Review

A Review of Resilient IoT Systems: Trends, Challenges, and Future Directions

Bandar Alotaibi *

Posted Date: 19 December 2025

doi: 10.20944/preprints202512.1717.v1

Keywords: internet of things; resilience; reliability; information security

 doi:10.20944/preprints202512.1717.v1

Review

A Review of Resilient IoT Systems: Trends, Challenges, and Future Directions

Bandar Alotaibi

Department of Information Technology, University of Tabuk, Tabuk 47512, Saudi Arabia; b-alotaibi@ut.edu.sa

Abstract

The Internet of Things (IoT) is increasingly embedded in critical infrastructures across healthcare, energy, transportation, and industrial automation, yet its pervasiveness introduces substantial security and resilience challenges. This paper presents a comprehensive review of recent advances in IoT resilience, focusing on developments reported between 2022 and 2025. A layered taxonomy is proposed to organize resilience strategies across hardware, network, learning, application, and governance layers, addressing adversarial, environmental, and hybrid stressors. The survey systematically classifies and compares more than forty representative studies encompassing deep learning under adversarial attack, generative and ensemble intrusion detection, hardware- and protocol-level defenses, federated and distributed learning, and trust- and governance-based approaches. A comparative analysis shows that while adversarial training, GAN-based augmentation, and decentralized learning improve robustness, they often have limitations, being confined to specific datasets or attack scenarios without extensive validation in large-scale deployments. The study highlights challenges in adaptive benchmarking, cross-layer integration, and explainable resilience, concluding with future directions for creating antifragile IoT systems that can self-heal and adapt to evolving cyber-physical threats.

Keywords: internet of things; resilience; reliability; information security

1. Introduction

The Internet of Things (IoT) has rapidly evolved from a network of connected sensors into the backbone of modern cyber–physical ecosystems [1]. It now powers critical infrastructures across healthcare, energy, transportation, industry, and public safety—enabling billions of devices to collect, communicate, and act upon data in real time [2,3]. This pervasive connectivity promises smarter cities, more efficient industries, and safer environments [4]. Yet as IoT scales and complexity increases, it introduces new points of fragility. Devices often operate with limited energy, unstable connectivity, and minimal computational resources [5], making them vulnerable to both cyber and physical disruptions [6,7]. Ensuring that such systems can continue functioning reliably—even when attacked, damaged, or degraded—has therefore become a central design imperative [8]. This capability is known as resilience: the capacity of a system to withstand, recover from, and adapt to disruptions without losing essential functionality [9–12].

Over the past few years, researchers have made notable progress in strengthening IoT resilience [13,14]. Efforts range from defending deep learning models against adversarial attacks and data poisoning [15,16] to hardening networks against packet loss, interference, and energy depletion [17,18]. Advances in hardware-based trust anchors—such as Physical Unclonable Functions (PUFs)—offer new mechanisms for secure device authentication [19,20]. Likewise, explainable and trust-aware governance frameworks are emerging to ensure transparency and accountability in autonomous IoT decision-making [21]. However, despite these developments, research on IoT resilience remains disconnected, with limited integration across different layers and threat models. Most studies focus on a single layer of the IoT stack or address isolated threats such as model evasion or communication faults. Little attention has been paid to hybrid stressors —cyberattacks coinciding with environmental

 doi:10.20944/preprints202512.1717.v1

or operational disturbances —or to antifragility, the concept that systems improve through exposure to stress.

This gap motivates the present review. While several surveys exist on IoT security or adversarial machine learning, no prior work has comprehensively analyzed original research (2022–2025) across the full spectrum of IoT resilience—from hardware reliability and network protocols to learning, application, and governance layers. To address this need, this paper makes three primary contributions:

1. A unified taxonomy of IoT resilience, categorizing methods by stressor type (adversarial, environmental, hybrid) and by the IoT layer they protect (hardware, protocol, learning, application, or governance).
2. A systematic, critical analysis of recent research, including federated adversarial learning, PUF-based authentication, GAN-driven defenses, explainable-AI governance, and multi-agent recovery frameworks.
3. A forward-looking discussion of challenges and future directions, emphasizing the need for hybrid-stress benchmarks, cross-layer defense orchestration, and the integration of antifragile design principles into practical IoT deployments.

The remainder of the paper is organized as follows: Section 2 compares our survey with the related reviews; Section 3 introduces the taxonomy of IoT resilience; Section 4 discusses stressor and layer dimensions in detail; Section 5 reviews resilience strategies across key domains; Section 6 demonstrates the practical implications of adversarial resilience in IoT through a controlled case study; Section 7 outlines open challenges and future research directions; and Section 8 concludes the survey paper.

2. Related Work

Research on resilience and security in connected systems spans multiple areas: machine learning security, cyber-physical systems (CPS), industrial networking, 6G, and IoT trust. Below, we position our survey relative to seven representative reviews and domain surveys, and Table 1 compares these surveys and contrasts them with our survey.

Chakraborty et al. [22] map the adversarial ML landscape across vision, speech, and other modalities, cataloging threat models (white and black-box), adversarial attacks (e.g., FGSM), and defenses (e.g., adversarial training, feature squeezing, and defensive distillation). Its strength is a unified taxonomy of attack/defense mechanics and evaluation pitfalls. However, it is mainly domain-agnostic and only briefly touches on IoT-specific constraints (energy, latency, multi-hop communication, and device heterogeneity). Our survey differs from this review because we focus on end-to-end IoT resilience, integrating adversarial robustness with environmental stressors, protocol/hardware anchors, and governance. Then, we assess deployability under edge constraints and realistic testbeds.

Goyal et al. [23] focus on language models and review defense strategies (i.e., adversarial training and detection, perturbation detection, and robustness certificate-based). The survey comprehensively highlights issues such as the use of adversarial training in most defense strategies, the lack of automatic generation of adversarial instances, and the generalization of adversarial training. However, it is scoped to text pipelines and does not engage with network telemetry, RF signals, or cyber-physical dynamics central to IoT. Our survey differs from this one because we cover non-text IoT data (traffic flows, RF/in-phase and quadrature (IQ) signals, images, time series) and cross-layer defenses, tying robustness to operational constraints (e.g., synchronization loss, packet drops, thermal limits).

Aaqib et al. [24] reviewed trust and reputation in IoT devices. The authors introduced a taxonomy to categorize models by trust management approach, including traditional systems and those based on artificial intelligence. Additionally, the authors compare and analyze various system methods and applications using performance metrics such as scalability, delay, cooperativeness, and efficiency. While rich in governance and policy aspects, it treats adversarial ML and cross-layer technical robustness only tangentially. This survey differs from theirs because we integrate trust controllers and XAI with

 doi:10.20944/preprints202512.1717.v1

technical defenses (PUFs, lightweight crypto, robust FL), demonstrating how trust scoring couples with adversarial detection and on-device constraints.

Table 1. Comparison of Related Surveys and Our Contributions.

Survey	Primary Scope	Layers Covered	Adversarial ML Depth	IoT Constraints Considered	What Our Survey Adds
Chakraborty et al. [22]	General ML (vision and speech)	Model level	High (attacks and defenses taxonomy)	Low (domain-agnostic)	IoT-specific datasets and testbeds, cross-layer integration, deployability on constrained devices
Goyal et al. [23]	Text and NLP	Model, data	High (text attacks and defenses)	Low (NLP-centric)	Non-text IoT data (traffic, RF, images), cross-layer coupling to protocols and hardware
Aaqib et al. [24]	Trust and reputation systems	Application and governance	Low (conceptual)	Medium (trust overheads)	Bridges trust controllers and XAI with adversarial defenses and edge feasibility
Segovia-Ferreira et al. [25]	CPS resilience phases	System and architecture	Medium (anomaly)	Medium (CPS ops focus)	Latest adversarial and federated IoT methods, dataset-driven comparisons
Khaloopour et al. [26]	6G network resilience	Network and service mgmt	Low (device learning)	Medium (6G orchestration)	Links radio and edge learning defenses to protocol and hardware anchors for IoT
Alrumaih et al. [27]	Industrial IoT (IIoT)	Network and ops	Low–Medium (anomaly)	High (industrial constraints)	Industrial adversarial ensembles, robust FL, deployability analysis
Berger et al. [28]	General IoT resilience	Multi-layer (high level)	Low–Medium (pre-2022 focus)	Medium (broad)	Updated 2022–2025 coverage, paper-by-paper summaries, five-section taxonomy
Ours	IoT resilience 2022–2025	Hardware → Governance	High (DL and ViT, GAN, FL, certified trends)	High (latency, energy, thermal, radio noise, packet loss)	Unified cross-layer synthesis, testbed-aware analysis, actionable gaps and roadmap

Segovia-Ferreira et al. [25] explore cyber-resilience techniques that enhance the resilience of cyber-physical systems (CPS) against cyber-attacks. Additionally, the article highlights challenges related to practical aspects of cyber-resilience, including metrics, evaluation methods, and testing environments. The survey emphasizes system-level strategies and standards but provides limited coverage of modern adversarial ML and federated learning in the context of IoT data/compute realities. This survey differs from theirs in that we reviewed recent adversarial and federated methods, added empirical comparisons (e.g., IoT-23, ToN-IoT, CICIoT2023), and assessed edge feasibility.

Khaloopour et al. [26] review resilient systems and introduce the resilience-by-design (RBD) concept for 6G communication networks. The survey outlines RBD principles, proposes an interdisciplinary approach for integrating them across 6G layers, and discusses associated challenges. The review illustrates RBD through 6G use cases and presents open research problems on 6G resilience. While it provides a valuable network-centric perspective, device-level learning resilience, dataset practices, and attack models are not its primary focus. This survey differs from theirs because we connect radio/edge learning with protocol and hardware mitigations (PUFs), bridging 6G network goals with device-side robustness.

Alrumaih et al. [27] analyze cyber resilience strategies for industrial networks, particularly Industrial Control Systems (ICSs) and the IIoT. They assess resilient network components and evaluate current cyber resilience frameworks based on defense mechanisms and survivability strategies. The survey exposes operational gaps (legacy stacks, safety constraints) but provides limited depth on adversarial ML and federated defenses now entering industrial analytics. Key challenges and current needs in cyber resilience are identified, along with requirements for future schemes and research directions. This survey differs from theirs because we synthesize industrial adversarial defenses (e.g., ensembles, robust FL aggregation) with hardware/protocol anchors to assess deployability on factory-floor devices.

 doi:10.20944/preprints202512.1717.v1

Berger et al. [28] organize failures and countermeasures across IoT layers and discuss reliability and safety considerations. However, it predates, or only lightly covers, the latest GAN-based augmentation, vision-transformer defenses, robust FL, and trust-XAI integration that have emerged since 2022. This survey differs from their survey because we deliver an updated, data-backed taxonomy spanning five focused subsections (deep learning under attack; generative and ensemble IDS; hardware or protocol; federated; trust, XAI, and governance).

3. Background and Definitions

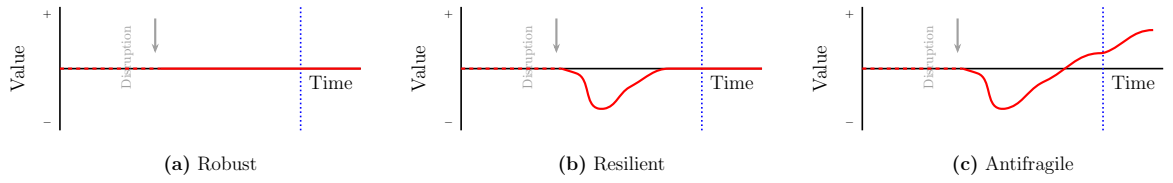
3.1. From Robustness to Resilience to Antifragility

In IoT devices and cyber-physical systems, three progressively advanced paradigms—robustness, resilience, and antifragility—play a crucial role in guiding the design and evaluation of trustworthy and dependable systems [29,30]. Understanding these concepts and their practical implications is essential for researchers and practitioners who are building next-generation IoT deployments. Table 2 compares these three concepts by defining them, providing a typical metric, and providing a typical IoT example for each.

Table 2. Comparison of robustness, resilience, and antifragility in IoT systems.

Property	Definition	Typical Metric	Example in IoT
Robustness	Withstands bounded disturbances	Deviation in accuracy or service under fixed faults or noise	Sensor fusion tolerant to up to 10% packet loss
Resilience	Recovers or adapts after disruptions	Area under resilience curve, time to recovery	Intrusion detector recovers accuracy after network attack
Antifragility	Improves through exposure to shocks	Increase in performance after perturbation, negative regret	intrusion detection system (IDS) that becomes more accurate after adversarial training with real attack traffic

Robustness is the most fundamental property. As illustrated in Figure 1a, robustness refers to a system's ability to endure disturbances or uncertainties without a significant decline in performance [31–33]. In simple terms, a robust Internet of Things (IoT) device or algorithm is capable of operating as intended, as long as any environmental changes, faults, or attacks fall within predetermined, manageable limits. For instance, a sensor fusion algorithm used in a smart home may be designed to handle up to 10% packet loss before the accuracy of its estimations drops below acceptable levels. Robustness is typically static: if the perturbation remains within the designed safe region, the system output is largely unaffected [34].



(a) Robust (b) Resilient (c) Antifragile

Figure 1. (**a**) the robust system maintains performance under bounded perturbations, (**b**) the resilient system recovers after a performance dip, (**c**) after a disruption, the normal performance is restored in an antifragile system, and the system surpasses its original baseline by learning and adapting from the stressor.

Resilience enhances the concept of trustworthiness by not only focusing on the ability to withstand disruptions but also on the dynamic processes of recovery and adaptation [35–37]. In resilient systems,

 doi:10.20944/preprints202512.1717.v1

when a disturbance—like a cyberattack or sensor fault—causes a drop in performance, the system can absorb the impact, adapt, and restore or even improve its performance [38]. Recovery may involve switching to backup modes, using redundant data sources, or activating adaptive controls. The resilience curve visualizes this process by plotting system performance over time. While a robust system's curve remains flat during disruptions, as shown in Figure 1a, a resilient system may dip but gradually returns to its baseline performance, as shown in Figure 1b. The area under this curve during the disruption and recovery period quantifies the loss of resilience. The speed and completeness of recovery distinguish a highly resilient system from one that is only robust [39].

Antifragility is the most advanced and ambitious concept. Coined by Nassim Nicholas Taleb [40] and now adapted in AI [41,42], IoT [43], and cyber-physical systems (CPSs) [44] literature, as shown in Figure 1c, an antifragile system not only survives stress and disruption but actually improves because of it [45]. For instance, an antifragile IoT anomaly detector could use real-world attack traffic as new training data, adjusting its detection thresholds and classifier boundaries to improve accuracy over time. Similarly, a resilient CPS may leverage environmental disturbances, such as rare sensor failures, to identify new fault modes and strengthen its redundancy mechanisms, thereby expanding its operational range.

