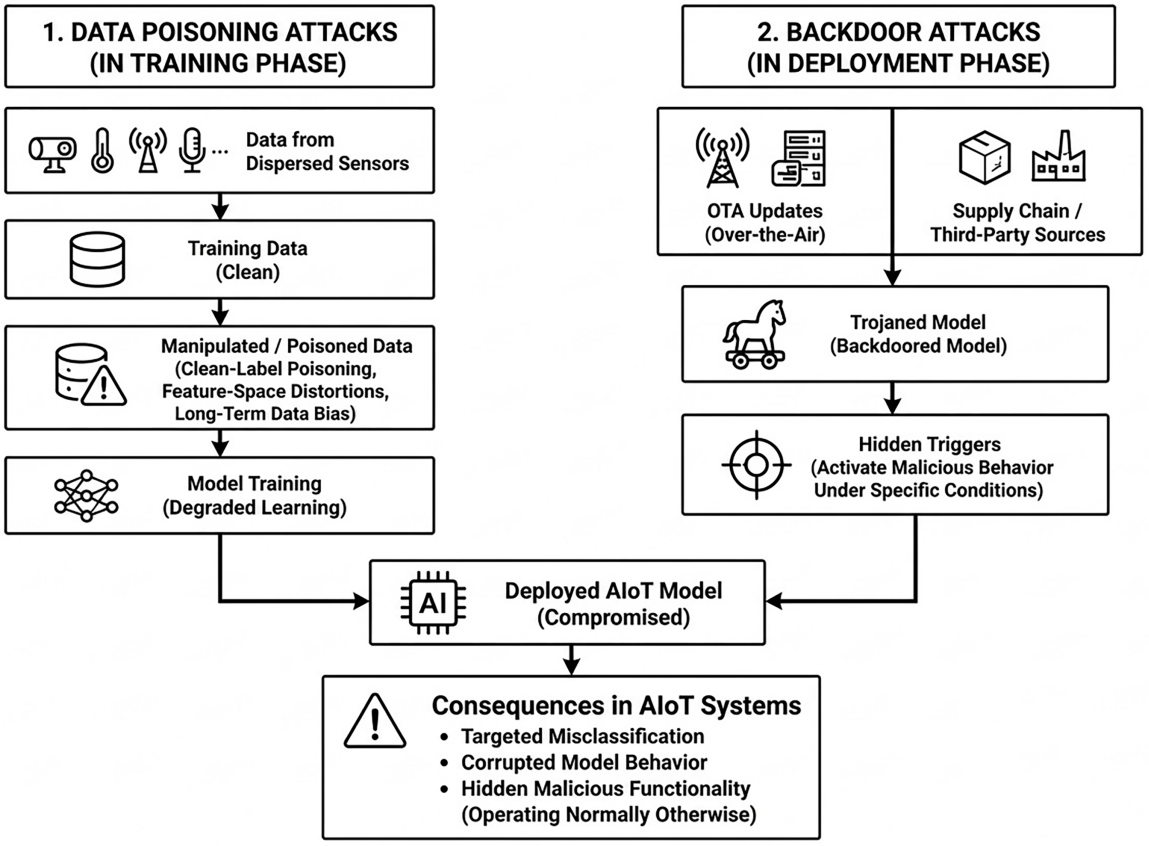

Figure 8: Poisoning and Backdoor attack pipeline in AIoT systems.

4.3 Byzantine and Federated Learning Attacks

Federated learning (FL) protects data privacy while enabling decentralized training among AIoT devices [67]. However, Byzantine participants who can manipulate model updates are exposed to serious vulnerabilities due to its distributed nature.

Gradient poisoning attacks and model replacement attacks, in which malicious clients submit properly constructed updates to completely dominate or skew the global model, continue to be an ultimate danger [59]. Interestingly, previous research has demonstrated that persistent backdoors can be effectively installed under conventional FedAvg (federated averaging) aggregation by a single malicious client [68]. Amplification considerations unique to AIoT include:

- Since honest updates inherently display significant variance, non-IID (Independent and Identically Distributed) data distributions make anomaly detection challenging [69].
- The influence of individual malicious nodes is increased by sparse participation per training cycle [70].
- Edge device intermittent connectivity weakens consensus-based defences [71].

Sign-flipping and scaled gradient attacks under adaptive aggregation defences [72], in which adversaries dynamically modify their updates based on observable server-side filtering methods, represent a recent development in attack design. The attack surface in large-scale AIoT deployments is further increased by sybil-based FL attacks, in which a single adversary controls numerous virtual or compromised edge nodes [73].

4.4 Model Extraction and Privacy Leakage Attacks

Model extraction attacks use deployed AI systems as black-box oracles to reconstruct functionally similar surrogate models [74]. Multi-channel leakage vectors that go beyond basic input-output enquiries greatly increase the impact of these attacks in AIoT.

The number of queries needed for high-fidelity extraction can be significantly decreased by using side-channel data from edge inference hardware, such as power usage, timing fluctuations, cache behavior, and electromagnetic emissions [75]. After being extracted, surrogate models can be applied to:

- Increase in adversarial attack transferability,
- Theft of intellectual property,
- System reverse engineering.

AIoT privacy threats go beyond model extraction. Membership inference attacks put healthcare IoT and smart home environments at risk by enabling adversaries to ascertain if a particular sample is included in the training dataset [76]. In FL, gradient inversion attacks pose a more serious risk since they allow adversaries to directly recreate raw sensor inputs (such as voice, physiological data, or pictures) from shared gradients [77]. This essentially calls into question the fundamental privacy premise of FL-based AIoT systems.

Temporal privacy leakage is an emerging field where user behavior can be reconstructed even in the absence of direct model access, thanks to long-term inference patterns in continuous AIoT data streams [78].

Table 6 lists the main attack types in AIoT systems, together with their attack vectors, target layers, and salient features, so as to give an organized summary of the numerous adversarial threats covered in this section. Particular obstacles presented by AIoT environments are highlighted in this tabular depiction, especially the integration of the physical and cyber domains, resource-constrained edge devices, and distributed learning frameworks.

Table 6: Overall summary of adversarial attacks in the AIoT ecosystem.

Attack Category	Sub-Type	Attack Vector	Target Layer	Key Characteristics in AIoT
Evasion Attacks	Digital Evasion	PGD, C&W, quantization-aware attacks.	Edge Inference	Under resource-constrained compress models, exploit model risks.
	Physical Attacks	LiDAR spoofing, adversarial patches, RF interference, and audio injection.	Sensor/Physical Layer	Impact physical world perception systems; robust to environmental variations.
	Cross-modal Attacks	Multi-sensor perturbations.	Sensor Fusion Layer	Exploit dependencies between modalities in AIoT systems.
Poisoning Attacks	Data Poisoning	Clean-label, feature manipulation.	Training Pipeline	Hard to detect due to correct labelling and subtle feature shifts.
	Stream Poisoning	Injection of online data, concept drift.	Continuous Learning Systems	Steady degradation of model performance over time.
Backdoor Attacks	OTA Backdoors	Malicious updates of models.	Deployment Layer	Exploit insecure firmware/model update mechanisms.
	Supply Chain Attacks	Compromised pertained models.	Model Acquisition	Even after compression or fine-tuning, hidden triggers persist.
Federated Learning Attacks	Byzantine Attacks	Malicious updates of the gradient.	FL Aggregation	With a few compromised devices, manipulate the global model.
	Model Replacement	Full model override.	FL Server	A single attacker can dominate the aggregation process.
	Sybil Attacks	Multiple fake clients.	Distributed FL System	Amplify adversarial influence in decentralized settings.
Model & Privacy Attacks	Model Extraction	Query-based + side-channel leakage.	Inference Interface	Enables surrogate model creation and IP theft.
	Membership Inference	Training data identification.	Model Output	Privacy leakage in sensitive AIoT applications.
	Gradient Inversion	Data reconstruction from gradients.	FL Training	Severe privacy breach in distributed learning.
	Side-channel Attacks	Power, timing, EM leakage.	Hardware Layer	Exploit physical leakage in edge devices.
	Temporal Leakage	Behavioral pattern reconstruction.	System Level	Long-term user activity inference.

5 Adversarial Defences in AIoT: Analysis

Strong and flexible defence mechanisms that can function in the face of resource limitations and changing environmental conditions are essential for countering adversarial attacks in AIoT systems [79]. AIoT implementations, in contrast to traditional AI systems, have to deal with a wider risk surface that includes distributed learning frameworks, edge devices, and physical sensors. Therefore, a comprehensive, multi-layered approach that incorporates algorithmic resilience, system-level protection, and hardware-assisted security is necessary for effective defence measures.

5.1 Unified Taxonomy of Defence Mechanisms

In AIoT systems, adversarial defences must function under severe limitations, such as constrained processing power, constant data flows, and exposure to both physical and cyber-attack surfaces [79]. AIoT defences, in contrast to traditional cloud-based AI, must simultaneously handle hardware-level trust, distributed learning security, and real-time inference robustness. Defence mechanisms can be methodically divided into five broad families to capture these requirements:

- Input transformation and filtering methods, which clean inputs before inference;
- Robustness-through-training strategies, which improve intrinsic model resilience throughout training;
- Hardware- and system-level safeguards that guarantee the integrity of models and execution environments;
- Federated and distributed defences, especially for decentralized AIoT learning frameworks;
- Detection and monitoring systems that spot abnormal or hostile behavior at runtime.

These defence families are summarized in Table 7 based on their threat coverage, adversarial robustness, computational overhead, memory requirements, and deployment feasibility within resource-constrained AIoT environments. The qualitative assessments (e.g., Low, Medium, High) are derived through comparative analysis of the characteristics, performance trade-offs, scalability, and implementation constraints reported in the referenced studies to provide a consistent evaluation framework for AIoT-oriented adversarial defence mechanisms.

Table 7: Comparative analysis of adversarial defence mechanisms in AIoT.

Defense Mechanism	Primary Threat Addressed	Robustness	Compute Cost	Memory Cost	Key Limitations/AIoT Challenges
Adversarial Training (PGD-AT) [80]	Digital Evasion	High	Very High	Low	Expensive; limited transfer to physical attacks
TRADES/Robust Loss [81]	Digital Evasion	High	Medium	Low	Trade-off tuning required; still digital-domain focused
Fast Adversarial Training (FGSM-RS) [82]	Evasion	Medium-High	Low-Medium	Low	Slight robustness drop vs. PGD-AT
Randomized Smoothing [83]	Evasion (L2 Certified)	Medium	Very High	Medium	Requires many forward passes; impractical on the edge

(Continued)

Table 7 (continued)

Defense Mechanism	Primary Threat Addressed	Robustness	Compute Cost	Memory Cost	Key Limitations/AIoT Challenges
Interval Bound Propagation (IBP) [84]	Certified Robustness	Medium	High	Medium	Limited scalability for complex models
Input Preprocessing (JPEG, smoothing) [85]	Evasion (Digital/Physical)	Low-Medium	Low	Low	Easily bypassed by adaptive attacks
Feature Denoising/Autoencoders [86]	Evasion + Noise	Medium	Medium	Medium	May distort clean data; domain-specific tuning needed
Lightweight Anomaly Detection (OCSVM, TinyML IDS) [87]	Evasion + DoS	Medium	Low	Low	Limited generalization to unseen attacks
Federated Adversarial Training [88]	FL + Evasion	Medium	High	Medium	Communication overhead; non-IID instability
Robust Aggregation (Krum, FLTrust, FLARE) [89]	Byzantine Attacks	High	Low	Low	Performance degrades under a high attacker ratio
Secure Aggregation + DP [90]	Privacy Leakage	Medium	Medium	Medium	Reduces model utility; adds communication cost
Hardware Trusted Execution Environment (TEE) (TrustZone, Software Guard Extension (SGX)) [91]	Model Extraction, Backdoor	High	Medium	High	Not available on ultra-low-end devices
PUF-based Attestation [92]	Integrity/Cloning	High	Low	Low	Limited to identity verification (not full defense)
Model Watermarking [93]	IP Theft/Extraction	High (IP)	Low	Low	Post-hoc protection only
Runtime Monitoring (Edge IDS + telemetry) [94]	Multi-Vector Attacks	Medium-High	Low-Medium	Low	Requires system-level integration

Fig. 9 demonstrates the multi-layered nature of adversarial defence techniques in AIoT systems by mapping defensive mechanisms across several system layers. Hardware-rooted security, at its most basic

level, guarantees device authenticity and builds confidence. System-level protections that ensure model execution and storage come next. By using adversarial training and input manipulations, the algorithmic layer improves the robustness of the model. The top layer concentrates on real-time monitoring and anomaly detection, whereas federated and distributed defences address weaknesses related to collaborative learning. The illustration also shows how various assault types might target several layers at once, highlighting the necessity of a thorough defence plan.

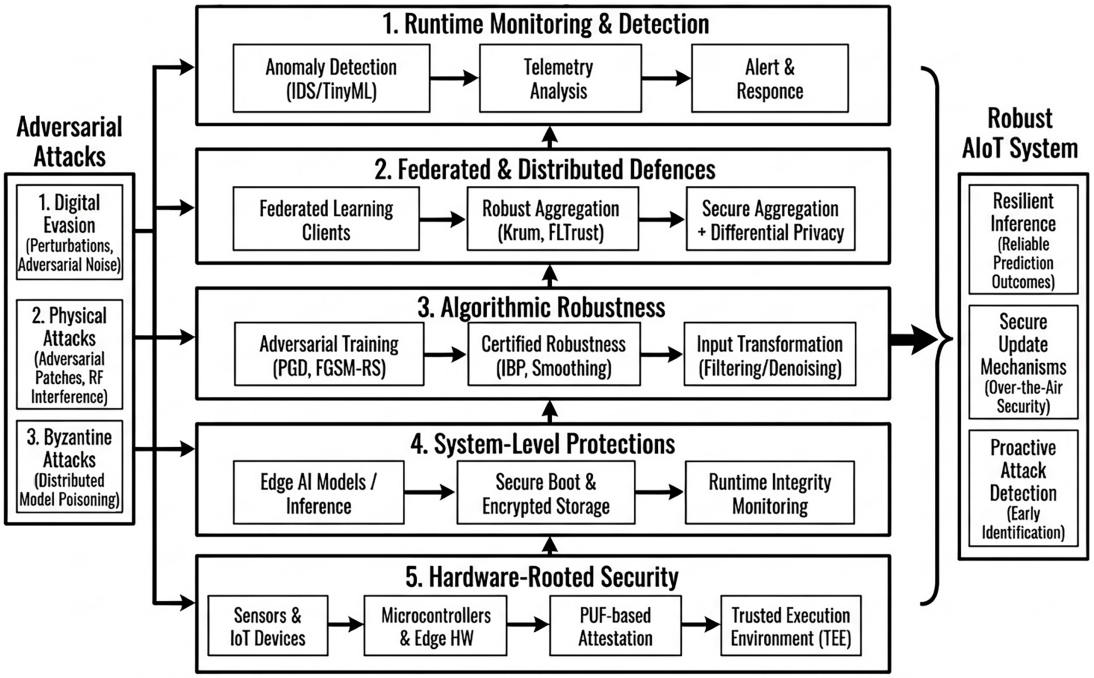

Figure 9: Block diagram of AIoT systems' multi-layered adversarial defense mechanisms.

5.2 Adversarial Training under AIoT Constraints

The most effective empirical defence against evasion attacks is still adversarial training (AT), which incorporates adversarial instances during model optimization [95]. Although PGD-based AT offers excellent robustness guarantees, many AIoT workflows cannot use it because of its computational complexity, especially in edge-based or on-device training scenarios.

Efficiency-aware adversarial training has been the subject of recent developments, including cyclic learning rate strategies and single-step techniques like (Fast Gradient Sign Method) FGSM with random beginnings (FGSM-RS) [68]. These methods are appropriate for TinyML pipelines with limited resources since they drastically lower training overhead while preserving acceptable robustness levels. The discrepancy between digital and physical threat models is a significant AIoT drawback. While real-world AIoT systems are subject to physically realizable attacks (such as patches, acoustic waves, and RF interference), standard AT mostly handles norm-bounded disturbances [96].

By integrating environmental changes during training, new methods like sensor-aware augmentation and physically grounded adversarial training seek to close this gap. Continuous adversarial training, in which models adjust to changing attack patterns in streaming AIoT environments without catastrophic forgetting, is a new approach.

5.3 Certified and Probabilistic Robustness

The goal of certified defences is to offer explicit assurances regarding the behavior of the model under bounded perturbations [97]. A popular probabilistic technique, randomized smoothing aggregates predictions over noisy inputs to provide robustness guarantees over an L2 radius. However, because certification requires hundreds to thousands of forward passes, significant inference latency limits its use in AIoT. It is therefore inappropriate for real-time edge inference.

Although they are currently restricted to very small architectures, deterministic certification techniques like Interval Bound Propagation (IBP), CROWN, and DeepPoly provide tighter assurances with fewer evaluations [98]. Scalability to more complicated AIoT workloads is still a barrier, notwithstanding the potential for TinyML models.