In summary, while traditional IoT systems have focused on robustness — resisting known stressors—future systems must be designed for resilience, capable of rapid recovery, and ultimately for antifragility, where each disruption becomes an opportunity to learn, adapt, and strengthen the system for the future.

3.2. Stressors in IoT Systems

IoT systems operate at the intersection of the digital and physical worlds [46]. They sense, compute, and communicate under real-world constraints [47]—meaning they face not only algorithmic attacks but also physical and environmental disruptions [48]. Sensors can drift or fail, wireless channels can become noisy or congested, batteries can drain unpredictably [49], and adversaries continuously adapt to system defenses. To reason about these diverse challenges, it is helpful to categorize stressors—factors that degrade reliability or performance—into three broad families: Adversarial stressors, caused intentionally by malicious agents; Environmental or operational stressors, arising naturally from the physical or logistical environment; and Hybrid stressors, where cyber and physical disruptions co-occur.

Figure 2 shows the various stressors affecting IoT systems, organized along three axes: Intent, Origin, and Severity. The x-axis distinguishes natural stressors, like sensor drift, from malicious ones, such as model poisoning. The y-axis indicates whether stressors originate in the physical environment or in the data and model layers, while the z-axis measures their severity. Blue data points represent environmental stressors, such as hardware degradation; red data points indicate adversarial stressors, such as poisoning; and green data points illustrate hybrid stressors, such as adversarial traffic and signal interference. This figure highlights the diverse threats based on their intent, origin, and impact on system performance. It serves as a framework for designing targeted resilience strategies—ensuring that defenses deployed at one layer (e.g., protocol-level rate limiting) complement those at another (e.g., adversarially robust learning).

Figure 3 shows a heatmap that categorizes IoT ecosystem stressors by their impact on system layers. The vertical axis includes stressors like packet loss, interference, and data poisoning, while the horizontal axis represents IoT stack layers, such as hardware, protocols, and governance. Color intensity indicates impact level (0 = low, 3 = high), allowing for quick identification of critical vulnerabilities. For example, packet loss and energy scarcity have the highest effect at the device and network layers, while data poisoning and inference-time evasion are more prevalent at the learning and application layers. This visualization shows that resilience strategies should be tailored to the specific stress points within the stack, rather than using a one-size-fits-all defense approach.

Figure 4 shows the coupling strength between IoT layers, illustrating how disturbances in one layer can affect the entire system. Diagonal cells indicate dependencies within layers, while off-diagonal

 doi:10.20944/preprints202512.1717.v1

values highlight interactions between layers. The results indicate that while each layer retains primary sensitivity to its own dynamics, significant coupling exists between adjacent layers—such as between network and learning layers or between application and governance layers—where disruptions can cascade upward or downward. This coupling map shows that resilience in IoT is fundamentally systemic: merely strengthening individual components is not enough unless it is aligned with a cross-layer design and adaptive feedback systems that work together to prevent cascading failures.

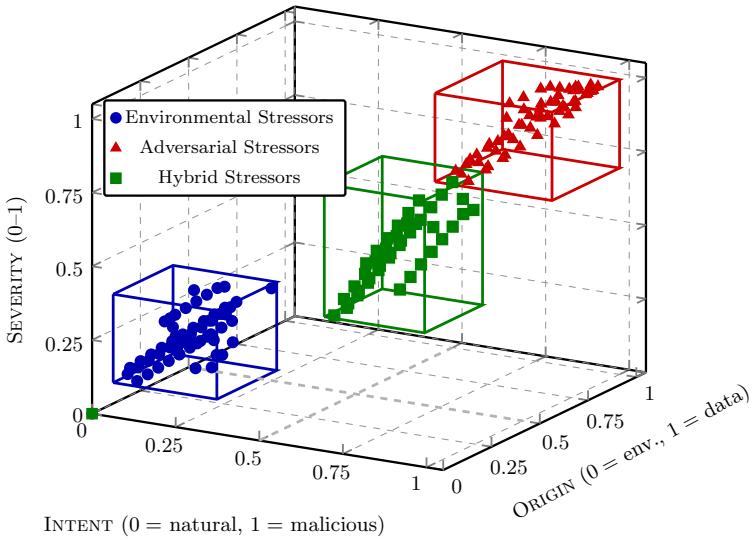


Figure 2. Three-dimensional landscape of IoT stressors organized by intent, origin, and severity.

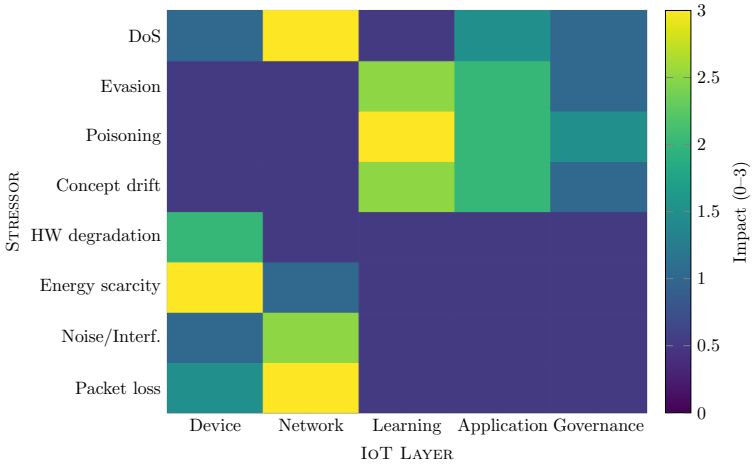


Figure 3. Mapping of IoT stressors across the system stack.

Adversarial stressors target and exploit vulnerabilities in algorithms, communication protocols, or operational assumptions, compromising system and network integrity [50]. These stressors are intentional and adaptive, aiming to mislead, exhaust, or subvert IoT components, and they can be categorized into specific types.

 doi:10.20944/preprints202512.1717.v1

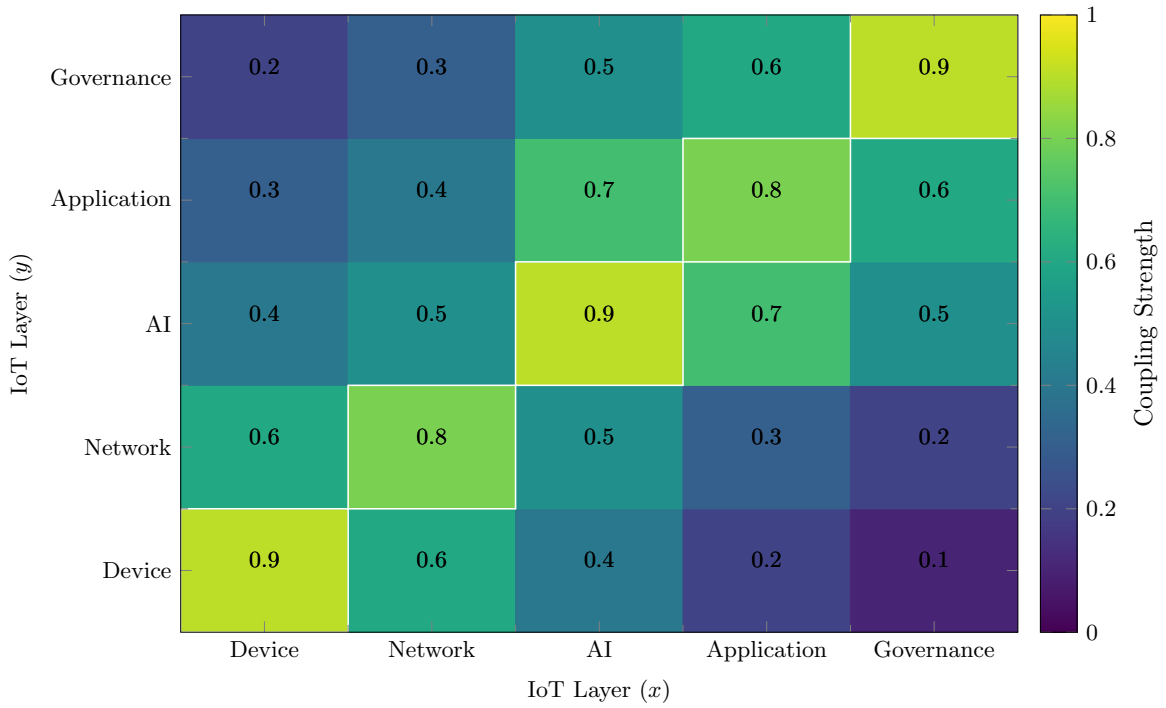


Figure 4. Cross-layer coupling strength among IoT resilience mechanisms.

- Data poisoning: before or during training, an adversary corrupts a fraction $p \in [0, 1]$ of the training set so that the learned model exhibits degraded or adversarially biased behavior at deployment [51,52]. Common tactics include label flipping, clean labeling, backdoor or trigger attacks, and model or gradient poisoning. Label-flip poisoning changes ground-truth labels while leaving features intact [53]. Clean-label poisoning creates what appear to be legitimate examples that manipulate the decision boundary [54]. Backdoor or trigger attacks exploit an uncommon pattern, leading the model to misclassify any input that includes it [55]. Model or gradient poisoning in federated environments can inject biases into global aggregation (for instance, as a result of Byzantine attacks). As shown in Figure 5, we simulate label-flip poisoning on ToN-IoT by flipping a fraction $p \in \{0, 0.01, 0.05, 0.10, 0.20\}$ of training labels uniformly at random while keeping validation and test sets clean. Test Accuracy and Macro-F1 decrease smoothly as p increases. The small absolute drops (e.g., $\Delta\text{Acc} \approx 0.0026$ at $p = 0.20$) indicate that random label flips primarily act as label noise, to which this tabular model is relatively robust—likely due to high class separability and redundancy in features. This is a lower bound on risk. Structured attacks, such as class-conditional or rare-class targeting, clean-label poisoning, and backdoor triggers, can cause larger errors at lower probabilities p. The effects may also amplify in federated training without robust aggregation methods. Attack goals when using poisoning attacks range from availability (overall error inflation) to targeted or class-specific failures on chosen classes or trigger patterns. Mitigations include distributional validation and data filtering (outlier and influence diagnostics), robust training losses and regularization, differential clipping/noise in FL, robust aggregation (trimmed mean, coordinate-wise median, Krum), holdout audits with canaries, and cryptographic or signed provenance logs to trace data lineage [51,52,56].

 doi:10.20944/preprints202512.1717.v1

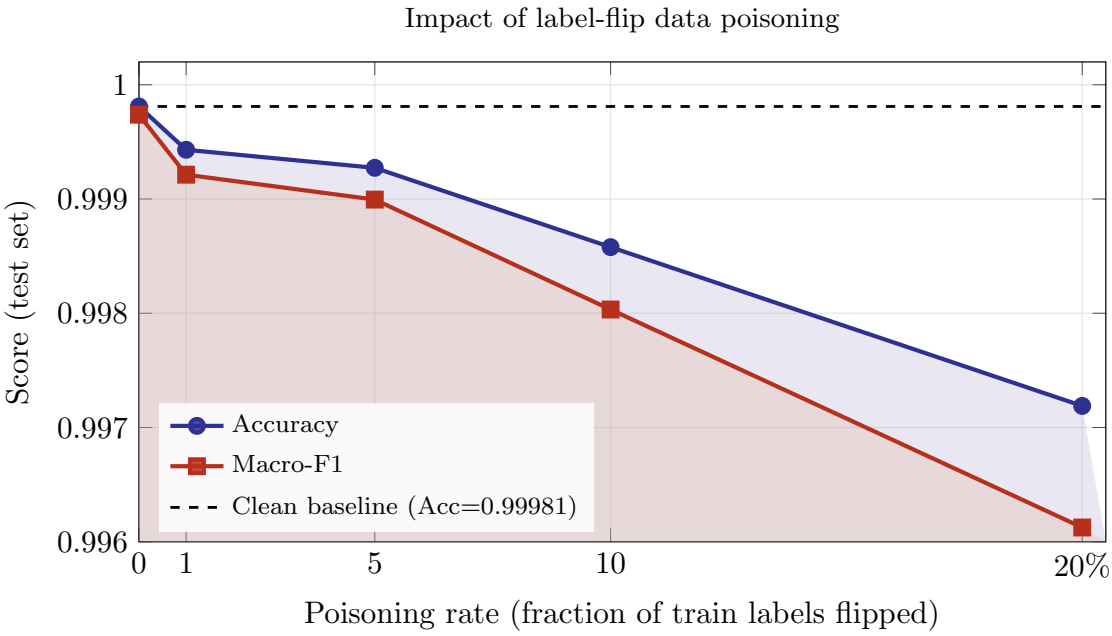


Figure 5. Simulated label-flip poisoning on ToN-IoT by flipping a fraction of training labels.

- Evasion at inference: once models are deployed, an adversary can add carefully tuned, norm-bounded noise to inputs so that the perturbations remain visually or statistically imperceptible but still induce misclassification [57]. Representative attacks include: fast gradient sign method (FGSM) [58], basic iterative method (BIM) [59], projected gradient descent (PGD) [60], momentum iterative method (MIM) [61], and DeepFool [62]. FGSM takes a single step that prompts each input feature in the direction that most increases the loss (the sign of the gradient), with step size ϵ, yielding a small ℓ_∞-bounded change that can already flip the prediction. BIM/Iterated-FGSM repeats FGSM in many small steps (step size α), clipping after each step to keep the total perturbation within the ℓ_∞ ball of radius ϵ. This iterative refinement typically finds stronger adversarial examples than a single step. PGD further strengthens BIM by starting from a random point inside the ϵ-ball and then iterating gradient steps with projection back to the ball; random restarts help avoid weak local optima, which is why PGD is a widely used (strong first-order) baseline [63]. MIM (Momentum Iterative FGSM) also iterates, but it accumulates a momentum term (an exponential average of recent gradients) to stabilize the update direction; this often improves transferability, making the crafted examples more effective even on unseen models [64]. DeepFool approximates the classifier's decision boundary and iteratively moves the input in the smallest direction that crosses that boundary, aiming for near-minimal ℓ_2 change; unlike the ϵ-bounded methods above, it does not fix a budget in advance but adapts the perturbation to reach misclassification with minimal effort [65]. To illustrate the effect of evasion at inferecne attacks, we experimented using ToN-IoT dataset and control the maximum perturbation size by a budget $\epsilon > 0$ (e.g., $\epsilon \in \{0.01, 0.025, 0.05, 0.10, 0.15\}$), where ϵ bounds the per-feature deviation under the chosen norm (typically ℓ_∞); here, the symbol $\in$ means is an element of (i.e., chosen from the set). Figure 6 shows the test accuracy under ℓ_∞ evasion as the perturbation budget ϵ increases. Relative to the clean baseline where the accuracy is 0.9996, FGSM declines gradually to 0.299 at $\epsilon = 0.15$. At the same time, PGD and BIM collapse rapidly (e.g., at $\epsilon = 0.05$, accuracy is 0.006 and 0.170, respectively, and ≈ 0 for $\epsilon \geq 0.10$). MIM degrades slowest among iterative methods (0.962 at $\epsilon = 0.025$, 0.923 at $\epsilon = 0.05$, 0.656 at $\epsilon = 0.15$). The shaded band ($\epsilon \leq 0.05$) denotes our imperceptible regime, where iterative attacks already cause large drops (e.g., PGD 0.273, BIM 0.288 at $\epsilon = 0.025$). Even with small budgets within our imperceptible range, iterative attacks such as PGD and BIM can substantially degrade accuracy, while MIM tends to degrade more slowly on our model; DeepFool, which does not use an explicit ϵ, instead adapts its steps to cross the

 doi:10.20944/preprints202512.1717.v1

nearest decision boundary. A representative example of evasion at inference, a small modulation change in a wireless packet could bypass an intrusion detection model. To address such attacks, it is a valuable practice to integrate one or more of the following techniques: adversarial training, randomized smoothing, and/or cross-modal input consistency checks.