To establish a useful trade-off between robustness and efficiency in edge deployments, recent work investigates hybrid certification, which combines empirical defences with lightweight bounds [99].

5.4 Federated and Distributed Defence Mechanisms

Federated learning's decentralized structure creates special weaknesses [100]. By filtering harmful updates using geometric or similarity-based criteria, robust aggregation solutions like Krum, Multi-Krum, and FLTrust reduce the impact of Byzantine attacks [101]. The efficacy of traditional robust aggregation is diminished by AIoT-specific issues such as non-IID data distributions, low client engagement, and sporadic connectivity. Recent expansions consist of:

- Adaptive aggregation procedures that dynamically modify clients' trust ratings [102].
- Systems of client reputation based on past performance.
- Coordinated assault detection using cluster-based filtering (FLARE, Flame).
- Sybil-resilient defences that use graph-based analysis and identity verification.

Furthermore, gradient leakage is prevented using secure aggregation in conjunction with differential privacy (DP), although at the expense of decreased model accuracy [103].

5.5 Hardware-Rooted and System-Level Defences

In AIoT systems, hardware-based defences offer a vital trust anchor, especially against firmware-level attacks, model extraction, and tampering [104]. Model inference can be securely isolated thanks to Trusted Execution Environments (TEEs) like RISC-V Keystone, Intel SGX, and ARM TrustZone [105]. Since ARM TrustZone is compatible with Cortex-M and Cortex-A processors, it is most commonly used in AIoT [106]. Secure inference is possible even on limited systems with moderate overhead, as shown by lightweight implementations (e.g., TinyTEE, Sherloc).

For ultra-low-cost microcontrollers, when TEEs are not available, there is a sizable gap [107]. Physical Unclonable Functions (PUFs) offer an alternative in these situations by facilitating safe key creation and hardware-rooted authentication. Cross-layer defence integration is a contemporary development that combines:

- System-level anomaly detection,
- Encrypted model storage,
- Secure boot and attestation,
- Runtime integrity monitoring.

These all-encompassing strategies work especially well in AIoT, because attackers frequently target several layers at once.

Table 8 provides a comparative overview of the main adversarial defence methods in AIoT systems to facilitate understanding of the different defence strategies discussed in this section. The table highlights their target threats, operational layers, deployment suitability within resource-constrained and distributed AIoT environments, and associated implementation limitations. The qualitative assessments are derived through comparative analysis of the characteristics, computational requirements, scalability, and deployment constraints reported in the literature to ensure consistency and reproducibility of the evaluation framework.

Table 8: Summary of main Adversarial methods in AIoT.

Defense Family	Technique	Target Threats	Operational Layer	AIoT Suitability	Key Limitations
Robust Training	PGD-based Adversarial Training	Digital evasion	Model Training	Moderate	High computational cost; weak against physical attacks
	TRADES/Robust Loss	Evasion (digital)	Model Training	Moderate-High	Requires careful tuning of the accuracy-robustness trade-off
	Fast AT (FGSM-RS)	Evasion	Model Training	High	Slightly lower robustness than full AT
	Federated Adversarial Training	Evasion + Byzantine	Distributed Training	Moderate	Communication overhead; non-IID issues
Input Transformation	JPEG Compression/Smoothing	Digital & Physical Evasion	Input Preprocessing	High	Vulnerable to adaptive attacks
	Feature Denoising/Autoencoders	Noise & perturbations	Input/Feature Layer	Moderate	May degrade clean input quality
Detection-Based	Lightweight Anomaly Detection (OCSVM, TinyML IDS)	Evasion, DoS	Edge Runtime	High	Limited generalization to unseen attacks
	Runtime Monitoring & Telemetry	Multi-vector attacks	System Level	High	Requires integration across system layers
Federated Defenses	Krum/Multi-Krum	Byzantine poisoning	FL Aggregation	High	Less effective against many attackers
	FLTrust/FLARE	Gradient manipulation	FL Server	High	Needs a trusted reference or clustering overhead
	Sybil-Resilient Mechanisms	Sybil attacks	Distributed FL	Moderate	Identity management complexity
Certified Robustness	Randomized Smoothing	Evasion (L2)	Inference	Low	High latency (multiple forward passes)
	IBP/CROWN	Certified robustness	Training/Inference	Low-Moderate	Limited scalability for large models
Hardware & System-Level	Trusted Execution Environments (TEE)	Model extraction, tampering	Hardware Layer	Moderate	Not available on low-end devices
	PUF-based Attestation	Device/model integrity	Hardware Layer	High	Limited to authentication, not full defence
	Secure Boot & Encryption	Model integrity	System Layer	High	Added implementation complexity

(Continued)

Table 8 (continued)

Defense Family	Technique	Target Threats	Operational Layer	AIoT Suitability	Key Limitations
Privacy Preservation	Differential Privacy (DP)	Data leakage	Training (FL)	Moderate	Reduces model accuracy
	Secure Aggregation	Gradient leakage	FL Communication	High	Communication overhead
	Model Watermarking	IP theft/extraction	Model Level	High	Only post-attack attribution

6 Emergent Cross-Layer Adversarial Threats

The AIoT systems' growing integration of sensing, processing, and actuation creates a new class of adversarial threats that go beyond isolated attack surfaces. These risks arise from the way system layers interact, giving adversaries the ability to take advantage of dependencies across network, computational, and physical components [108]. Localized perturbations can therefore spread and intensify across the system, producing complicated and frequently unanticipated results. Designing strong and resilient AIoT architectures thus requires an understanding of these cross-layer interactions. The emergent cross-layer adversarial threats in AIoT systems are depicted in Fig. 10, which illustrates how attacks might spread throughout several closely related levels of an integrated architecture.

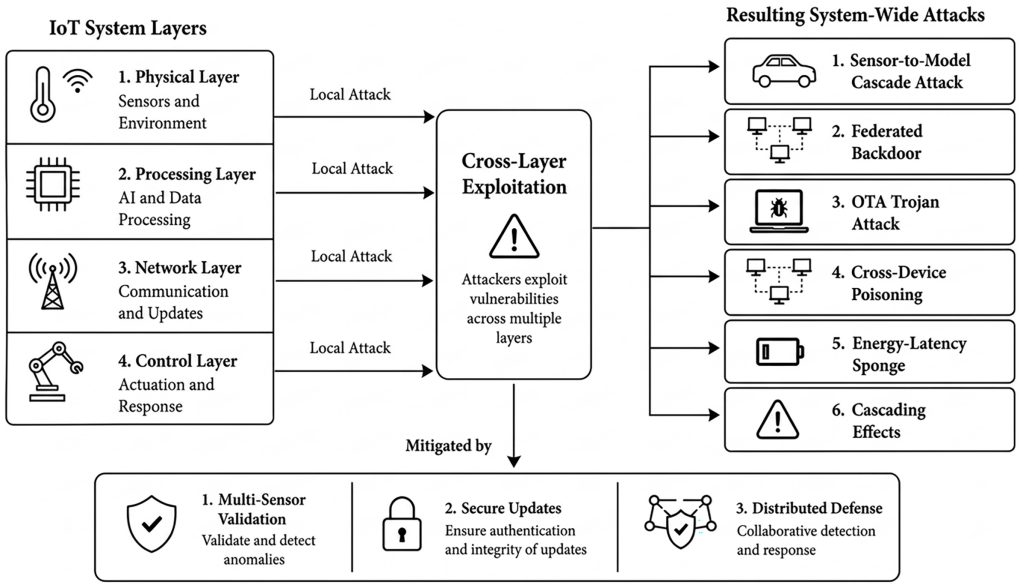

Figure 10: Emergent cross-layer adversarial threats in AIoT systems.

A local attack can start at the physical or sensing level and spread through higher layers in the AIoT system, which is depicted on the left as a four-layer stack made up of the Physical Layer (sensing and environmental interaction), Processing Layer (AI computation and data analysis), Network Layer (communication and model updates), and Control Layer (actuation and system response). The idea of cross-layer exploitation is emphasized in the figure's center, showing how minor disruptions can spread through linked layers and intensify into system-wide failures. Five major cross-layer adversarial threat classes are summarized on the right side of the diagram: Sensor-to-Model Cascade Attacks, Federated Backdoor via

Sensor Homogeneity, OTA Trojan or Mode Switch Attacks, Cross-Device Poisoning Amplification, and Energy-Latency Sponge Attacks. Mitigation techniques, including dispersed defence systems, secure OTA update mechanisms, and multi-sensor validation, are shown at the bottom. Overall, the image highlights the cross-layer nature of AIoT security vulnerabilities, which can spread in a cascade fashion across the sensing, learning, communication, and control domains.

6.1 Defining Cross-Layer Threats

The close integration of physical sensing, embedded intelligence, network communication, and real-time actuation is a distinguishing feature of AIoT systems [109]. Adversarial behaviors that go beyond the conventional bounds of either AML or standard IoT security are produced by this deep coupling. Attacks, for example, can spread throughout several system layers and increase their impact by taking advantage of the interdependencies between perception, learning, and control.

These are what we refer to as cross-layer adversarial threats, which are attack patterns that concurrently take advantage of interactions across system-level, algorithmic, and physical components [38]. These threats, in contrast to single-layer attacks, frequently show cascading effects, where a small disturbance at one layer has exaggerated repercussions at another.

Our analysis highlights developing vulnerabilities specific to closely linked intelligent systems and identifies five representative cross-layer threat classes, as presented in Table 9, which are mostly missing from current AIoT and AML taxonomies.

Table 9: Cross-Layer adversarial threat classes in the AIoT.

Cross-Layer Threat Class	Mechanism Description	AIoT Application Scenarios	Mitigating Approach
Sensor-to-Model Cascade Attack	Physical perception manipulation cascading to AI-driven control miss action	Smart vehicle obstacle avoidance; industrial robot safety stops	Physical barrier + model-level anomaly detection
Federated Backdoor via Sensor Homogeneity	Exploitation of correlated sensor data distribution in FL to implant backdoors surviving aggregation	Smart grid demand prediction; fleet health monitoring	Data heterogeneity injection; per-layer gradient analysis
OTA-Triggered Adversarial Mode Switch	Trojaned OTA update activates dormant adversarial vulnerability in previously benign model	Smart home device reclassification; medical wearable alert suppression	Cryptographic model attestation; behavioral regression testing post-update
Cross-Device Poisoning Amplification	Coordinated poisoning across multiple devices sharing a common AIoT platform exploits aggregation to amplify attack influence	Smart city camera networks; agricultural sensor arrays	Decentralized trust scoring; contribution auditing
Energy-Latency Sponge Cascade	Sponge adversarial inputs designed to maximize inference latency cascade to real-time control failures (DoS via model)	Autonomous systems; medical monitoring	Inference time budgeting; input complexity gating

6.2 Sensor-to-Model Cascade Attacks

The closed-loop nature of AIoT systems, where perception outputs directly affect control decisions, gives rise to sensor-to-model cascade attacks. Mispredictions in AIoT systems, in contrast to traditional classification errors, might result in dangerous or unintentional physical actions [110]. For instance, a well-constructed physical disturbance (such as an adversarial patch or signal interference) may cause a perception model to incorrectly identify a crucial object, which then spreads to the control layer and causes dangerous actuation. The main problem is that the control loop amplifies the attack's impact, turning minor prediction errors into dangerous system behaviors.

Cross-layer co-design is necessary to mitigate these risks, where robustness is assessed both at the model level and in relation to downstream control strategies. To avoid single-point failures, emerging strategies include redundant multi-sensor validation systems and risk-aware control algorithms.

6.3 Federated Backdoor via Sensor Homogeneity

Conventional Byzantine-resilient FL makes the assumption that benign clients have enough multiple information distributions to allow anomalous updates to be detected [111]. But in AIoT contexts, where devices frequently function under very similar sensing conditions, this assumption fails.

Gradient updates from trustworthy devices naturally correlate in certain situations, making it possible for a malicious client to create backdoor updates that match this distribution and avoid detection. Physical deployment homogeneity (sensing layer) and statistical aggregation processes (learning layer) intersect to provide a unique cross-layer vulnerability [112].

According to recent research, rather than depending only on global similarity measurements, it is necessary to perform fine-grained gradient consistency checks across layers and incorporate artificial variety into training (e.g., data augmentation, client clustering) in order to mitigate this issue.

6.4 OTA-Triggered Adversarial Mode Switch

A crucial cross-layer attack vector that connects cloud-level model distribution and edge-level execution is introduced via OTA updates [113]. A Trojan model that operates properly under normal circumstances but initiates harmful behavior upon receiving a specific trigger can be deployed by an adversary who compromises the update pipeline.

These triggers in AIoT systems can involve physical environmental factors like temperature thresholds, audio signals, or temporal events in addition to digital patterns [32]. As such, triggers may be missed by conventional input-based detection systems, which greatly boosts stealth. Strengthening end-to-end update integrity, which includes secure boot standards, cryptographic attestation, and post-deployment behavioral verification under various environmental circumstances, is necessary to counter this danger. Continuous runtime attestation, in which model behavior is routinely verified against reliable reference profiles, is a promising approach.

6.5 Cross-Device and Resource-Aware Cascading Attacks

Beyond the identified categories, emerging AIoT threats increasingly exploit system-wide coordination and resource constraints [114]. When compared to standalone attacks, cross-device poisoning amplification shows how several compromised devices can work together to affect distributed learning systems, greatly enhancing attack efficacy.

Similar to this, energy-latency sponge attacks create inputs that maximize inference complexity to target edge devices' computing limits [115]. This causes delayed responses in real-time AIoT applications,

thereby generating a model-induced DoS condition that impairs system functionality without the need for conventional network flooding.

These patterns show a more general trend toward resource-aware adversarial methods, in which attackers take advantage of system-level limitations like latency, energy, and bandwidth in addition to model flaws.

7 Discussion

This section incorporates the survey's main findings by analyzing them in light of the research questions. It concentrates on the fundamental behavior, system consequences, and structural features of adversarial AIoT systems rather than simply restating attack or defence taxonomies.

7.1 RQ1: The AIoT Adversarial Threats' Structural Nature

The analysis shows that rather than being separate algorithmic flaws, adversarial threats in AIoT are better understood as system-interaction phenomena. AIoT creates new connections between perception, communication, and actuation in contrast to traditional AI security settings, where attacks mainly target model inputs or training data.

A key insight is that the attack surface is defined not only by the model but by the interaction between physical signals and computational decision pathways. This leads to emergent behaviors that are not present in conventional AI pipelines, like cascade misinterpretations and environment-driven model alteration.

7.2 RQ2: Defence Efficiency under Deployment Limitations

Deployment reality, in particular, resource constraints and system heterogeneity, significantly influence defence performance in AIoT. The results point to a fragmented defence environment, where each mechanism offers some protection against particular threat classes, rather than a single dominant defence technique. Because of this fragmentation, resilience in AIoT is implied to be compositional rather than absolute, necessitating integrated defence measures across system, inference, and training levels.

The analysis further reveals that the practicality of adversarial defence mechanisms in AIoT strongly depends on deployment constraints such as memory availability, computational capability, latency sensitivity, communication overhead, and energy consumption. Lightweight mechanisms, including sensor filtering, input sanitization, lightweight anomaly detection, and quantization-aware robustness, are comparatively feasible for constrained edge devices due to their reduced computational and memory requirements. In contrast, computationally intensive approaches such as certified robustness verification, large-scale adversarial retraining, federated aggregation auditing, and ensemble-based runtime analysis typically require edge-server or cloud-assisted infrastructures. These observations indicate that defence suitability in AIoT is deployment-dependent, where security effectiveness must be balanced against operational efficiency and real-time system requirements.

7.3 RQ3: Dynamics of Cross-Layer Security Emerging

The finding of cross-layer adversarial dynamics, in which vulnerabilities result from interactions across system components rather than within individual layers, is a significant contribution of this study. Three significant characteristics are revealed by these dynamics:

- Propagation: Even tiny disturbances can spread through layers and intensify at the control level.
- Coupling: Dependencies between the sensing, learning, and communication modules lead to security flaws.

- Concealment: Adversarial behavior can be concealed from conventional detection methods through multi-layer interactions.

As a result, end-to-end system resilience modelling replaces component-level robustness in the security paradigm.

8 Conclusion

This study presented a systematic review of adversarial threats and defence mechanisms in AIoT systems, with particular emphasis on the transition from isolated machine learning vulnerabilities to complex cross-layer security challenges emerging within tightly integrated intelligent environments. The analysis demonstrates that conventional AI security or IoT security frameworks alone are insufficient for comprehensively characterizing adversarial behaviours in AIoT ecosystems. Instead, adversarial threats in AIoT arise through complex interactions among sensing, communication, learning, and actuation components operating across heterogeneous system layers. To address this challenge, the proposed 3D-AIoT-TT framework is introduced as a unified taxonomy for systematically modeling adversarial threats across three fundamental dimensions: AI pipeline stages, IoT architectural layers, and adversarial knowledge levels. The framework further enables structured characterization of compound and cross-layer attack scenarios that are typically overlooked in traditional single-layer threat models.