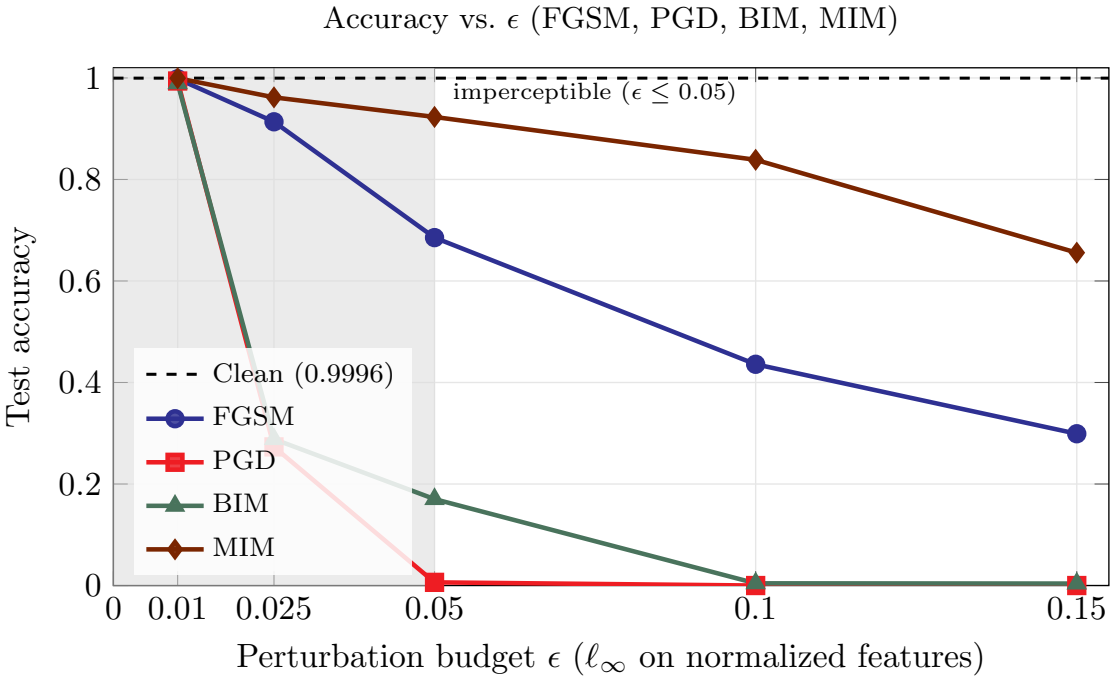


Figure 6. ToN-IoT accuracy vs. perturbation budget for inference-time evasion attacks.

- Model extraction and inversion: in cloud/edge APIs, an adversary can iteratively query a deployed (black-box) model to extract a high-fidelity surrogate (recovering decision boundaries or even approximating parameters), or to invert the model to reconstruct features of sensitive training records [66]. Leakage surfaces include top-1 labels, confidence scores (soft probabilities), and auxiliary signals (temperature-scaling artifacts, calibration curves), which together facilitate knowledge distillation and amplify privacy risks—potentially exposing patient attributes, user habits, or network signatures [67,68]. As shown in Figure 7, to illustrate model extraction, we train a high-accuracy teacher (test Acc $=$ 0.99981) and simulate an attacker who can fit a student using teacher-labeled queries under three disclosures: label-only (hard top-1), soft-probs (full confidences), and noisy-soft (soft-probs with Laplace noise, $T=2$). Across budgets $\{1\,\text{k}, 5\,\text{k}, 10\,\text{k}, 25\,\text{k}, 50\,\text{k}\}$, student accuracy rises monotonically toward the teacher: for example, at 1 k queries the student attains $\{0.9824, 0.9785, 0.9831\}$ for {label-only, soft-probs, noisy-soft}, and at 50 k reaches $\{0.9985, 0.9988, 0.9986\}$ respectively—still below the dashed teacher baseline. Soft-prob disclosure yields the strongest extraction at high budgets (e.g., 0.9988 at 50 k), while injecting calibrated noise slightly suppresses student performance (noisy-soft 0.9986) with minimal utility loss for benign users, and label-only sits in between. The x-axis is log-scaled to emphasize early-query efficiency: most gains occur by 10 k–25 k queries, underscoring the value of early throttling and score suppression in production APIs. Practical approches that can be used to mitigate model extraction and inversion include API governance (authentication, rate or volume limits, burst throttling, per-class quota), output minimization (label-only, confidence truncation or quantization, randomized response on scores), noise mechanisms (Laplace or Gaussian noise on logits or probabilities), and privacy-preserving training (DP-SGD or post-training DP), complemented by extraction watermarking, server-side audit tests (monitor agreement patterns vs. natural data), and adaptive blocking when query statistics deviate from benign usage [67,68].

 doi:10.20944/preprints202512.1717.v1

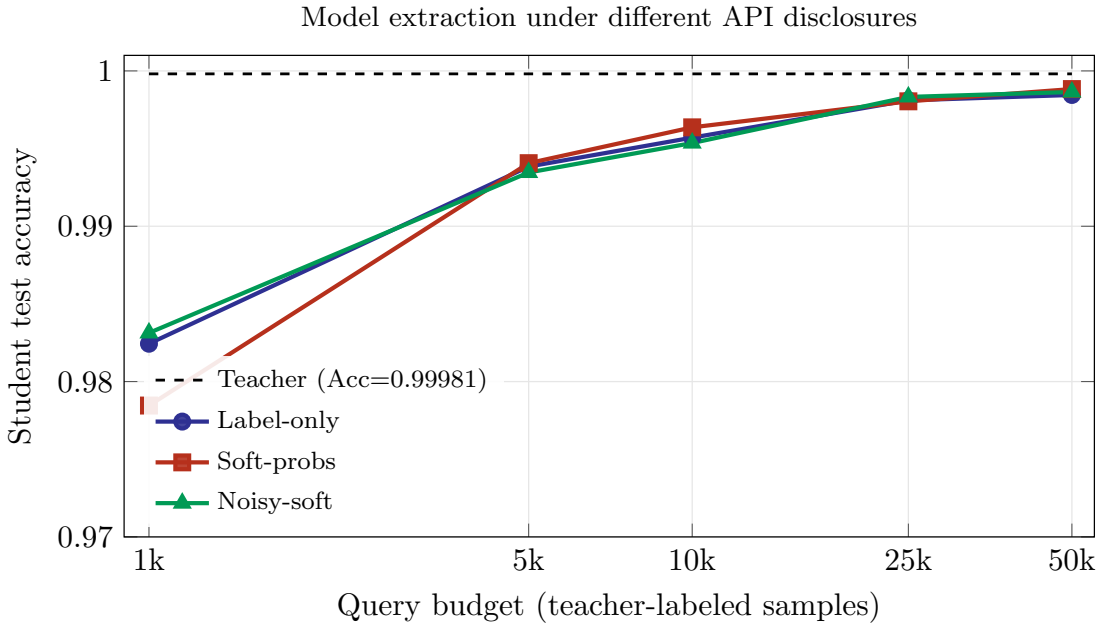


Figure 7. Model extraction using a teacher and a simulated attacker under three disclosures.

- Protocol spoofing: beyond software-level API abuse, adversaries can impersonate endpoints by manipulating the RF channel itself. Common attack vectors include satellite deception, such as GPS spoofing, link-layer replay attacks like Roll-Jam, and waveform injections that imitate a device's modulation [69–71]. In a typical replay attack, illustrated in Figure 8, an attacker captures a fob's rolling code R_n at time t_0, preventing the vehicle's receiver from decoding it at time t_1. The attacker then replays R_n within the receiver's acceptance window $[n, n+W]$ at time t_2, causing the vehicle to respond with an unlock or acknowledgment (ACK) at time t_3. To mitigate protocol spoofing, various methods can be used, including cross-layer defenses, desynchronization-resilient rolling-code updates, PHY hardening, spectrum-level defenses, and RF fingerprinting. Defense at the cross-layer can be employed, such as cryptographic freshness with nonce-based challenge–response and strict single-use counters (no grace window or $W = 0$ for critical actions). Desynchronization-resilient rolling code updates use session-bound keys and monotone counters with limited resynchronization attempts. Physical layer security is enhanced through time-of-flight and multi-antenna angle-of-arrival checks to limit plausible emitter geometry. RF fingerprinting that exploits device-specific imperfections (carrier-frequency offset, I/Q imbalance, transient shape) and channel-state/timing features to reject cloned waveforms [72,73]. Spectrum-level defenses (frequency-hopping spread spectrum (FHSS) or direct-sequence spread spectrum (DSSS), adaptive carrier sensing) with anomaly analytics on energy, inter-frame timing, and Doppler patterns. Operational controls—lockout or backoff after failed frames, localized rate-limits, and out-of-band second factors (e.g., proximity ultra wide band (UWB) ranging)—further shrink the attack surface while preserving usability [74].
- Denial-of-Service/Distributed DOS (DoS/DDoS): in IoT environments, attackers exploit resource limitations to overwhelm bandwidth at gateways or drain device resources (CPU, memory, battery), resulting in service interruptions or cascading failures [75–77]. Botnet-driven swarms of compromised endpoints (e.g., cameras or smart plugs) can generate high-rate floods or carefully timed bursts that defeat naive token buckets, overwhelm queueing buffers, and trigger retransmission storms, further reducing goodput. Let L denote the offered load (benign and malicious) and C the gateway capacity. In the benign regime, goodput G roughly increases with $\min\{L, C\}$. However, under attack, issues like queue overflows and packet drops cause G to drop sharply below a critical threshold $L^\star < C$, well before reaching nominal capacity C. This behavior is reflected in the goodput versus offered load curve in Figure 9. Rate limiting delays collapse goodput; puzzles

 doi:10.20944/preprints202512.1717.v1

add slight latency but reduce bot amplification; and edge filtering maximizes performance beyond C by eliminating unnecessary traffic before it occupies limited buffer space. In practice, robust deployments combine the following practical countermeasures with lightweight anomaly scoring, short control loops for threshold tuning, and fail-open exceptions for safety-critical flows to avoid overblocking. Practical countermeasures span admission control and in-network enforcement, trading reactivity against collateral damage. Per-packet/flow rate limiting caps burstiness and bounds worst-case load, raising $L^{\star}$ while keeping implementation lightweight at edge routers [78]. Client puzzles that are stateless and adjustable in difficulty shift computational tasks to suspected sources, limiting the impact of bot swarms on CPU usage and reducing unnecessary gateway processing. The difficulty of the puzzles can be modified based on observed queue occupancy to maintain quality of service for compliant devices [79]. In-network filtering at access gateways (e.g., prefix-/behavior-based filters, Bloom-filter aggregates, or programmable data-plane rules) removes malicious traffic near its ingress and prevents backpressure into constrained subnets, preserving goodput even when $L > C$ [80].

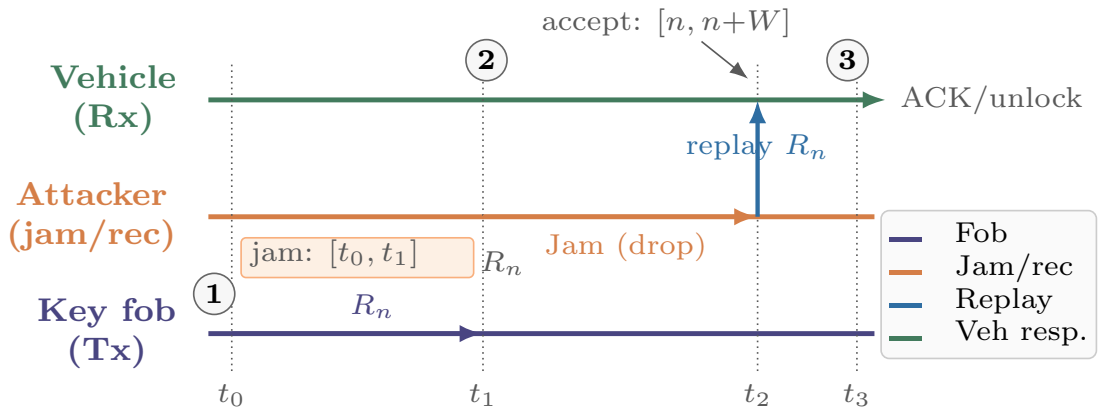


Figure 8. An illustration of link-layer replay attack in the form of Roll–Jam on rolling codes.

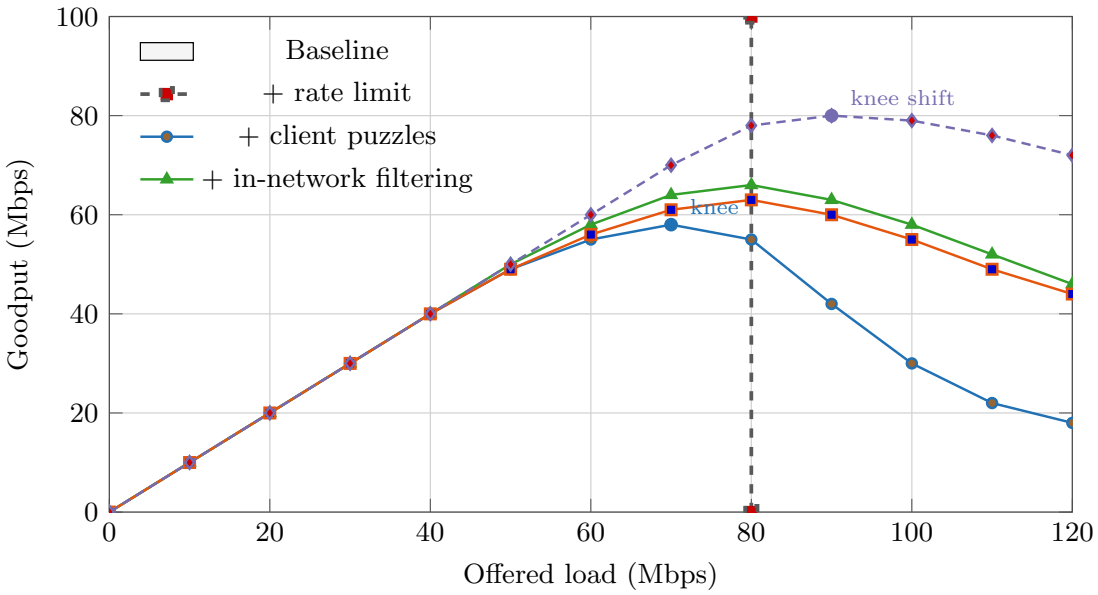


Figure 9. Goodput vs. offered load under IoT DoS/DDoS.