By synthesizing existing adversarial attacks, defence strategies, and emergent cross-layer vulnerabilities, this work highlights the growing need for comprehensive, deployment-aware, and resource-efficient security mechanisms suitable for real-world AIoT environments. The findings further indicate that future progress in AIoT security will depend on bridging the gap between theoretical adversarial robustness and practical system-level resilience under heterogeneous deployment constraints. Overall, this study establishes a structured foundation for next-generation adversarial AIoT security research, where robustness must be treated as an integrated system-level property rather than an isolated model-centric characteristic. Future research directions include the development of scalable cross-layer defence frameworks, lightweight resource-aware security mechanisms, and standardized real-world evaluation methodologies for intelligent AIoT infrastructures.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm their contributions to the paper as follows: literature review and analysis: Ali Hassan, Syed Rizwan Hassan; study conception and design: Ali Hassan, Syed Rizwan Hassan, Ammar Rafiq; data collection: Ali Hassan, Ammar Rafiq; critical analysis and interpretation of results: Syed Rizwan Hassan; draft manuscript preparation: Ali Hassan, Syed Rizwan Hassan, Ammar Rafiq. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: This study did not involve any human or animal participants.

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

Supplementary Materials: The supplementary material is available online at https://www.techscience.com/doi/10.32604/cmc.2026.084672/s1. PRISMA abstract checklist and PRISMA 2020 checklist are available in the supplementary materials.

Abbreviations

Term	Interpretation
AIoT	Artificial Intelligence of Things
3D-AIoT-TT	Three-Dimensional AIoT Threat Taxonomy
MCUs	Microcontroollers
RL	Reinforcement Learning
XAI	Explainable Artificial Intelligence
ToM	Theory of Mind
DT	Digital Twin
EM	Electromagnetic
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RFID	Radio Frequency Identification
FGSM	Fast Gradient Sign Method
DoS	Denial-of-Service
BBox	Black Box
C&W	Carlini & Wagner
LiDAR	Light Detection and Ranging
FedAvg	Federated Averaging
FDT	Fast Adversarial Training
OCSVM	One Class Support Vector Machine
SGX	Software Guard Extension
DP	Differential Privacy
AML	Adversarial Machine Learning
IoT	Internet of Things
NAS	Neural Architecture Search
SMGA	Sense-Map-Generate-Act
FL	Federate Learning
GAI	Generative Artificial Intelligence
TinyML	Tiny Machine Learning
NIST	National Institute of Standards and Technology's
IMUs	Inertial Measurement Units
PGD	Projected Gradient Descent
APIs	Application Programming Interfaces
OTA	Over-the-Air
MITM	Man-in-the-Middle
DMA	Direct Memory Access
RF	Radio Frequency
IID	Independent and Identically Distributed
IBP	Interval Bound Propagation
TEE	Trusted Execution Environment
PUFs	Physical Unclonable Functions
IDS	Intrusion Detection System

Appendix A

Representative database-specific search queries used during literature retrieval.

Database	Example Query Formulation
IEEE Xplore	(“Adversarial Machine Learning” OR “Adversarial Attack” OR “Poisoning Attack” OR “Evasion Attack”) AND (“AIoT” OR “Artificial Intelligence of Things” OR “Internet of Things”) AND (“Security” OR “Defense” OR “Robustness”)
Scopus	TITLE-ABS-KEY ((“Adversarial Machine Learning” OR “Adversarial AI” OR “Poisoning Attack” OR “Evasion Attack”) AND (“AIoT” OR “Artificial Intelligence of Things” OR “Edge AI”) AND (“Security” OR “Trustworthy AI” OR “Attack Detection”))
Web of Science	TS = ((“Adversarial Machine Learning” OR “Adversarial Attack”) AND (“AIoT” OR “Artificial Intelligence of Things”) AND (“Security” OR “Defense” OR “Robustness”))
ACM Digital Library	Abstract: (“AIoT” AND “Adversarial Machine Learning” AND Security)
ScienceDirect	(“Artificial Intelligence of Things” OR “AIoT”) AND (“Adversarial Attack” OR “Poisoning Attack”) AND Security
SpringerLink	(“AIoT” AND “Adversarial Machine Learning”) OR (“AI-enabled IoT” AND Security)
Wiley Online Library	(“AIoT” OR “Edge AI”) AND (“Adversarial Learning” OR “Attack Detection”)
MDPI	AIoT AND Adversarial Attack AND Defense Mechanisms
Google Scholar	All in title: AIoT Adversarial Machine Learning Security

Appendix B

Quality assessment rubric used for study selection.

Assessment Criterion	Description	Score Range
Relevance	Relevance to AIoT adversarial security domain	0–2
Novelty	Originality and contribution significance	0–2
Methodological Rigor	Soundness of research methodology	0–2
Experimental Validation	Quality of experiments/results/evaluation	0–2
Reporting Clarity	Clarity of presentation and technical explanation	0–2
Total Score	Maximum achievable score	**10**

References

1. Hassan A. Quantum-IoT security: advances, challenges, and future directions. J Hardw Syst Secur. 2026;10(1):6. doi:10.1007/s41635-026-00179-z.

2. Hassan A, Nizam-Uddin N, Quddus A, Hassan SR, Rehman AU, Bharany S. Navigating IoT security: insights into architecture, key security features, attacks, current challenges and AI-driven solutions shaping the future of connectivity. CMC. 2024;81(3):3499–559. doi:10.32604/cmc.2024.057877.
3. Kavitha Devi MK, Murugan T, Indira K, Lavanya R, Abinaya S. Intelligent mobile and IoT ecosystems: bridging cloud, fog, edge, and AI. Boca Raton, FL, USA: CRC Press; 2026, doi:10.1201/9781003608653.
4. Herglotz C, Jabłoński I, Karnapke R, Shahin K, Reichenbach M. Designing intelligent sensor networks: a comprehensive survey of sensing, edge AI, and 5G+/6G connectivity. IEEE Access. 2025;13(6):172306–25. doi:10.1109/ACCESS.2025.3615205.
5. Dhilleswararao P, Boppu S, Manikandan MS, Cenkeramaddi LR. Efficient hardware architectures for accelerating deep neural networks: survey. IEEE Access. 2022;10:131788–828. doi:10.1109/ACCESS.2022.3229767.
6. Ali Khan S, Mazhar T, Shah SFA, Ahmad W, Khan S, BiBi A, et al. Security and privacy challenges, solutions, and performance evaluation in AIoT-enabled smart societies. Comput Model Eng Sci. 2026;146(3):1–10. doi:10.32604/cmes.2026.075882.
7. Huma ZE, Jan SU, Ahmad J, Buchanan W, Pitropakis N. Adversarial machine learning in IoT security: a comprehensive survey. ACM Comput Surv. 2026;58(8):1–35. doi:10.1145/3785665.
8. Malik J, Muthalagu R, Pawar PM. A systematic review of adversarial machine learning attacks, defensive controls, and technologies. IEEE Access. 2024;12:99382–421. doi:10.1109/ACCESS.2024.3423323.
9. Liu Y, Guo B, Fang Y, Li H, Yu Z. The emergence of IoT 3.0: from IoT, AIoT to MIoT. IEEE Internet Things Mag. 2026:1–10. doi:10.1109/miot.2025.3646765.
10. Shanthi AVK, Swain S, Nandigama NC, Thenmozhi L. Artificial intelligence and internet of things. New Delhi, India: BR Publications; 2026.
11. Alahi MEE, Sukkuea A, Tina FW, Nag A, Kurdthongmee W, Suwannarat K, et al. Integration of IoT-enabled technologies and artificial intelligence (AI) for smart city scenario: recent advancements and future trends. Sensors. 2023;23(11):5206. doi:10.3390/s23115206.
12. López Delgado JL, López Ramos JA. A comprehensive survey on generative AI solutions in IoT security. Electronics. 2024;13(24):4965. doi:10.3390/electronics13244965.
13. Braga CM, Suarez-Bárcena Á, Serrano MA, Fernández-Medina E. TrustAIoT: a framework for building trustworthy AIoT platforms. Internet Things. 2025;34(5):101751. doi:10.1016/j.iot.2025.101751.
14. Lin YD, Lin TH, Sudyana D, Lai YC. AI for AIoT as a service: AI to configure models, capacities, and tasks. IEEE Internet Things J. 2026;13(11):25210–23. doi:10.1109/JIOT.2026.3676376.
15. Nakamura Y. AIoT-driven health behavioral security: vision and challenges. ACM Trans Comput Healthc. 2026;7(1):1–7. doi:10.1145/3771551.
16. Liu W. AIoT: the integrating of artificial intelligence and the Internet of Things. ACE. 2025;118(1):64–8. doi:10.54254/2755-2721/2025.20805.
17. Luo X, Xiong S, Jia X, Zeng Y, Chen X. AIoT-enabled data management for smart agriculture: a comprehensive review on emerging technologies. IEEE Access. 2025;13(2):102964–93. doi:10.1109/ACCESS.2025.3578751.
18. Huang YP, Khabusi SP. Artificial intelligence of things (AIoT) advances in aquaculture: a review. Processes. 2025;13(1):73. doi:10.3390/pr13010073.
19. Luo X, Wang A, Zhang X, Huang K, Wang S, Chen L, et al. Toward intelligent AIoT: a comprehensive survey on digital twin and multimodal generative AI integration. Mathematics. 2025;13(21):3382. doi:10.3390/math13213382.
20. Paolone G, Pilotti F, Piazza A. Evolution of the concept of device moving from Internet of Things to artificial intelligence of things. IJECE. 2024;14(6):7236–43. doi:10.11591/ijece.v14i6.
21. Li K, Li C, Yuan X, Li S, Zou S, Sohail Ahmed S, et al. Zero-trust foundation models: a new paradigm for secure and collaborative artificial intelligence for Internet of Things. IEEE Internet Things J. 2025;12(22):46269–93. doi:10.1109/jiot.2025.3603957.
22. Guerraoui R, Gupta N, Pinot R. Byzantine machine learning: a primer. ACM Comput Surv. 2024;56(7):1–39. doi:10.1145/3616537.
23. Szadeczky T, Bederna Z. Risk, regulation, and governance: evaluating artificial intelligence across diverse application scenarios. Secur J. 2025;38(1):35. doi:10.1057/s41284-025-00495-z.