Environmental and operational stressors originate from physical reality rather than malice [81]. Yet, their cumulative effect can be equally damaging, especially in large-scale or remote deployments. These stressors can be categorized as follows.

 doi:10.20944/preprints202512.1717.v1

- Packet loss and desynchronization: wireless IoT connections—especially within low-power wide-area networks (LPWANs)—often encounter burst losses caused by interference, duty-cycle limitations, and synchronization drift [82,83]. In rolling-code systems, missing even one frame can lead to a permanent authentication failure [84,85]. To mitigate these problems, it is advantageous to use self-synchronizing codes [86], selective retransmissions [87], and interleaved packet scheduling [88].
- Noise and interference: the unlicensed industrial, scientific, and medical (ISM) bands that most IoT devices rely on are congested [89], leading to signal collisions, higher bit-error rates, and increased timing jitter. In industrial control systems, this congestion can destabilize control loops [90]. To deal with noise and interference, the following measures can be employed: frequency hopping, adaptive modulation/coding, and sensor fusion with uncertainty-aware weighting [91,92]. To better illustrate the impact of physical-layer resilience mechanisms under realistic channel conditions, Figure 10 shows the packet loss probability in relation to the signal-to-noise ratio (SNR) for uncoded, forward error correction (FEC)-coded, and FEC and automatic repeat request (ARQ) transmission modes. The uncoded link shows the classical waterfall region where even small SNR degradations cause orders-of-magnitude increases in loss. By introducing FEC (rate 1/2), the reliability curve shifts toward lower SNR values—representing a coding gain of several decibels. Adding a single ARQ layer significantly reduces effective loss, demonstrating that lightweight hybrid error-control strategies can enhance environmental resilience without redesigning protocols. The shaded area represents the typical SNR range of -3 to $+5$ dB for LPWAN, crucial for maintaining connectivity in noisy industrial and outdoor settings.
- Energy scarcity: battery-powered and energy-harvesting devices often enter aggressive sleep modes [93], resulting in sparse or delayed data streams. For instance, a remote soil sensor may report only once per hour on cloudy days. It is helpful to implement event-driven sensing, compressive sampling, and lightweight on-device learning (i.e., tiny machine learning (TinyML)) [94–97].
- Hardware degradation: over time, sensors drift due to temperature fluctuations, aging, or wear [98]. PUFs used for device identification can also lose reliability under thermal stress [99]. To address these issues, specific measures can be utilized, such as periodic recalibration, helper data schemes for PUF correction, and redundant sensing with majority voting [100–102].
- Non-stationary data (concept drift): IoT data often evolves as environments, users, or firmware change [103]. A model trained on winter energy patterns may perform poorly during the summer months [104]. To mitigate this issue, the following measures can be employed: sliding-window retraining, online learning, and drift detection algorithms such as adaptive windowing (ADWIN) or the drift detection method (DDM) [105,106].

Real-world IoT incidents rarely involve a single, isolated stressor. Therefore, hybrid stressors, which are a combination of cyber and physical stressors, are common. Instead, disruptions often combine physical degradation and adversarial manipulation. For example, adversarial traffic might occur during a network outage, or GPS spoofing could coincide with heavy radio frequency (RF) interference. In autonomous drone fleets, an attacker might exploit simultaneous packet loss and model drift to induce collisions or miscoordination. These combined effects are hazardous because traditional countermeasures tend to address one dimension at a time. Figure 11 illustrates a simulated IoT scenario experiencing simultaneous adversarial and environmental disturbances. The shaded disruption window corresponds to a period in which a projected gradient descent (PGD) adversarial attack occurs concurrently with 30% network packet loss. Three cross-layer indicators are logged in parallel: physical layer, where signal-to-noise ratio (SNR) drops from $\approx$20 dB to 8 dB due to interference and energy depletion. Network layer, where packet loss rises sharply to 30%, representing congestion or wireless fading. Application layer, where model confidence (e.g., from a classifier or anomaly detector) declines from 0.94 to 0.60, showing the combined impact of noise and malicious perturbation. The red resilience curve tracks system-level utility or normalized performance over time.

 doi:10.20944/preprints202512.1717.v1

During the disruption, the performance falls by a magnitude $\Delta \approx 0.34$, marking the immediate drop. After mitigation and adaptation mechanisms are triggered (e.g., retransmission, adversarial retraining, redundancy), the system gradually recovers, reaching its baseline level within about 48 s—the recovery time. In some cases, the curve may show an overshoot, where the system slightly exceeds its original baseline due to adaptive learning or parameter re-tuning. The area under the curve between the onset of disruption and recovery represents the loss of resilience, which can be quantified as a function of the recovery speed and the magnitude of the drop. Overall, this figure illustrates how cross-layer monitoring (physical, network, and application) can aid in characterizing the resilience behavior of IoT systems under realistic compound stressors, providing a framework for quantitative benchmarking of recovery and adaptation mechanisms.

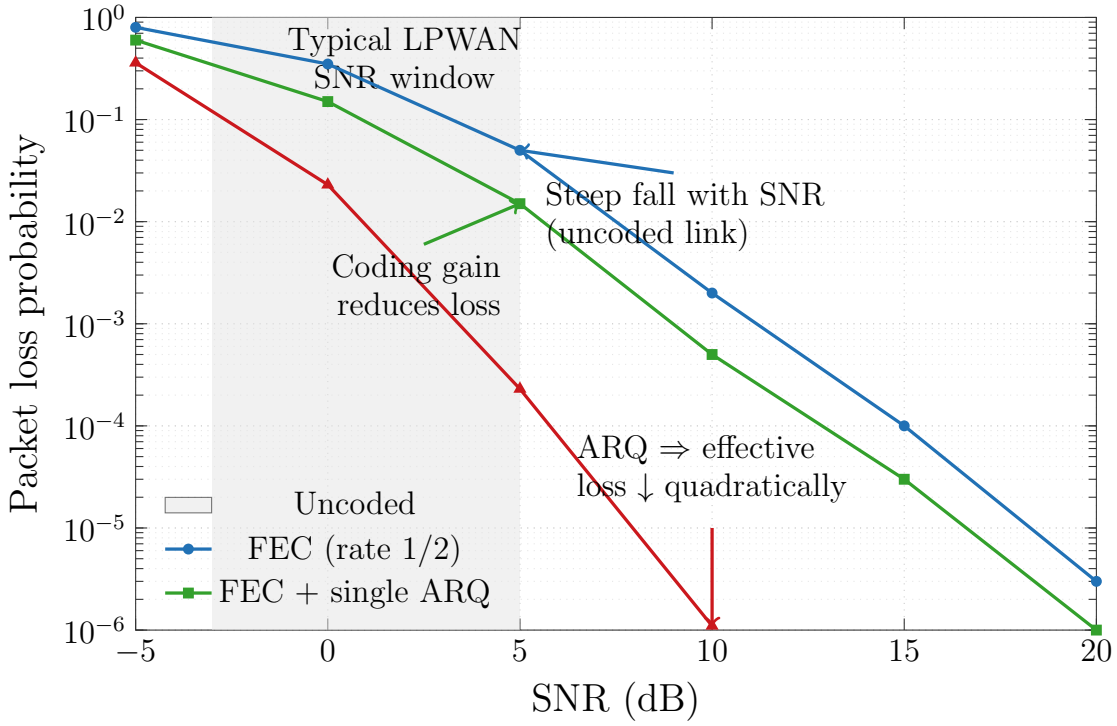


Figure 10. Packet loss probability versus SNR for different transmission strategies in low-power IoT networks.

In summary, IoT resilience cannot be understood by analyzing a single layer or stressor in isolation. Actual robustness emerges only when adversarial, environmental, and hybrid stressors are jointly modeled, tested, and mitigated across the whole system stack—from hardware and protocol layers up to AI-driven decision logic and governance mechanisms.

 doi:10.20944/preprints202512.1717.v1

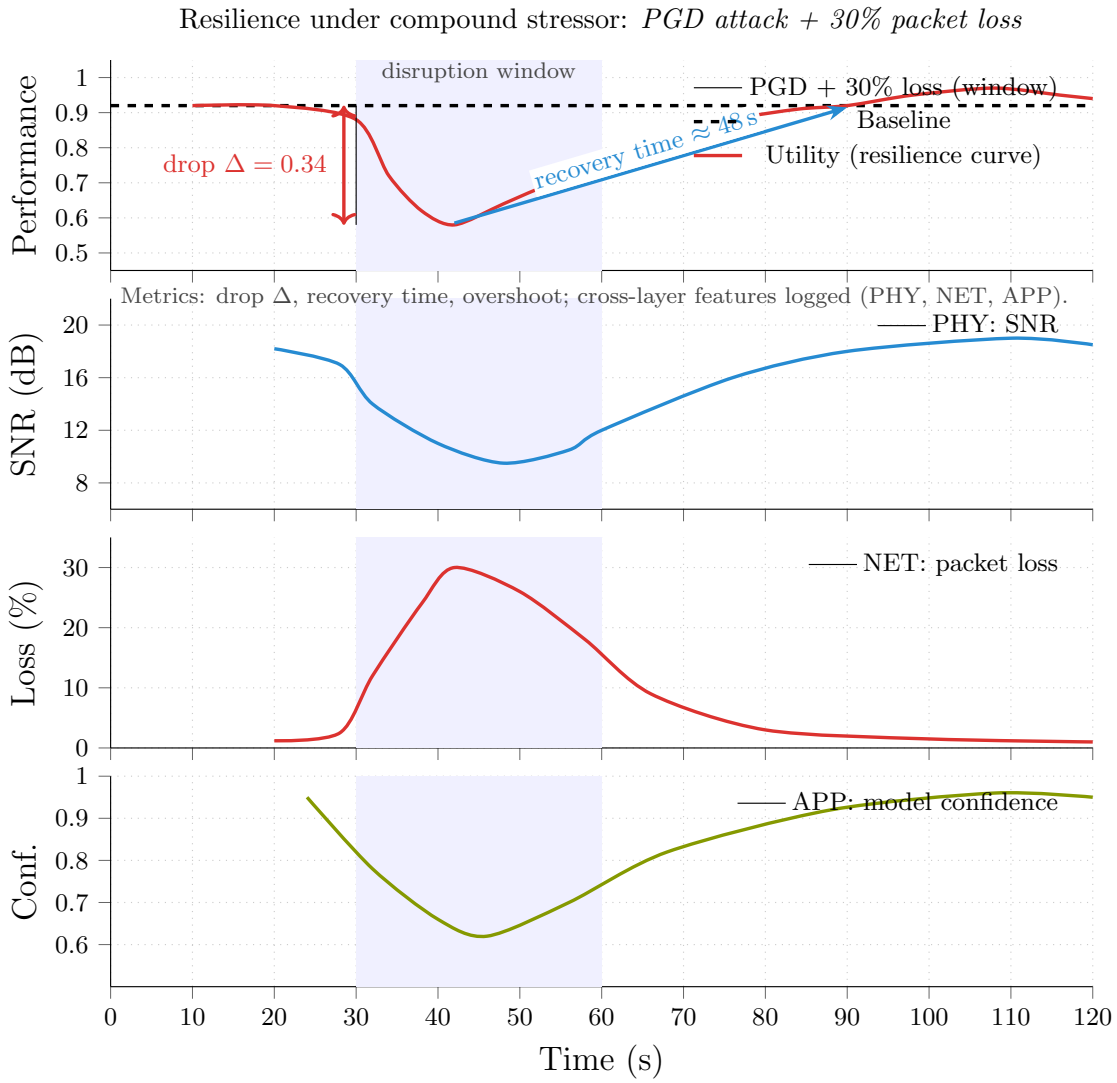


Figure 11. Resilience under Compound Stressors (PGD Attack and 30% Packet Loss).

3.3. Layers of IoT Resilience

Resilience in the IoT is not a single mechanism but an emergent property that arises from coordinated behavior across architectural layers. Each layer—from low-level sensors to high-level governance—contributes to anticipating disruptions, absorbing their impact, restoring functionality, and, ideally, improving through adaptation.

The device and hardware Layer comprises physical devices—sensors, actuators, and embedded controllers—operating under tight constraints in power, memory, and processing [107]. Typical stressors include noise, wear and tear, temperature drift, and physical tampering. Resilience strategies like Physically Unclonable Functions (PUFs), lightweight authentication, and self-calibrating redundant sensing secure identities and ensure reliable operation. PUFs create unique, device-specific keys from manufacturing variations, removing the need for vulnerable stored secrets [108]. Lightweight authentication (e.g., hash-based challenge–response) fits kilobyte-scale memory and low-MHz CPUs found in microcontrollers. Side-channel protections (masking, jitter, current flattening) reduce leakage through timing or power profiles. Redundant sensing and self-calibration help reduce drift and counteract the effects of aging components. By employing multiple sensors that cross-validate readings and periodically re-baseline, we can ensure ongoing accuracy. For instance, in smart agriculture, a soil moisture sensor utilizes redundant probes and supplementary data to maintain precision, even when faced with temperature fluctuations or partial sensor failures.

 doi:10.20944/preprints202512.1717.v1

Above the hardware lies the protocol and network layer that moves data through LPWAN, mesh, and 5G/edge links [109]. Stressors in this layer include packet loss, interference, congestion, and selective jamming. Resilience methods, such as resilient routing, flooding/DoS defenses, decentralized consensus, and cognitive radio/adaptive spectrum, focus on maintaining end-to-end delivery and trustworthy coordination. Resilient routing (multi-path, opportunistic forwarding) re-routes around failed or jammed nodes. Flooding and Denial-of-Service (DoS) defenses differentiate between legitimate bursts of activity—such as firmware rollouts—and malicious overloads through techniques like rate limiting and in-network filtering. Decentralized consensus mechanisms, including directed acyclic graph (DAG) ledgers and Proof-of-Authority, can tolerate network partitions while maintaining the auditability of updates and commands. Additionally, cognitive radio technology can adaptively shift to cleaner channels and modify coding and modulation in response to interference [110]. For instance, an IIoT mesh automatically detours telemetry through backup gateways during a channel-specific jamming incident, preserving control-loop stability.

The third layer is the learning and AI layer (inference and adaptation/processing layer) [111]. IoT increasingly relies on machine learning (ML) for anomaly detection, prediction, and closed-loop control. The stressors we face are dynamic and adversarial, including concept drift, data imbalance, poisoned updates, and evasion attacks. Resilience focuses on models that can adapt, recover, and resist manipulation, such as adversarial training, ensembles, generative augmentation, federated learning, continual learning, and graph neural networks (GNNs). Adversarial training and ensembles improve the robustness of classifiers against perturbed inputs and single-point failures [112]. Generative augmentation, such as generative adversarial networks (GANs) and variational autoencoders (VAEs), helps reconstruct missing modalities and balances rare events to stabilize learning when faced with loss or sparsity [113]. It is a good practice to apply federated and continual learning with a local adaptation enabler to ensure privacy [114]; robust aggregation methods (e.g., trimmed mean, median) mitigate the impact of poisoned clients [115]. GNNs leverage the topology of devices and flows to detect correlated anomalies and localize faults. For example, a federated smart-grid forecaster adapts to regional disturbances without sharing raw customer data, while robust aggregation downweights suspicious client updates.

Applications translate data into domain-specific actions (healthcare, transportation, manufacturing, energy) [116]. Stressors include partial outages, delayed data, and degraded sensing. Resilience in sectors like smart healthcare, Industrial IoT, and smart grids focuses on maintaining service continuity and ensuring a graceful degradation of performance. In smart healthcare, redundant biosignals from wearables are combined, so if a sensor fails, monitoring can continue with alerts that account for uncertainty. In IIoT, predictive maintenance and digital twins are utilized to simulate potential faults and develop pre-planned recovery strategies [117]. Smart grids manage distributed generation and demand response to ensure stability during cyber or weather-related disturbances [118]. For example, during a hospital network outage, electrocardiogram (ECG) wearables buffer data locally and synchronize later, maintaining clinical oversight with bounded data loss.

At the top sits the organizational, regulatory, and ethical backbone of resilience (i.e., governance and trust). Stressors include opaque decision-making, ambiguity of provenance, and policy non-compliance. Methods such as blockchain-anchored logs, explainable AI (XAI), trust scoring, and compliance engines enhance transparency, auditability, and adaptive policy. Blockchain-anchored logs document device identity, software lineage, and security events for post-incident forensics and incident response [119]. XAI supports operator trust during incident response by clarifying model rationale and failure modes [120]. Trust scoring continuously estimates the reliability of devices, links, and data sources, and decays scores for anomalous behavior [121]. Compliance engines codify jurisdictional rules and update enforcement policies as regulations evolve. For instance, in autonomous mobility, immutable event logs ensure that braking decisions are based on authentic, time-synchronized sensor data rather than spoofed inputs.

 doi:10.20944/preprints202512.1717.v1

True resilience emerges when these layers operate together. A noisy sensor reading (device layer) can be flagged by a GNN-based detector (AI layer), quarantined by a routing policy (network layer), explained to operators (XAI), and recorded for audit (governance). Conversely, a policy change (governance) can tighten model thresholds (AI), which in turn reconfigures sampling rates (device) to conserve energy during sustained attacks.

3.4. Metrics and Evaluation Frameworks

Assessing resilience in the IoT is fundamentally multi-faceted. Unlike evaluations of static security or performance, measuring resilience requires reflecting the evolving behavior of systems over time—how they deteriorate, recover, and adjust in response to challenges. Conventional metrics like accuracy or throughput offer only fleeting glimpses; genuine resilience requires measures that account for temporal, structural, and interpretability aspects, demonstrating both immediate robustness and enduring adaptability.

Figure 11 conceptually illustrates a typical resilience evaluation, where system performance drops after a disruption and then recovers over time. Key metrics, such as performance under stress, scalability, system-level trust, transparency, and interpretability, as well as hybrid and compound Benchmarks, can be used to evaluate resilience.

- Performance metrics under stress: the first dimension assesses how well a model or system maintains operational quality during and after disruptions. Common indicators include: accuracy or macro-F1 under perturbation, area under the resilience curve (AURC), and latency and energy overhead. Accuracy, or macro-F1, evaluates prediction consistency under difficult conditions, such as data degradation or network issues (e.g., 30% packet loss). The AURC measures performance from the start of a disruption to recovery, with a higher AURC indicating a faster or more complete restoration. Latency and energy overhead capture the efficiency cost of resilience mechanisms, such as self-healing routing or retraining after poisoning. For instance, a resilient intrusion detection model might drop from 95% to 70% accuracy under a PGD adversarial attack but recover to 90% within 200 s—yielding a higher AURC than a static model that stagnates at 75%.
- Scalability and system-level metrics: resilience must extend beyond individual devices to distributed IoT environments where hundreds of clients cooperate through federated or edge learning. Evaluations, therefore, include client scalability, network resilience index, and cross-layer coordination latency. Client scalability varies in performance as the number of participants increases (e.g., from 10 to 150 nodes in federated learning). Network resilience index is measured by the ratio of sustained throughput or model convergence speed under partial connectivity loss (e.g., 20% clients offline). Cross-layer coordination latency measures the time between fault detection at one layer and adaptation at another, indicating the resilience mesh's interdependence. In a federated healthcare network, an adaptive aggregator that maintains over 85% accuracy with a 30% client dropout rate demonstrates better resilience than one that drops below 70%.
- Trust, transparency, and interpretability metrics: because resilience also involves human oversight, operators must trust the system's adaptation process. Interpretability metrics quantify this human–machine alignment using Shapley Additive exPlanations (SHAP) or local interpretable model-agnostic explanations (LIME) attribution stability, trust-score variance, and recovery transparency [122]. SHAP and LIME attribution stability assess the consistency of feature importance across model recoveries, highlighting semantic preservation. Trust-score variance measures fluctuations in model reliability under stress, with lower variance signifying more stable behavior. Recovery transparency is a qualitative or quantitative measure of how well recovery actions are logged, explained, and verifiable (e.g., via blockchain audit trails). For example, a resilient anomaly detector should not only regain performance after retraining but also maintain stable SHAP attributions—ensuring that its reasoning process remains interpretable and trustworthy.
- Hybrid and compound benchmarks: disruptions in real-world IoT environments result from multiple factors. Current assessment frameworks should use compound stress testing by intro-

 doi:10.20944/preprints202512.1717.v1

ducing various stressors—like adversarial perturbations, packet loss, and energy constraints—at the same time. Hybrid benchmarks, such as compound-scenario testing, cross-layer metrics, and resilience trade-off curves, remain scarce in the literature but are essential for realistic validation. Compound scenario testing involves two or more measures, e.g., PGD and 30% packet loss or concept drift and node dropout. Cross-layer metrics combine physical-layer link reliability, network-layer throughput, and model-layer recovery accuracy. Resilience trade-off curves visualize the balance between recovery speed, energy cost, and trust stability across scenarios. For example, an antifragile federated model might slightly reduce accuracy during an attack but significantly shorten recovery time and energy cost across combined network and model stressors.

Assessing the resilience of IoT demands a comprehensive framework that incorporates performance, scalability, interpretability, and the assessment of multiple stressors. In the absence of such multi-dimensional metrics, there is a danger that systems may be deemed resilient based solely on incomplete evidence. Using metrics like AURC for temporal, cross-layer coordination for structure, and trust stability for cognition marks a key advancement in standardizing resilience assessment in IoT research.

4. Taxonomy of IoT Resilience

Resilience in the IoT is the capacity of a system to sustain essential functionality and safety despite facing challenges such as cyberattacks, component failures, and environmental changes [28]. In contrast to conventional reliability that aims to avoid failure, resilience highlights the ability to absorb impacts, adapt to changes, and recover effectively [123]. In a resilient IoT ecosystem, a smart city or healthcare network continues to function—even if sensors fail, communication links degrade, or parts of the system are compromised [124]. Despite these developments, research on IoT resilience remains disconnected, with limited integration across different layers. Individual studies often target isolated layers—such as hardware security, network protocols, machine learning robustness, or governance—without a holistic framework. To address this, we propose a unified two-dimensional taxonomy of IoT resilience, organized along two orthogonal axes: the type of stressor, which represents the nature of the disruption (adversarial, environmental/operational, or hybrid), and layer of the IoT stack: which is the architectural level at which resilience mechanisms operate (hardware, network, learning or AI, application, and governance/trust).

This framework enables researchers to reason systematically about where resilience measures are applied and which types of disturbances they mitigate. IoT systems are exposed to three primary stressor classes, referred to as stressor dimensions: adversarial stressors, environmental or operational stressors, and hybrid stressors. Adversarial stressors are deliberate attacks—such as model poisoning [125], evasion [126], or GPS spoofing [127]—that exploit vulnerabilities to compromise security. Environmental or operational stressors arise naturally from real-world conditions such as interference, sensor drift, and hardware degradation. Hybrid stressors combine both domains, coupling cyber and physical disruptions (e.g., jamming during mechanical faults in a drone swarm). These categories are elaborated with real-world examples and mitigation strategies in the previous Section.

Resilience manifests differently across the layers of the IoT stack, including the hardware layer, network layer, learning layer, application layer, and governance and trust. The hardware layer is tamper-resistant, features redundant sensing, and incorporates PUFs for secure identity verification. The network layer is resilient against routing, featuring decentralized consensus and mechanisms to mitigate denial-of-service attacks. The learning layer incorporates adversarial training, data augmentation, and federated adaptation. The application layer ensures service continuity under degraded conditions through fault-tolerant orchestration. The governance and trust incorporate explainable AI (XAI), blockchain-based auditability, and trust scoring.

By combining these two dimensions, researchers can identify areas of coverage and resilience gaps. For example, adversarial stressors are best addressed through AI-layer defenses, while environmental

 doi:10.20944/preprints202512.1717.v1

stressors often require hardware and protocol-level redundancy. Hybrid stressors demand cross-layer coordination—linking sensing, communication, and learning for joint adaptation.

Figure 12 illustrates how various types of stressors impact distinct layers of the IoT stack. Each row represents a common stressor (e.g., packet loss, poisoning, or concept drift). At the same time, each column corresponds to one of the five resilience layers, ranging from the physical device layer to governance and trust. The color intensity encodes the relative impact of each stressor on that layer, where darker shades indicate higher vulnerability or performance degradation (scaled from 0 to 3). The figure reveals three clear patterns: Network-layer fragility, where packet loss, interference, and DoS attacks register the highest impact due to their ability to disrupt connectivity and throughput; AI-layer sensitivity, where poisoning and evasion attacks dominate, emphasizing that learning-centric IoT components (e.g., anomaly detection or federated models) remain highly exposed to adversarial inputs and data drift; hardware-layer degradation, where energy depletion and physical aging impose consistent operational stress, often preceding higher-layer failures. The governance or trust layer, on the other hand, consistently exhibits low impact, illustrating its role as an oversight mechanism (such as explainable AI and blockchain auditing) that regulates systemic behavior rather than incurring direct failures. Overall, the heatmap underscores the need for resilience strategies that span layers, where recovery methods at one layer (for instance, adaptive routing or continual learning) can alleviate cascading effects across the IoT stack.

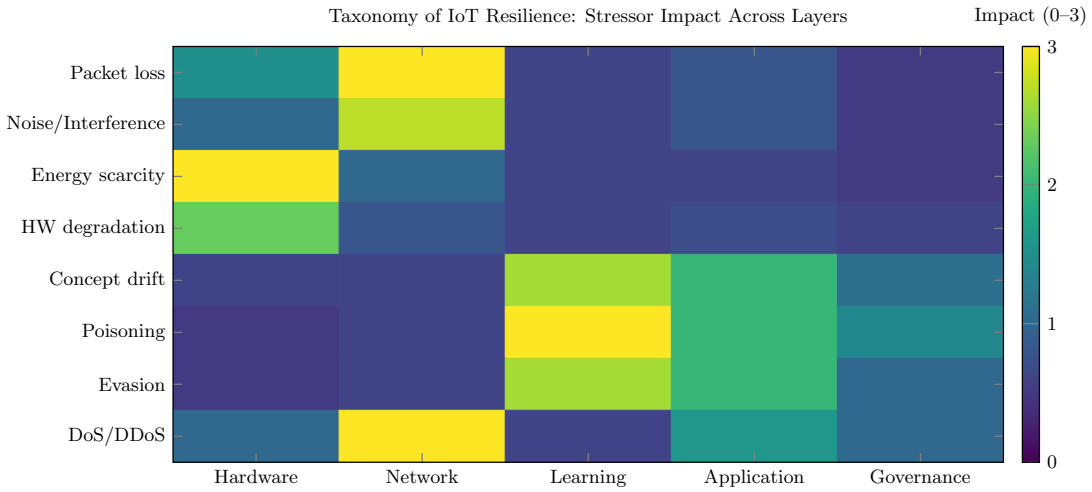


Figure 12. Stressor–Layer Heatmap of IoT Resilience.

Our classification highlights two main insights. The majority of current research emphasizes resilience within a single layer, especially at the AI level, while cross-layer collaboration remains mostly unexamined. Additionally, resilience focused on governance—guaranteeing that adaptive actions are transparent, can be audited, and align with ethical standards—has not been sufficiently developed. Future research should move toward unified, multi-layered frameworks that treat resilience not as isolated protection, but as a dynamic, system-wide property evolving toward antifragility.

5. Adversarial Robustness in IoT Learning and Networking

The integration of advanced learning models and IoT improves automation; however, it exposes systems and environments to adversarial threats. Recent studies concentrate on adversarial robustness via diverse approaches, including deep learning defenses, generative methods, hardware/protocol strategies, federated settings, and mechanisms for trust and explainability. We review the field by grouping research into five categories, summarizing each key paper, and distilling cross-cutting lessons. This subsection reviews deep learning adversarial defence approaches in IoT, and Table 3 compares these techniques.

 doi:10.20944/preprints202512.1717.v1

5.1. Deep Learning Under Adversarial Attack

Deep learning has become the analytical backbone of many IoT applications, from medical diagnostics to industrial automation and wireless communication. However, its data-driven nature makes it highly sensitive to adversarial perturbations—imperceptible input manipulations that can mislead models into incorrect classifications. The fragility of deep architectures under such attacks has motivated the development of defense mechanisms focused on detection, robust training, and architectural adaptation for resource-constrained IoT environments.