24. Parums DV. Editorial: review articles, systematic reviews, meta-analysis, and the updated preferred reporting items for systematic reviews and meta-analyses (PRISMA) 2020 guidelines. Med Sci Monit. 2021;27:e934475. doi:10.12659/msm.934475.
25. Zainuddin AA, Zakirudin MAZ, Syafiq Zulkefli AS, Mazli AM, Al Syahmi Mohd Wardi M, Fazail MN, et al. Artificial intelligence: a new paradigm for distributed sensor networks on the Internet of Things: a review. IJPCC. 2024;10(1):16–28. doi:10.31436/ijpcc.v10i1.414.
26. Chi C, Yin Z, Liu Y, Chai S. A trusted cloud–edge decision architecture based on blockchain and MLP for AIoT. IEEE Internet Things J. 2024;11(1):201–16. doi:10.1109/jiot.2023.3300845.
27. Parihar V, Malik A, Bhawna, Bhushan B, Chaganti R. From smart devices to smarter systems: the evolution of artificial intelligence of things (AIoT) with characteristics, architecture, use cases and challenges. In: AI models for blockchain-based intelligent networks in IoT systems: concepts, methodologies, tools, and applications. Berlin/Heidelberg, Germany: Springer; 2023. p. 1–28. doi:10.1007/978-3-031-31952-5_1.
28. Villar E, Martín Toral I, Calvo I, Barambones O, Fernández-Bustamante P. Architectures for industrial AIoT applications. Sensors. 2024;24(15):4929. doi:10.3390/s24154929.
29. Rong G, Xu Y, Tong X, Fan H. An edge-cloud collaborative computing platform for building AIoT applications efficiently. J Cloud Comput. 2021;10(1):36. doi:10.1186/s13677-021-00250-w.
30. Cheng L, Gu Y, Liu Q, Yang L, Liu C, Wang Y. Advancements in accelerating deep neural network inference on AIoT devices: a survey. IEEE Trans Sustain Comput. 2024;9(6):830–47. doi:10.1109/TSUSC.2024.3353176.
31. Aloraini F, Javed A, Rana O, Burnap P. Adversarial machine learning in IoT from an insider point of view. J Inf Secur Appl. 2022;70(3):103341. doi:10.1016/j.jisa.2022.103341.
32. Hou KM, Diao X, Shi H, Ding H, Zhou H, de Vaulx C. Trends and challenges in AIoT/IIoT/IoT implementation. Sensors. 2023;23(11):5074. doi:10.3390/s23115074.
33. Tian Z, Cui L, Liang J, Yu S. A comprehensive survey on poisoning attacks and countermeasures in machine learning. ACM Comput Surv. 2023;55(8):1–35. doi:10.1145/3551636.
34. Zhang X, Hu M, Xia J, Wei T, Chen M, Hu S. Efficient federated learning for cloud-based AIoT applications. IEEE Trans Comput Aided Des Integr Circuits Syst. 2021;40(11):2211–23. doi:10.1109/TCAD.2020.3046665.
35. Wang S, Ko RKL, Bai G, Dong N, Choi T, Zhang Y. Evasion attack and defense on machine learning models in cyber-physical systems: a survey. IEEE Commun Surv Tutor. 2024;26(2):930–66. doi:10.1109/COMST.2023.3344808.
36. Guo Y, Tan Z, Lu S. Towards model-centric security for IoT systems. In: Proceedings of the 2020 29th International Conference on Computer Communications and Networks (ICCCN); 2020 Aug 3–6; Honolulu, HI, USA. p. 1–6. doi:10.1109/icccn49398.2020.9209689.
37. Xiong Z, Cai Z, Takabi D, Li W. Privacy threat and defense for federated learning with non-i.i.d. data in AIoT. IEEE Trans Ind Inform. 2022;18(2):1310–21. doi:10.1109/TII.2021.3073925.
38. Liu S, Guo B, Fang C, Wang Z, Luo S, Zhou Z, et al. Enabling resource-efficient AIoT system with cross-level optimization: a survey. IEEE Commun Surv Tutor. 2024;26(1):389–427. doi:10.1109/comst.2023.3319952.
39. Harbi Y, Medani K, Gherbi C, Aliouat Z, Harous S. Roadmap of adversarial machine learning in Internet of Things-enabled security systems. Sensors. 2024;24(16):5150. doi:10.3390/s24165150.
40. Alaba FA. IoT architecture layers. In: Internet of Things: a case study in Africa. Berlin/Heidelberg, Germany: Springer; 2024. p. 65–85. doi:10.1007/978-3-031-67984-1_4.
41. Lai Y, Zhou J, Zhang X, Zhou K. Toward certified robustness of graph neural networks in adversarial AIoT environments. IEEE Internet Things J. 2023;10(15):13920–32. doi:10.1109/JIOT.2023.3263384.
42. Cao X, Wang W, Chen Z, Wang X, Yang M. Resilient integrated control for AIOT systems under DoS attacks and packet loss. Electronics. 2024;13(9):1737. doi:10.3390/electronics13091737.
43. Vassilev A, Oprea A, Fordyce A, Anderson H, Davies X, Hamin M. Adversarial machine learning: a taxonomy and terminology of attacks and mitigations: NIST AI 100-2e2025. Gaithersburg, MD, USA: National Institute of Standards and Technology (NIST); 2025.
44. Krishna RR, Priyadarshini A, Jha AV, Appasani B, Srinivasulu A, Bizon N. State-of-the-art review on IoT threats and attacks: taxonomy, challenges and solutions. Sustainability. 2021;13(16):9463. doi:10.3390/su13169463.