Rahman et al. [128] introduce RAD-IoMT, a transformer-based framework for defending against adversarial attacks in the Internet of Medical Things (IoMT). The model consists of two submodels: the first is an adversarial attack detector that screens perturbed inputs, and the second is a transformer-based submodel for disease classification. Tested on more than 100,000 medical images from chest X-ray, retinal OCT, and skin cancer datasets [129], the detector attained an F1 score of 0.91 and an accuracy of 0.94. The disease classification model achieved an F1 score of 0.97 and an accuracy of 0.98. This modular pipeline facilitates explainable, attack-resistant diagnoses, though its computational demands still hinder deployment on low-power medical edge devices.

Güngör et al. [130] initially illustrate that cyber-attacks can severely affect the effectiveness of ML-based predictive maintenance (PDM) techniques, resulting in up to a 120× decrease in prediction performance. Following this, the authors introduce a stacking ensemble learning framework designed to remain robust against a range of white-box adversarial attacks. The findings indicate that their framework demonstrates strong performance even when faced with cyber-attacks and shows up to 60% greater resilience compared to the most robust individual ML method on NASA C-MAPSS and UNIBO Powertools [131] datasets. Despite its strong resilience and generalization across datasets, its inference latency and high memory footprint make it less practical for real-time factory-floor IoT controllers.

Zhang et al. [132] developed a defensive strategy for transformer-based modulation classification systems against adversarial attacks. This paper introduces a vision transformer (ViT) architecture featuring an adversarial indicator (AdvI) token to detect adversarial attacks. This is the first implementation of an AdvI token in ViT for defense. The proposed method merges adversarial training with a detection mechanism in a single neural network. It examines how the AdvI token affects attention weights for identifying unusual input features. Experimental results show that this approach is superior to various techniques in white-box attack scenarios, including basic iterative method, fast gradient method, and projected gradient descent attacks. Nonetheless, it is restricted to white-box scenarios and has not been validated on physical radio hardware or under adaptive attack conditions.

Zyane and Jamiri [133] propose a framework that combines adversarial training with feature squeezing to detect adversarial attacks such as PGD and FGSM. The authors use decision trees, SVMs, and CNNs to evaluate their method on the CICIoT2023 dataset [134]. The proposed method showed promising accuracy, increasing by about 30% against PGD attacks, demonstrating the model's resilience. Their work highlights the potential to improve defenses in real-world IoT applications, making it suitable for resource-constrained devices, although it remains susceptible to adaptive adversaries.

Efatinasab et al. [135] propose a framework for detecting smart grid instability using stable data with a Generative Adversarial Network (GAN). The generator creates Out-Of-Distribution (OOD) samples to illustrate unstable behavior, while the discriminator is trained solely on stable data. This approach allows the system to identify unstable conditions without needing unstable data for training. The framework includes an adversarial training layer for increased attack resilience. It achieves up to 98.1% accuracy in predicting grid stability using the Electrical Grid Stability Simulated Dataset [136] and 98.9% in detecting adversarial attacks, with real-time decision-making capabilities on a single-board computer, averaging response times of under 7 ms. However, the system was validated only in simulated environments and is limited to stability-related metrics.

Javed et al. [137] present a comprehensive analysis of robustness methods in deep learning for medical diagnostics. The study analyzes adversarial training, input preprocessing, uncertainty estima-

 doi:10.20944/preprints202512.1717.v1

tion, and privacy-preserving frameworks (e.g., TensorFlow Privacy and CleverHans [138,139]). The authors highlight that achieving robustness often comes at the expense of accuracy and computational efficiency. Although primarily focused on healthcare AI, the authors highlight lessons broadly applicable to IoT—particularly the need for multi-objective optimization that balances robustness and interpretability. However, the analysis lacks empirical evaluations.

Moghaddam et al. [140] propose a hybrid intrusion detection framework that merges transformer architectures with GAN-based data augmentation and a biogeography-based optimizer. Using CIC-IoT-2023 and TON_IoT [141] datasets, the model outperforms baseline models and achieves 99.67% and 98.84% accuracy, respectively, while handling severe class imbalance. The approach effectively improves detection accuracy in adversarially contaminated data; however, its computational demands pose challenges for real-time edge deployment.

Tian et al. [142] examine how gradient-based adversarial perturbations (e.g., forward-derivative-based adversarial attack (FAA) and random scaling attack (RSA)) corrupt neural network state estimation in power systems and evaluate defenses such as adversarial training and input sanitization. Using standard power-system benchmarks, i.e., the 2012 Global Energy Forecasting Competition dataset [143] (synthetic measurements on IEEE bus test cases), the authors show that small, targeted perturbations can induce significant estimation bias or incorrect topology decisions. In contrast, adversarial training recovers a substantial fraction of lost accuracy with a modest impact on clean-data error. The authors introduced a clear, principled threat model and defense benchmarking; however, the focus is on white-box settings and offline simulations (no embedded/real-time validation).

Tusher et al. [144] examine the vulnerabilities of deep learning and artificial neural network (ANN) models for sky-image-based nowcasting to adversarial attacks, including FGSM and PGD. The authors introduce a feature-extraction-based multi-unit solar (FEMUS)-Nowcast model. The model is integrated with adversarial training to provide resilience against advanced attacks. Under normal conditions using the SKIPP'D dataset [145], the model significantly outperforms existing models (i.e., reducing root mean square error (RMSE) by 48% and mean absolute error (MAE) by 25%) on one hand. On the other hand, under adversarial attacks, the model's accuracy has severely degraded. FGSM increased RMSE by 5–16 times and MAE by 4–12 times. To counteract this, adversarial training is applied to the FEMUS-Nowcast, enhancing its robustness without compromising performance. The adversarially trained model shows resilience against advanced attacks, confirming its reliability across various scenarios. However, there is a limited evaluation against explicit adversarial attacks and uncertain transferability to very different climates/cameras without fine-tuning.

Alsubai et al. [146] build an adversarial evaluation suite using ensemble learning for Internet-of-Medical-Things models—curating data, implementing canonical evasion attacks, and providing a digital twin pipeline for repeatable stress testing. The authors evaluated the impact of adversarial attacks on well-known machine learning models, using the WUSTL-EHMS-2020 dataset [147], demonstrating significant reductions in accuracy. However, adversarial training can partially recover performance. The proposed model achieved a 94% accuracy, outperforming baseline models. However, limitations include limited modality coverage, increased computational overhead for robust defenses, and a need for broader validation across hospitals.

Lessons learned: Current methods for defending IoT deep learning against adversarial attacks mainly concentrate on enhancing robustness in particular applications—such as medical imaging, industrial maintenance, wireless modulation, and intrusion detection. Modular detectors and ensemble architectures improve robustness, while features like squeezing maintain a balance between performance and efficiency. Hybrid frameworks like GRU-Bayesian LSTM and transformer-GAN fusion combine learning methods to address cyber threats and physical challenges. Reviews show that achieving adversarial resilience can conflict with the need for computational efficiency and interpretability, especially in edge-deployed IoT systems.

 doi:10.20944/preprints202512.1717.v1

Table 3. Comparison of Deep Learning Adversarial Defenses in IoT.

Paper	Methodology	Dataset(s)	Main Results	Limitations
Rahman et al. [128]	Transformer-based attack detector and disease classifier	Chest X-ray, Retinal OCT, Skin cancer	F1 = 0.97; Accuracy = 0.98; strong detection recovery	High compute cost; limited real-time feasibility
Güngör et al. [130]	Stacking ensemble (Deep learners and LR/RF/XGBoost)	NASA C-MAPSS, UNIBO Powertools	Up to 60% higher robustness vs. baselines	Increased latency and complexity; not edge-suitable
Zhang et al. [132]	Vision Transformer with adversarial indicator token	RML [148] and RDL [149]	Stronger resilience to FGSM/PGD/BIM; interpretable attention	Tested only under white-box conditions; no hardware validation
Zyane and Jamiri [133]	CNN with adversarial training and feature squeezing	IoT-23 intrusion dataset	Accuracy: 32% → 61% under PGD attack	Static defense; vulnerable to adaptive/black-box attacks
Efatinasab et al. [135]	GAN and OOD samples	Electrical Grid Stability Simulated	Accuracy = 0.981 (stability); robust to GAN-based perturbations (0.989)	Simulation-only; limited scope
Javed et al. [137]	Review: adversarial training, uncertainty, privacy tools	Medical diagnostic DL	Synthesizes robustness metrics; highlights trade-offs	Survey only; lacks empirical evaluation
Moghaddam et al. [140]	Transformer, GAN augmentation, and bio-inspired optimizer	CIC-IoT-2023, TON_IoT	Accuracy: 99.67%, 98.84%; handles imbalance	High computational load; limited edge scalability
Tian et al. [142]	Adversarial attack/defense study for NN state estimation (FAA/RSA); adversarial training and input sanitization	Synthetic power-system measurements (IEEE bus cases)	Small perturbations cause significant estimation bias; adversarial training substantially restores accuracy on attacked inputs with modest clean-data impact	White-box focus; offline simulation only; no embedded/real-time validation
Tusher et al. [144]	FEMUS-Nowcast: feature-enhanced multi-scale U-Net with robustness-oriented augmentation/denoising	SKIPP'D dataset for solar nowcasting	Higher nowcast accuracy than baselines; reduced sensitivity to image artifacts and noisy frames; practical inference latency	No explicit eval vs. strong adversarial attacks; transferability across climates/cameras requires fine-tuning
Alsubai et al. [146]	IoMT adversarial dataset and digital-twin pipeline; benchmarks CNN/Transformer with adversarial attacks and training	Curated IoMT adversarial dataset (classification tasks)	Reproducible stress tests; significant drops under attack; adversarial training recovers part of the loss; provides baselines/tools	Limited modality/task coverage; compute overhead for strong attacks; needs broader cross-institution validation

Recent studies further reinforce these trends. Studies show that minor disruptions in smart grids can greatly affect forecasts made by neural network-based state estimation. It is essential to integrate physics-informed constraints and adversarial retraining in cyber-physical deep learning models to improve accuracy. Ideas like FEMUS-Nowcast enhance resilience through domain-aware augmentation and denoising. Furthermore, digital twin frameworks for the Internet of Medical Things (IoMT) support dynamic trust evaluation and offer adversarial datasets for stress testing and vulnerability assessments.

Future research should focus on cross-layer integration—connecting sensor-level preprocessing, adaptive model calibration, and distributed learning—to develop self-healing, resource-aware IoT architectures that retain explainability in the face of unseen adversarial or hybrid stressors. Incorporating physics-guided constraints, continuous feedback from digital twins, and domain-specific adversarial datasets is key to ensuring that deep learning models in safety-critical IoT ecosystems achieve reliable and sustainable resilience.

5.2. Generative and Ensemble Intrusion Detection

Generative models and ensemble learning have recently emerged as two of the most prominent strategies for enhancing intrusion detection in adversarial IoT environments. Generative adversarial networks (GANs) and hybrid ensembles provide advantages over static classifiers by better modeling complex attack distributions, generating realistic adversarial variations, and improving robustness

 doi:10.20944/preprints202512.1717.v1

against unseen perturbations. This line of research bridges data augmentation, adversarial training, and meta-learning to enable more adaptive and resilient detection frameworks for large-scale, heterogeneous
IoT environments. This subsection reviews GAN-based and ensemble intrusion detection defences in IoT, and Table 4 compares these approaches.

Son et al. [150] introduce a framework for adversarial training to enhance AI model resilience against unexpected adversarial attacks. This framework employs SE-GAN, a self-attention-driven conditional generative adversarial network, to generate adversarial samples, which are then used to train the AI model. Classifiers such as Random Forest, XGBoost, CatBoost, and Extra Trees trained on augmented data (UNSW-NB15 [151], ToN-IoT [152], and power-system [153] datasets) demonstrate substantially higher robustness against previously unseen attack types. The method can generalize beyond the distributions of the original dataset. However, its latency and energy footprint on edge devices remain unreported, limiting practical evaluation.

Khatami et al. [154] proposed an intrusion detection system (IDS) based on a GAN to detect attacks in IoT environments. This system achieves promising accuracy on datasets such as IoT-23, NSL-KDD, and UNSW-NB15, even under adversarial attacks, outperforming traditional single-stage models. To improve performance, the 5-Dimensional Gray Wolf Optimizer (5DGWO) is combined with the GAN-based IDS, resulting in nearly perfect accuracy (around 100%) and significantly lower false-positive rates. However, despite these encouraging findings, the framework has been assessed only offline, leaving its operational feasibility on distributed IoT nodes unexamined.

Alwaisi [155] explores the identification of Mirai botnet attacks in dense IoT environments using a resource-efficient TinyML framework with 6G technology. The study evaluates several lightweight models, including K-Nearest Neighbors, Support Vector Machine, Naïve Bayes, and Random Forest, on Raspberry Pi and Arduino platforms. The KNN model achieves over 99% detection accuracy for various Mirai variants—such as scan, user datagram protocol (UDP) floods, transmission control protocol (TCP) floods, and acknowledgment (ACK) floods—while ensuring minimal memory usage. Despite the proposed approach being robust against attacks and suitable for real-time edge applications, the lack of deep models restricts its ability to adapt to evolving botnets.

Vajrobol et al. [156] propose a training method to detect Mirai attacks using various models: Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), LSTM combined with Random Forest, and LSTM with XGBoost. They tested these models on the CICIoT2023 dataset. The LSTM combined with XGBoost achieved an accuracy of 97.7%. The framework is resilient to attacks during training but has high training and inference costs, which limit its use in low-resource settings.

To defend smart city IoT networks against adversarial and data poisoning attacks, Alajaji [157] introduces FortiNIDS—a robust, GAN-aware Network Intrusion Detection System evaluated on the CICDDoS2019 dataset [158]. FortiNIDS enhances detection accuracy by utilizing models to generate transferable perturbations and integrating adversarial training with Reject on Negative Impact (RONI) filtering. This approach effectively maintains performance even in the presence of significant adversarial threats. The approach significantly reduces false positives compared to baseline NIDS architectures but incurs high computational costs and focuses primarily on DDoS-style traffic.

Omara and Kantarci [159] introduce a detection method based on GANs for Vehicle-to-Microgrid (V2M) IoT systems, where edge services are notably susceptible to subtle adversarial inputs. Conventional models such as SVM, LSTM, density-based spatial clustering of applications with noise (DBSCAN), and an attentive autoencoder (AAE) yield Adversarial Detection Rates (ADRs) below 60% when tested on simulated GAN attacks incorporated into the iHomeLab RAPT dataset [160]. In contrast, the GAN-based detector reaches an impressive 92.5%. This framework demonstrates resilience against adaptive and transfer attacks, though its reliance on synthetic data and lack of real-world edge validation remain limitations.

 doi:10.20944/preprints202512.1717.v1

Morshedi et al. [161] introduce an approach for anomaly detection in IoT network traffic using GANs. The authors evaluate the model using the CICIDS2017 dataset [162]. The model starts with preprocessing steps, including feature scaling, adding Gaussian noise for better generalization, and extracting the Hurst self-similarity parameter to analyze data behavior. The model includes a generator that generates pseudo-real data and a discriminator that distinguishes real from fake data, enabling anomaly detection. The proposed method achieved a high accuracy and recall of 99.88%, outperforming traditional detection techniques. The innovation lies in combining GAN with the Hurst parameter and noise addition, enhancing the model's ability to detect complex and low-frequency attacks while reducing false positives. However, the model has been evaluated offline and no deployment on real IoT or edge hardware.