45. Siam SI, Ahn H, Liu L, Alam S, Shen H, Cao Z, et al. Artificial intelligence of things: a survey. ACM Trans Sen Netw. 2025;21(1):1–75. doi:10.1145/3690639.
46. Doss S. AIoT and the governance of security. In: Information security governance using artificial intelligence of things in smart environments. Boca Raton, FL, USA: CRC Press; 2025.
47. Wang K, Xu P, Chen CM, Kumari S, Shojafar M, Alazab M. Neural architecture search for robust networks in 6G-enabled massive IoT domain. IEEE Internet Things J. 2021;8(7):5332–9. doi:10.1109/JIOT.2020.3040281.
48. Wang Y, Tan YA, Baker T, Kumar N, Zhang Q. Deep fusion: crafting transferable adversarial examples and improving robustness of industrial artificial intelligence of things. IEEE Trans Ind Inform. 2023;19(6):7480–8. doi:10.1109/TII.2022.3168874.
49. Patil A, Bansode R. EdgeAware: distributed threat detection and monitoring system for smart cities. Int J AI ML Data Sci. 2026;1(2):e005. doi:10.66261/fa1cep06.
50. Marchand A, Imine Y, Ouarnoughi H, Tarridec T, Gallais A. Firmware integrity protection: a survey. IEEE Access. 2023;11:77952–79. doi:10.1109/access.2023.3298833.
51. Hovakimyan G, Bravo JM. Evolving strategies in machine learning: a systematic review of concept drift detection. Information. 2024;15(12):786. doi:10.3390/info15120786.
52. Xu L, Han Z, Zhao D, Li X, Yu F, Chen C. Addressing concept drift in IoT anomaly detection: drift detection, interpretation, and adaptation. IEEE Trans Sustain Comput. 2024;9(6):913–24. doi:10.1109/TSUSC.2024.3386667.
53. Wei X, Kaust, Pu B, Zhao S, Lu J, Wu B. Visual adversarial attacks and defenses in the physical world: a survey. ACM Comput Surv. 2026;58(10):1–36. doi:10.1145/3793659.
54. Imtiaz N, Wahid A, Ul Abideen SZ, Muhammad Kamal M, Sehito N, Khan S, et al. A deep learning-based approach for the detection of various Internet of Things intrusion attacks through optical networks. Photonics. 2025;12(1):35. doi:10.3390/photonics12010035.
55. Alzuhair A, Alghaihab A. The design and optimization of an acoustic and ambient sensing AIoT platform for agricultural applications. Sensors. 2023;23(14):6262. doi:10.3390/s23146262.
56. Son BD, Hoa NT, Van Chien T, Khalid W, Ferrag MA, Choi W, et al. Adversarial attacks and defenses in 6G network-assisted IoT systems. IEEE Internet Things J. 2024;11(11):19168–87. doi:10.1109/jiot.2024.3373808.
57. Ouyang N, Shatte A, Lu Z, Chen C, Xiang W. Artificial intelligence in mitigating security threats for lightweight IoT devices: a survey of technologies, protocols, and future challenges. IEEE Internet Things J. 2026;13(7):14096–111. doi:10.1109/JIOT.2025.3649316.
58. Qiu F. Beyond physical constraints: AI-powered cross-modal interference suppression in wireless multi-modal communication systems. In: Proceedings of the 2025 IEEE 5th International Conference on Computer Communication and Artificial Intelligence (CCAI); 2025 May 23–25; Haikou, China. p. 361–6. doi:10.1109/ccai65422.2025.11189794.
59. Han C, Yang T, Sun X, Cui Z. Secure hierarchical federated learning for large-scale AI models: poisoning attack defense and privacy preservation in AIoT. Electronics. 2025;14(8):1611. doi:10.3390/electronics14081611.
60. Kim S, Lim CY, Rho Y. Spatio-temporal analysis of dependent risk with an application to cyberattacks data. Ann Appl Stat. 2024;18(4):1952. doi:10.1214/24-aoas1952.
61. Yang J, Zheng J, Baker T, Tang S, Tan YA, Zhang Q. Clean-label poisoning attacks on federated learning for IoT. Expert Syst. 2023;40(5):e13161. doi:10.1111/exsy.13161.
62. Zhao Y, Gong X, Lin F, Chen X. Data poisoning attacks and defenses in dynamic crowdsourcing with online data quality learning. IEEE Trans Mob Comput. 2023;22(5):2569–81. doi:10.1109/TMC.2021.3133365.
63. Cheng SM, Hong BK, Hung CF. Attack detection and mitigation in MEC-enabled 5G networks for AIoT. IEEE Internet Things Mag. 2022;5(3):76–81. doi:10.1109/IOTM.001.2100144.
64. Rahaman H, Chatterjee A, Bhunia S. Fortifying AI systems: emerging threats and security countermeasures. SN Comput Sci. 2026;7(3):227. doi:10.1007/s42979-025-04658-y.
65. Chen J, Gao Y, Shan J, Peng K, Wang C, Jiang H. Manipulating supply chain demand forecasting with targeted poisoning attacks. IEEE Trans Ind Inform. 2023;19(2):1803–13. doi:10.1109/TII.2022.3175958.

66. Zhou Y, Wang H, He C, Li X. LoRA-augmented ConvMixed-ViT architecture for adaptive compressive sensing in resource-constrained AIoT scenarios. IEEE Internet Things J. 2025;12(17):35848–60. doi:10.1109/JIOT.2025.3580548.
67. Xu H, Seng KP, Ang LM, Smith J. Decentralized and distributed learning for AIoT: a comprehensive review, emerging challenges, and opportunities. IEEE Access. 2024;12:101016–52. doi:10.1109/ACCESS.2024.3422211.
68. Jia Y, Lin F, Sun Y. A novel federated learning aggregation algorithm for AIoT intrusion detection. IET Commun. 2024;18(7):429–36. doi:10.1049/cmu2.12744.
69. Wang Z, Zhu Y, Wang D, Han Z. Federated analytics informed distributed industrial IoT learning with non-IID data. IEEE Trans Netw Sci Eng. 2023;10(5):2924–39. doi:10.1109/TNSE.2022.3187992.
70. Wang Z, Liu S, Guo B, Yu Z, Zhang D. CrowdLearning: a decentralized distributed training framework based on collectives of trusted AIoT devices. IEEE Trans Mob Comput. 2024;23(12):13420–37. doi:10.1109/TMC.2024.3427636.
71. Zhang C, Yang S, Mao L, Ning H. Anomaly detection and defense techniques in federated learning: a comprehensive review. Artif Intell Rev. 2024;57(6):150. doi:10.1007/s10462-024-10796-1.
72. Sharma A, Marchang N. Probabilistic sign flipping attack in federated learning. In: Proceedings of the 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT); 2024 Jun 24–28; Kamand, India. p. 1–6. doi:10.1109/icccnt61001.2024.10725463.
73. Xiao X, Tang Z, Li C, Jiang B, Li K. SBPA: sybil-based backdoor poisoning attacks for distributed big data in AIoT-based federated learning system. IEEE Trans Big Data. 2024;10(6):827–38. doi:10.1109/TBDATA.2022.3224392.
74. Xu J, Guo B, Chen F, Shen Y, Dai S, Dai C, et al. A defense mechanism for federated learning in AIoT through critical gradient dimension extraction. Comput Commun. 2025;236(5):108114. doi:10.1016/j.comcom.2025.108114.
75. Tu T, He Z, Zheng Z, Zheng Z, Jiang J, Gong Y, et al. Toward lifelong unseen task processing with a lightweight unlabeled data Schema for AIoT. IEEE Internet Things J. 2025;12(4):3441–52. doi:10.1109/jiot.2024.3396282.
76. Bai L, Hu H, Ye Q, Li H, Wang L, Xu J. Membership inference attacks and defenses in federated learning: a survey. ACM Comput Surv. 2025;57(4):1–35. doi:10.1145/3704633.
77. Yang W, Wang S, Wu D, Cai T, Zhu Y, Wei S, et al. Deep learning model inversion attacks and defenses: a comprehensive survey. Artif Intell Rev. 2025;58(8):242. doi:10.1007/s10462-025-11248-0.
78. Mei Y, Wang W, Liang Y, Liu Q, Chen S, Wang T. Privacy-enhanced cooperative storage scheme for contact-free sensory data in AIoT with efficient synchronization. ACM Trans Sen Netw. 2024;20(4):1–19. doi:10.1145/3617998.
79. Hina S, Abbas Q, Ahmed K. Adversarial attacks on artificial Intelligence of Things-based operational technologies in theme parks. Internet Things. 2025;32(1):101654. doi:10.1016/j.iot.2025.101654.
80. Zhao W, Alwidian S, Mahmoud QH. Adversarial training methods for deep learning: a systematic review. Algorithms. 2022;15(8):283. doi:10.3390/a15080283.
81. Zhao S, Wang X, Wei X. Mitigating accuracy-robustness trade-off via balanced multi-teacher adversarial distillation. IEEE Trans Pattern Anal Mach Intell. 2024;46(12):9338–52. doi:10.1109/TPAMI.2024.3416308.
82. Jia X, Li J, Gu J, Bai Y, Cao X. Fast propagation is better: accelerating single-step adversarial training via sampling subnetworks. IEEE Trans Inf Forensics Secur. 2024;19:4547–59. doi:10.1109/TIFS.2024.3377004.
83. Zhong M, Tandon R. SPLITZ: certifiable robustness via split Lipschitz randomized smoothing. IEEE Trans Inf Forensics Secur. 2025;20:9099–112. doi:10.1109/TIFS.2025.3595402.
84. Faza GA, Wauters J, Cuzzolin F, Hallez H, Moens D. Direct interval propagation methods using neural-network surrogates for uncertainty quantification in physical systems surrogate model. Knowl Based Syst. 2026;341(1):115824. doi:10.1016/j.knosys.2026.115824.
85. Liu T, Xia J, Ling Z, Fu X, Yu S, Chen M. Efficient federated learning for AIoT applications using knowledge distillation. IEEE Internet Things J. 2023;10(8):7229–43. doi:10.1109/JIOT.2022.3229374.
86. Alrayes FS, Zakariah M, Amin SU, Iqbal Khan Z, Helal M. Intrusion detection in IoT systems using denoising autoencoder. IEEE Access. 2024;12(2):122401–25. doi:10.1109/access.2024.3451726.
87. Pelekis S, Koutroubas T, Blika A, Berdelis A, Karakolis E, Ntanos C, et al. Adversarial machine learning: a review of methods, tools, and critical industry sectors. Artif Intell Rev. 2025;58(8):226. doi:10.1007/s10462-025-11147-4.