Lakshminarayana et al. [163] present two data-driven algorithms for detecting and identifying compromised nodes and attack parameters of load-altering attacks (LAAs). The first algorithm uses a Sparse Identification of Nonlinear Dynamics approach to identify attack parameters through sparse regression. The second utilizes a physics-informed neural network (PINN) to infer these parameters from measurements. Both methods are designed for decentralized edge computing architectures. Simulations on IEEE-bus systems show that these algorithms detect and identify attack locations more quickly than existing methods, including unscented Kalman filters and support vector machines. However, the model relies on simulated data and lacks evaluation under adaptive adversaries.

Table 4. Comparison of GAN-Based and Ensemble Intrusion Detection Defenses in IoT.

Paper	Methodology	Dataset(s)	Main Results	Limitations
Son et al. [150]	Self-Attention Conditional GAN with ensemble classifiers	UNSW-NB15, ToN-IoT, Power system	Strong detection of unseen attacks; improved generalization	Latency/energy on edge not analyzed
Khatami et al. [154]	GAN with 5-Dimensional Gray Wolf Optimizer for hyperparameter tuning	NSL-KDD, UNSW-NB15, IoT-23	∼100% detection; reduced false positives	Offline evaluation only; no scalability proof
Alwaisi [155]	TinyML anomaly detector (KNN, SVM, NB, RF) for Mirai variants	Real 6G smart-home and industrial testbeds	KNN >99% accuracy with minimal memory	Excludes deep models; limited adaptability
Vajrobol et al. [156]	Adversarial training with hybrid LSTM with XGBoost	CICIoT2023	Accuracy = 97.7%; robust to adversarial samples	High computational overhead
Alajaji [157]	Surrogate adversarial training + RONI filtering	CICDDoS2019	Recovered accuracy under severe attack; reduced false positives	Compute-intensive; DDoS-focused
Omara and Kantarci [159]	GAN-based detector for adversarial V2M attacks	V2M edge simulations	Attack Detection Rate up to 92.5%; resilient to adaptive attacks	Synthetic data; untested in real-world edge
Morshedi et al. [161]	GAN combined with controlled Gaussian noise for anomaly detection	CICIDS2017	achieved a high accuracy and recall of 99.88%; lower false negatives on unseen attacks	Evaluated offline only; no IoT/edge deployment
Lakshminarayana et al. [163]	sparse regression and PINN for IoT-enabled load-altering attack detection	IEEE-bus smart-grid simulations	3% error rate; precise attack localization; interpretable fusion	Simulation-based; untested under adaptive adversaries

Lessons learned: Generative and ensemble intrusion-detection frameworks substantially enhance IoT resilience by enabling both the generation of synthetic adversarial data and the establishment of multi-model consensus. GANs offer diversity and realism for adversarial training, while meta-ensembles mitigate the overfitting of single models. However, most frameworks remain compute-intensive and lack evaluation on energy-limited or real-time platforms.

Recent studies expand this viewpoint. The GAN and Gaussian-Noise framework shows that stochastic regularization can enhance the stability of adversarial training and boost generalization to zero day exploits. Likewise, semi-supervised ensembles for load-altering-attack detection show that combining data-driven learning with interpretable physical models can simultaneously boost accuracy and explainability in critical-infrastructure IoT. Together, these advances underscore a growing shift

 doi:10.20944/preprints202512.1717.v1

from static, dataset-specific IDS design toward adaptive, self-calibrating, and physically informed generative defenses.

Future research should focus on developing adaptable, lightweight online generative modules and integrating stochastic and physics-aware priors into ensemble learning. It's important to validate these methods in diverse, latency-sensitive IoT environments to ensure their robustness and feasibility.

5.3. Hardware and Protocol-Level Defenses

The lower layers of the IoT stack—hardware, firmware, and communication protocols—act as the first line of defense against cyber and physical threats. Given that IoT devices operate in often untrusted environments, it is essential to ensure the resilience of hardware identifiers, authentication methods, and lightweight cryptographic protocols. Recent research has focused on strengthening these layers against adversarial manipulation, hardware degradation, and synchronization loss. This subsection reviews hardware and protocol-level defences, and Table 5 compares these approaches.

Bao et al. [164] investigate the robustness of convolutional neural network (CNN)-based radio-frequency (RF) fingerprinting. The performed experiments reveal that even minimal adversarial perturbations—crafted via FGSM, basic iterative method (BIM), PGD, or momentum iterative method (MIM) attacks—cause significant drops in classification accuracy, demonstrating that current RF-based identification systems are highly vulnerable to adversarial noise. Although the study identifies critical weaknesses, it proposes no defense strategy, underscoring the urgent need for hardware-aware adversarial training and cross-domain signal validation.

Sánchez et al. [165] investigate the vulnerabilities associated with hardware-based device identification related to context-aware and machine learning attacks. The authors developed a hybrid approach that merges LSTM networks with CNNs. This system achieved a promising F1 score of 0.96 using the LwHBench dataset [166] when identifying 45 Raspberry Pi devices. It showed good resistance to temperature-based attacks but had difficulty with advanced evasion attacks. To improve its performance, the researchers used techniques called adversarial training and model distillation, which raised the score against evasion attacks from 0.88 to 0.17. This study shows how using hardware-level telemetry can greatly improve the resilience of devices. However, its evaluation is restricted to homogeneous testbeds, and performance across diverse platforms and heterogeneous IoT hardware remains unexplored.

Cao et al. [167] introduce S2-Code, a symmetric authentication protocol aimed at ensuring communication integrity even in cases of significant desynchronization and packet loss. The system utilizes dual authentication windows along with the ChaCha20-Poly1305 authenticated encryption algorithm, which has been confirmed through both ProVerif formal verification [168] and practical hardware testing. The method effectively recovers from 50-step desynchronization and tolerates 30% packet loss, achieving latencies of less than 17 ms and about 101 µJ of energy consumption per session. Despite its high efficiency, the protocol relies on pre-shared keys and may face vulnerabilities to side-channel attacks in hostile hardware conditions.

Hemavathy et al. [169] propose a lightweight authentication protocol for hardware security using Arbiter Physical Unclonable Functions (PUFs). This approach enhances device-level trust and effectively counters model-learning attacks on 16-, 32-, and 64-bit field-programmable gate array (FPGA) implementations. The approach reduced attackers' prediction accuracy to about 50%. The design is also power-efficient and ideal for embedded systems. However, large-scale validation under hybrid stress conditions (e.g., combined thermal drift and adversarial probing) is still unexamined.

Aribilola et al. [170] introduce a stream cipher that uses a Möbius S-box, designed for visual IoT data streams. The approach consists of substitution, permutation, XOR, and shift operations capable of creating strong confusion and diffusion. It was tested on IoT video datasets using an Intel NUC testbed. The results show that this method performs better than both Advanced Encryption Standard with cipher feedback (AES-CFB) and ChaCha20 in terms of security and efficiency while still providing secure multimedia encryption. Nevertheless, further cryptanalysis and validation on multimodal IoT workloads (e.g., audio-visual fusion) are required to confirm long-term resilience.

 doi:10.20944/preprints202512.1717.v1

Elhajj et al. [171] extend resilience to the protocol layer by developing a three-layer blockchain framework that combines a Proof of Stake and Practical Byzantine Fault Tolerance to reduce energy consumption and maintain stability, and a K-means clustering algorithm to enable an energy-saving strategy. Evaluated on smart city traffic data (UK-DALE [172] and PECAN Street [173]), their design achieves low latency, energy efficiency (i.e., 80% lower energy use), and resistance to DoS, Sybil, and man-in-the-middle (MITM) attacks. The framework reinforces integrity and non-repudiation in cross-node data exchange but depends on trusted consortium members and remains challenging to scale to thousands of devices.

Alnfiai [174] presents SecureNet-RL, a security orchestration system for 5G IoT that utilizes reinforcement learning. Utilizing a multi-agent RL configuration (including deep Q-networks (DQN) and proximal policy optimization (PPO)), this system adjusts defensive measures like rate limiting and rerouting in real-time based on identified anomalies. In NS3 simulations, SecureNet-RL demonstrates a detection accuracy of 95.8%, a response latency under 50 ms, and a false-positive rate of 4.3%, significantly surpassing traditional IDS methods. Although the proposed method shows scalability in simulations, however, there are challenges to address for actual deployment, such as adversarial reward manipulation and the intricacies of policy synchronization across extensive 5G slices.

Dong et al. [175] present a hierarchical optimization framework designed to strategically position cyber-decoy sensors within IoT networks of the power grid, aimed at misleading adversaries and enhancing defense against intrusions. A multi-objective solver is employed to simultaneously minimize operational costs and misdirect the efforts of attackers through optimal resource allocation and adaptive decoy placement. The framework incorporates various methods, including multi-agent deep reinforcement learning (MADRL), distributionally robust optimization, and Bayesian inference, to tackle the continuously evolving patterns of attacks and uncertainties. Testing on the IEEE 123-bus system and utilizing real phasor measurement units (PMUs) data, the approach enhances resilience metrics by 35% and decreases the success rates of attacks by 40%. The framework provides security and integrity against cyber physical threats. However, the proposed method has two limitations: high computational cost during re-optimization and a lack of hardware-in-the-loop validation.

Table 5. Comparison of Hardware and Protocol-Level Resilience Mechanisms in IoT Systems

Paper	Methodology	Dataset/Testbed	Key Strengths	Limitations
Bao et al. [164]	CNN-based RF fingerprinting for device ID	RF signal traces	Reveals brittleness of RF-only identifiers under FGSM/PGD	No defense proposed
Sánchez et al. [165]	LSTM-CNN with adversarial distillation	Raspberry Pi 4 cluster	F1 = 0.96; attack success reduced 0.88→0.17	Limited to homogeneous hardware; needs wider validation
Cao et al. [167]	Dual-window symmetric authentication (ChaCha20-Poly1305 AEAD)	Hardware, Mininet, ProVerif simulation	50-step desync recovery; <17 ms latency; 101 µJ/session	Relies on pre-shared keys; potential side-channel exposure
Hemavathy et al. [169]	Lightweight authentication protocol to counter model learning attacks on FPGA PUFs	FPGA (16-, 32-, and 64-bit)	ML attacker success reduced to 50%; lightweight	Not validated under hybrid stress or large-scale deployment
Aribilola et al. [170]	Möbius S-box stream cipher	IoT video data on NUC	More secure and efficent than AES-CFB and ChaCha20	Requires deeper cryptanalysis and multimodal testing
Elhajj et al. [171]	Three-layer blockchain for IoT data integrity and access control	UK-DALE and PECAN Street	Energy-efficient; resilient to DoS, Sybil, and MITM; low latency	Requires trusted consortium; scalability limits
Alnfiai [174]	Multi-agent RL for dynamic 5G defense orchestration	NS3-based 5G simulation	95.8% detection; <50 ms mitigation; adaptive learning	Reward poisoning risk; deployment complexity
Dong et al. [175]	Multi-layered optimization for adaptive cyber-decoy placement; MADRL and Bayesian inference for evolving attakcs and uncertanties	IEEE 123-bus grid, real PMU traces	35% better resilience metrics; 40% success rates of attacks decrease	High computation during re-optimization; no hardware-in-loop validation

 doi:10.20944/preprints202512.1717.v1

Lessons learned: To build effective IoT systems for challenging environments, it is essential to concentrate on hardware-aware learning and protocol innovations. Future research shoud focus on using physical identifiers, dynamic authentication, and efficient encryption strengthens the trust around IoT devices. However, most evaluations happen only in controlled lab settings, lacking the real-world diversity of device architectures and environmental stresses.

Recent contributions further expand this foundation. Multi-layered optimization for adaptive decoy positioning shows that dynamic, deception-oriented defenses at the protocol layer can significantly decrease detection delays and improve situational awareness across the grid without incurring high bandwidth costs. Adversarial testing shows that data-driven control systems in power networks are vulnerable. It is crucial to combine adversarial retraining with communication-level sanitization to address these weaknesses.Together, these findings point toward a new generation of hybrid hardware–protocol defenses that merge physical security primitives with intelligent, optimization-driven resilience mechanisms.

Future research should focus on large-scale hardware-in-the-loop testing for IoT platforms. It should incorporate decoy orchestration, adaptive cryptography, and robust signal estimation to strengthen resilience against cyber-physical threats.

5.4. Federated and Distributed Resilience

Federated learning (FL) and distributed training paradigms have become central to privacy-preserving intelligence in IoT systems, allowing edge devices to collaboratively train global models without sharing raw data. However, the decentralized nature of FL introduces new resilience challenges—ranging from model poisoning and Byzantine behavior to communication failures and device heterogeneity. Recent studies have therefore focused on both identifying vulnerabilities and developing robust aggregation mechanisms that preserve accuracy under adverse or adversarial conditions. This subsection reviews federated and distributed resilience frameworks, and Table 6 compares them.

Reis [176] introduce Edge-FLGuard, an anomaly detection framework based on federated learning and edge AI designed for real-time security in IoT environments supported by 5G. This framework employs lightweight deep learning architectures—specifically autoencoders and LSTM networks—for inference on devices, along with a privacy-preserving federated training approach to facilitate scalable and decentralized threat detection without the need for sharing raw data. The authors assess the proposed method using two public datasets (i.e., CICIDS2017, TON_IoT) and synthetic datasets in various attack situations, such as spoofing, DDoS, and unauthorized access. The experimental findings indicate good detection performance (an F1-score of more than 0.91 and AUC-ROC of more than 0.96), and minimal inference latency of less than 20 ms, and resilience against data variability and adversarial scenarios. By merging edge intelligence with secure and collaborative learning, Edge-FLGuard offers a viable and scalable cybersecurity option for future IoT implementations. Although the research illustrates the vulnerability of small-scale federated learning networks to local breaches, it does not include any defensive measures, highlighting the necessity for adaptive aggregation and anomaly filtering on devices.