88. Qiao Y, Sathyanarayana NB, Shi C, He Z, Wang T, Hou T. A survey on adversarial machine learning: attacks, defenses, real-world applications, and future research directions. Neurocomputing. 2026;671(6433):132670. doi:10.1016/j.neucom.2026.132670.
89. Ahmad M, Habib S, Tariq F. Enhancing model robustness in federated learning: a systematic literature review of Byzantine-resilient aggregation methods. VFAST Trans Softw Eng. 2025;13(2):196–227. doi:10.21015/vtse.v13i2.2163.
90. Aga DT, Chintanippu R, Mowri RA, Siddula M. Exploring secure and private data aggregation techniques for the Internet of Things: a comprehensive review. Discov Internet Things. 2024;4(1):28. doi:10.1007/s43926-024-00064-7.
91. Kornaros G, Martini S, Spyridakis N, Polixronidou A. Trusted execution, updating of AI/ML models and sensor data through isolation on IoT devices. In: Proceedings of the 2025 IEEE Annual Congress on Artificial Intelligence of Things (AIoT); 2025 Dec 3–5; Osaka, Japan. p. 861–5. doi:10.1109/aiot66900.2025.00139.
92. Román R, Arjona R, Baturone I. A lightweight remote attestation using PUFs and hash-based signatures for low-end IoT devices. Future Gener Comput Syst. 2023;148(3):425–35. doi:10.1016/j.future.2023.06.008.
93. Ural O, Yoshigoe K. SecurePoL: integration of watermarking with proof-of-learning to enhance security against spoofing attacks. IEEE Access. 2025;13(3):213067–91. doi:10.1109/ACCESS.2025.3642198.
94. Jamshidi S, Wahab OA, Herrero R, Khomh F, Bellaïche M, Keivanpour S, et al. Think fast: real-time IoT intrusion reasoning using IDS and LLMs at the edge gateway. IEEE Internet Things J. 2026;13(8):15485–513. doi:10.1109/jiot.2026.3656738.
95. Lin L, Xu Z, Chen CM, Wang K, Hassan MR, Alam MGR, et al. Understanding the impact on convolutional neural networks with different model scales in AIoT domain. J Parallel Distrib Comput. 2022;170(12):1–12. doi:10.1016/j.jpdc.2022.07.011.
96. Liu Y, Li S, Wang X, Xu L. A review of hybrid cyber threats modelling and detection using artificial intelligence in IIoT. CMES. 2024;140(2):1233–61. doi:10.32604/cmes.2024.046473.
97. Vo QV, Haq TM, Montague P, Abraham T, Abbasnejad E, Ranasinghe DC. Certified but fooled! breaking certified defenses with ghost certificates. AAAI. 2026;40(12):9621–9. doi:10.1609/aaai.v40i12.37924.
98. Liang Z, Wu T, Liu W, Xue B, Yang W, Wang J, et al. Towards robust neural networks via a global and monotonically decreasing robustness training strategy. Front Inform Technol Electron Eng. 2023;24(10):1375–89. doi:10.1631/fitee.2300059.
99. Abdelhady G, Ghandoura A, Motwakel A, Alajmi A. Hybrid machine learning anomaly detection and lightweight zero trust authentication for LoRaWAN networks. IEEE Access. 2026;14(8):18387–407. doi:10.1109/ACCESS.2026.3660731.
100. Hallaji E, Razavi-Far R, Saif M, Wang B, Yang Q. Decentralized federated learning: a survey on security and privacy. IEEE Trans Big Data. 2024;10(2):194–213. doi:10.1109/TBDATA.2024.3362191.
101. Ahmad Abbas S, Khan MJ, Rao MS. FedTrust-Net: a federated learning–based trust-aware security framework for resilient *Ad Hoc* wireless networks. Int J Intell Eng Syst. 2025;18(11):979–91. doi:10.22266/ijies2025.1231.60.
102. Van Landuyt D, Halasz D, Verreydt S, Weyns D. Towards understanding trust in self-adaptive systems. In: Proceedings of the 19th International Symposium on Software Engineering for Adaptive and Self-Managing Systems; 2024 Apr 15–16; Lisbon, Portugal. p. 207–13. doi:10.1145/3643915.3644100.
103. Li X, Chen Y, Wang C, Shen C. When deep learning meets differential privacy: privacy, security, and more. IEEE Netw. 2021;35(6):148–55. doi:10.1109/MNET.001.2100256.
104. Sánchez Sánchez PM, Huertas Celdrán A, Bovet G, Martínez Pérez G. Adversarial attacks and defenses on ML- and hardware-based IoT device fingerprinting and identification. Future Gener Comput Syst. 2024;152(12):30–42. doi:10.1016/j.future.2023.10.011.
105. Suzaki K, Nakajima K, Oi T, Tsukamoto A. TS-perf: general performance measurement of trusted execution environment and rich execution environment on intel SGX, arm TrustZone, and RISC-V keystone. IEEE Access. 2021;9:133520–30. doi:10.1109/ACCESS.2021.3112202.
106. Lucan Orăşan I, Seiculescu C, Căleanu CD. A brief review of deep neural network implementations for ARM cortex-M processor. Electronics. 2022;11(16):2545. doi:10.3390/electronics11162545.

107. Alotaibi B. A review of resilient IoT systems: trends, challenges, and future directions. Appl Sci. 2026;16(4):2079. doi:10.3390/app16042079.
108. Vivek Menon U, Babu Kumaravelu V, Vinoth Kumar C, Rammohan A, Chinnadurai S, Venkatesan R, et al. AI-powered IoT: a survey on integrating artificial intelligence with IoT for enhanced security, efficiency, and smart applications. IEEE Access. 2025;13(2):50296–339. doi:10.1109/access.2025.3551750.
109. Fu H, Rao J, Deng F, Wang Y, Zhao B, Liu Z, et al. AIoT: artificial intelligence and the Internet of Things for monitoring and prognosis of systems and structures. IEEE Trans Instrum Meas. 2025;74(3):9700232. doi:10.1109/TIM.2025.3557124.
110. Awaisi KS, Ye Q, Sampalli S. A survey of industrial AIoT: opportunities, challenges, and directions. IEEE Access. 2024;12(9):96946–96. doi:10.1109/ACCESS.2024.3426279.
111. Gouissem A, Abualsaud K, Yaacoub E, Khattab T, Guizani M. Collaborative Byzantine resilient federated learning. IEEE Internet Things J. 2023;10(18):15887–99. doi:10.1109/JIOT.2023.3266347.
112. Mengistu TM, Kim T, Lin JW. A survey on heterogeneity taxonomy, security and privacy preservation in the integration of IoT, wireless sensor networks and federated learning. Sensors. 2024;24(3):968. doi:10.3390/s24030968.
113. Mohammed BA, Al-Shareeda MA, Hamzah AE, Alhasnawi BN, Homod RZ, Alkhabra YA, et al. Security challenges and solutions in Internet of Medical Things (IoMT) communication: a review. J King Saud Univ Comput Inf Sci. 2026;118(5):2719. doi:10.1007/s44443-026-00658-x.
114. Pasin M, Ménétrey J, Felber P, Schiavoni V, Heyn HM, Knauss E, et al. Methods for requirements engineering, verification, security, safety, and robustness in AIoT systems. In: Shaping the future of IoT with edge intelligence: how edge computing enables the next generation of IoT applications. New York, NY, USA: River Publishers; 2023. p. 197–228. doi:10.1201/9781032632407-12.
115. Cinà AE, Demontis A, Biggio B, Roli F, Pelillo M. Energy-latency attacks via sponge poisoning. Inf Sci. 2025;702(6):121905. doi:10.1016/j.ins.2025.121905.