Albanbay et al. [177] conduct a large-scale study to determine the optimal deep learning model and data volume for IDS on resource-limited IoT devices using FL. The study shifts the focus from accuracy to addressing the computational constraints of IoT hardware. It evaluates three deep learning architectures —DNN, CNN, and hybrid CNN combined with BiLSTM —on the CICIoT2023 dataset in a simulated environment with up to 150 IoT devices. The assessment includes detection accuracy, convergence speed, and inference costs. The CNN demonstrates an accuracy of approximately 98% and maintains low latency. The CNN-BiLSTM model reaches about 99% accuracy but has a higher computational cost. Testing on Raspberry Pi 5 devices shows both models can be effectively used on IoT edge hardware. However, the experiments were conducted in a controlled offline setting with a static IoT dataset, which doesn't reflect real deployment challenges. The current setup assumes reliable devices and does not account for adversarial scenarios like model poisoning.

 doi:10.20944/preprints202512.1717.v1

Shabbir et al. [178] studied two types of federated learning frameworks—centralized (CFL) and decentralized (DFL)—to forecast smart grid loads. They used a three-layer artificial neural network (ANN) with three sub-datasets: APE_hourly, PJME_hourly, and COMED_hourly (i.e., part of Hourly Energy Consumption dataset [179]). They found that during poisoning attacks, DFL setups, including line, ring, and bus topologies, had mean absolute percentage errors (MAPEs) of less than 0.5%, 4.5%, and 1%, respectively. In contrast, CFL models showed much higher errors of over 6%, 18%, and 10%. This shows that decentralized systems are better at handling attacks from harmful or faulty sources. However, the evaluation remains simulation-based, lacking real hardware or communication noise.

Haghbin et al. [180] introduce Auxiliary Federated Adversarial Learning (AuxiFed) as a solution to challenges in federated learning. It uses pre-trained auxiliary-classifier GANs (AC-GANs) and probabilistic logic to create diverse synthetic data, improving model resilience and accuracy while protecting against adversarial attacks. AuxiFed improves training effectiveness by leveraging both real and synthetic data. The authors evaluate the method on the MNIST [181] and EMNIST [182] datasets and show that AuxiFed outperforms baseline algorithms such as Federated Averaging (FedAvg) and its variants (FedAvg combined with variational autoencoder (VAE) and FedAvg combined with C-GAN) across all metrics in both homogeneous and heterogeneous environments. Variants such as AuxiFed-PGD and AuxiFed-FGSM also show strong performance. Overall, AuxiFed improves model resilience against adversarial attacks and enhances generalization to unseen data. Despite its strong defense against data heterogeneity and model poisoning, the scalability and computational cost of its IoT-scale deployments remain untested.

Al Dalaien et al. [183] introduce a dual-aggregation technique to improve security in federated learning. The proposed method uses existing machine learning techniques without requiring more computing power. It involves ensemble learning, where each client model first makes predictions using random forests and gradient boosting. Then, the projections from all clients are combined to create a complete global model. Experimental results using the CICIoT-2023 dataset show that the proposed technique achieves 91% accuracy, underscoring its robustness against model poisoning attacks. The proposed approach provides a lightweight, resilient framework for securing IoT systems against adversarial threats. However, its resilience against stealthy backdoor attacks has yet to be verified, leaving potential gaps in trustworthiness under adaptive adversarial conditions.

Mukisa et al. [184] discuss an intrusion detection system (IDS) for edge-IoT networks that combines blockchain technology with federated learning (FL) using a customized aggregation strategy, adaptive trimmed mean aggregation (ATMA). The system uses a permissioned blockchain for client authentication and secure model storage, allowing only verified participants to train the model. The ATMA strategy adjusts its trimming parameter based on client update variance, improving resistance to Byzantine faults while maintaining $O(n \log n)$ complexity. Tested against label-flipping and Gaussian-noise attacks with various adversarial rates on both independent and identically distributed (IID) (91.8% accuracy) and non-IID data (88.4% accuracy) using CICIoV2024 [185], Edge-IIoTset [186], and ForgeIIOT Pro [187] datasets, the system demonstrated strong detection performance (outperforming the standard FedAvg and Krum algorithms) and minimal overhead, making it a scalable and secure solution for Edge-IoT security. While conceptually elegant, the method introduces additional computational complexity and may face scalability challenges in large or rapidly fluctuating IoT networks.

Vinita [188] extends federated resilience beyond security to fairness and incentive stability through the Incentive-Aware Federated Bargaining (IAFB) framework. Leveraging Nash bargaining, Shapley-value-based incentives, and advanced encryption standard-galois/counter mode (AES-GCM) encryption for secure aggregation, IAFB ensures both fair participation and robust aggregation in IoT smart homes. On CASA Smart Home and MNIST datasets, it improves accuracy by 6.5%, fairness by 28%, and reduces communication overhead by nearly 40%. While scalable in controlled environments, incentive computation may become burdensome in large heterogeneous systems.

 doi:10.20944/preprints202512.1717.v1

Prasad et al. [189] introduce a Two-Tier Optimization Strategy for Robust Adversarial Attack Mitigation (TTOS-RAAM) to enhance IoT network security. The technique aims to detect adversarial attack behaviors by first normalizing input data using a min-max scaler. It employs a hybrid coati-grey wolf optimization (CGWO) for optimal feature selection and utilizes a conditional variational autoencoder (CVAE) for attack detection, with parameter adjustments made through an improved chaos African vulture optimization (ICAVO) algorithm. Extensive experimental analyses on the RT-IoT2022 dataset show that TTOS-RAAM achieves a remarkable accuracy of 99.91%, surpassing other existing methods. The proposed approach is accurate. However, the multi-stage property might increase training overhead, and the model requires high-quality labeled data.

ALFahad et al. [190] propose sequential learning-based algorithms using multi-armed bandit (MAB) systems to tackle the node selection problem. They introduce novel MAB algorithms for node selection that leverage deep learning expert models. To address the natural uncertainty associated with nodes, we propose ExpGradBand, a new expert-based gradient MAB algorithm that takes advantage of the selection efficiency offered by gradient bandits utilizing historical contextual data. They evaluate and compare ExpGradBand with various Multi-Armed Bandit (MAB) methods and benchmarks, both with and without contextual information. Nonetheless, the requirement for ongoing feedback and the computational demands might limit its use on ultra-low-power devices.

Table 6. Comparison of Federated and Distributed Resilience Frameworks in IoT Systems.

Paper	Methodology	Dataset/Testbed	Key Results	Limitations
Reis [176]	Edge FL testbed on Jetson Nano and Raspberry Pi	CICIDS2017, TON_IoT	F1-score (>91%); AUC-ROC (>96%); Latency (<20 ms)	No defenses; small-scale
Albanbay et al. [177]	DNN, CNN, CNN-BiLSTM profiling on edge nodes	CICIoT2023, Raspberry Pi 5	CNN: 98% accuracy; CNN-BiLSTM: 99% accuracy	controlled offline settings; reliable devices assumption
Shabbir et al. [178]	DFL vs CFL for smart grid forecasting	Hourly Energy Consumption	DFL MAPE < 0.5%, 4.5%, 1%; CFL MAPE >6%, 18%, 10% under poisoning	Simulation-only; no real-world noise
Haghbin et al. [180]	GAN-augmented FL with AC-GAN synthesis	MNIST, EMNIST	Best convergence under PGD/FGSM; higher robustness	Not IoT-scale validated; compute cost
Al Dalaien et al. [183]	Client ensemble (RF/GBM) + weighted server aggregation	CICIoT-2023	∼91% accuracy; moderate cost	Backdoor resistance untested
Wang et al. [184]	Adaptive Trimmed Mean aggregation	CICIoV2024, Edge-IIoTset, and ForgeIIOT Pro	>91% accuracy; better than FedAvg, Krum	Added complexity; scalability untested
Vinita [188]	Nash bargaining, Shapley-value incentives, AES-GCM	CASA Smart Home, MNIST	+6.5% accuracy, +28% fairness, -39.5% communication	Complex incentive computation; large-scale burden
Prasad et al. [189]	Two-tier hybrid optimization (Coati–GWO with CVAE)	RT-IoT2022	99.91% accuracy; robust adversarial mitigation	Multi-stage training cost; label dependency
ALFahad et al. [190]	Contextual multi-armed bandit node selection	Simulated edge environment	Adaptive, resilient to gradient noise; faster convergence	Requires feedback loop; limited low-power feasibility

Lessons learned: Federated and distributed learning architectures demonstrate strong potential for resilient intelligence in IoT networks by limiting data exposure and enabling fault-tolerant collaboration. Decentralized topologies inherently mitigate single-point failure and poisoning risks, while robust aggregation (e.g., GAN-based synthesis, adaptive trimming) strengthens model integrity. Nonetheless, real-world constraints—including communication delays, energy efficiency, backdoor resilience, and large-scale heterogeneity—remain underexplored. Future research should focus on cross-layer evaluation, on-device defensive adaptation, and standardized testbeds to benchmark the robustness of FL across diverse IoT scenarios.

Newer developments such as FEMUS-Nowcast extend these findings by demonstrating that federated deep architectures can sustain high accuracy and stability even under noisy communication

 doi:10.20944/preprints202512.1717.v1

and client dropout—conditions common in edge energy and environmental IoT. Such multi-sensor, site-adaptive models highlight the feasibility of federated inference that balances local autonomy with global coordination. Collectively, these advances signal a shift toward distributed IoT ecosystems that not only safeguard data privacy but also ensure operational continuity through redundancy, adaptive aggregation, and environment-aware optimization across resource-heterogeneous nodes.

5.5. Trust, Explainability, and Governance

As IoT systems become more independent, it is important to ensure that their decision-making is transparent, easy to understand, and reliable. This will help make sure they can be safely and socially accepted. While traditional intrusion detection and anomaly detection systems are generally accurate, they often function as black boxes, providing minimal interpretability for operators or regulators. As a result, recent studies have highlighted explainable AI (XAI), trust-aware controllers, and blockchain-based governance frameworks as essential components for building resilient and trustworthy IoT infrastructures. This subsection discusses trust, explainability, and governance approaches, and Table 7 compares them.

Recent pipelines proposed by Nugraha et al. [191] have integrated model-agnostic XAI methods such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) with analysis of variance (ANOVA) [192] to analyze feature contributions in network intrusion detection models and to select and reduce feature complexity without compromising detection performance. Using the CICDDoS2019, CICIoT2023, and 5G-PFCP [193] datasets, experiments show that eXtreme Gradient Boosting (XGB) achieves the highest F1-score of more than 99%. At the same time, the feature selection method has reduced the feature dimensionality by 70% to only 10 essential features. This interpretability exposes latent model biases and potential attack surfaces. However, such methods still struggle to generalize across multimodal IoT contexts—particularly when fusing traffic, sensor, and user-behavior data.

Naik et al. [121] introduced a hybrid approach that combines CNNs, LSTMs, and variational autoencoders (VAEs) in the healthcare sector, dynamically adjusting device permissions based on metrics such as anomaly rate, contextual entropy, and model confidence. These controllers implement real-time access control and isolation measures to avert data leakage and device impersonation. When evaluated on biomedical datasets such as MIMIC-III [194] and MIT-BIH Arrhythmia [195], the system achieves an average F1-score of 94.3% for anomaly detection while maintaining inference latency below 160 ms—satisfying edge deployment standards. Despite strong real-time performance, maintaining calibrated trust scores and adapting to non-stationary or adversarial data remain significant research challenges.

Majumdar and Awasthi [196] developed a secure system for tracking and logging events in IoT applications using blockchain technology and AI for anomaly detection. By leveraging Hyperledger Fabric, the system provides reliable GPS tracking and event logging essential for public safety, such as wildfire management, with a latency of around 250 milliseconds and a false-positive rate below 5% in simulated emergencies. While this approach enhances accountability and traceability, issues related to scalability for national-level IoT networks and compliance with data protection regulations remain.

He et al. [197] introduce NIDS-Vis, a new black-box algorithm for exploring decision boundaries in deep neural network-based network intrusion detection system (NIDS). This framework visualizes the decision boundaries and evaluates their effects on performance and adversarial robustness. The conducted experiments on the UQ-IoT dataset [198] reveal a trade-off between performance and robustness, and the authors propose two innovative training methods —feature space partition and a distributional loss function —to improve the generalized adversarial robustness of DNN-based NIDSes while maintaining performance levels. Yet, its effectiveness diminishes in high-dimensional or highly heterogeneous IoT networks, where visual interpretability and model convergence become increasingly complex.

Al-Fawa'reh et al. [199] introduce a semi-supervised learning approach, entropy-driven latent transformation (EDLS), designed to detect adversarial attacks and out-of-distribution (OOD) instances.

 doi:10.20944/preprints202512.1717.v1

This innovative method leverages variational autoencoders (VAEs) and normalizing flows to identify anomalous samples within the latent space. The framework's effectiveness was assessed using the KDD99 and X-IIOTID [200] datasets, yielding promising accuracy and F1 scores. The proposed method demonstrates superior performance compared to prior techniques in the context of black-box attacks. However, it is important to note that the method requires substantial computational resources, which may limit its practicality for real-time edge deployment without further optimization.

Alasmari et al. [201] improve the explainability of IoT security by integrating CNN-LSTM models with DistilBERT and SHAP-based explanations for detecting phishing and malicious URLs. The method was tested on extensive email and web datasets [202–205] (i.e., having more than 1.6 million records). The proposed method achieves an outstanding 98.64% accuracy and increases transparency by 41% compared to conventional methods. The method's explainable results enable analysts to understand the reasoning behind decisions, thereby building user trust. Nonetheless, a significant challenge remains in upholding interpretability as phishing tactics continue to evolve.

Jin and Lee [30] advocate for an antifragile approach to AI safety, emphasizing the need for systems to adapt to rare and out-of-distribution (OOD) events. They highlight the limitations of traditional benchmarks, including insufficient scenario coverage and the risk of reward hacking. They suggest using uncertainties to get ready for future challenges and call for a new way to measure AI safety. Their goal is to enhance current strategies for robustness by providing ethical and practical guidelines to support the AI safety community. Although conceptual, their work frames a forward-looking governance paradigm that complements explainable and blockchain-based resilience research by emphasizing learning-driven adaptation.

Table 7. Comparison of Trust, Explainability, and Governance Approaches in IoT Security.

Paper	Methodology	Dataset/Testbed	Key Results	Limitations
Nugraha et al. [191]	Model-agnostic XAI, LIME and ANOVA feature analysis	CICDDoS2019, CICIoT2023, and 5G-PFCP	XGB best accuracy; efficent feature importance method	Limited transfer to multimodal IoT
Naik et al. [121]	Context entropy + anomaly/confidence scoring	MIMIC-III, MIT-BIH	F1 = 94.3%; latency <160 ms	Calibration and adaptation remain difficult
Majumdar and Awasthi [196]	AI anomaly detection and Hyperledger Fabric ledger	Simulated emergency IoT	Tamper-proof logs; ∼250 ms latency; fewer false positives	Scalability and regulation gaps
He et al. [197]	Visualization-driven black-box tuning	UQ-IoT	Improved robustness and interpretability	Degrades on high-dimensional data
Al-Fawa'reh et al. [199])	Semi-supervised latent entropy transformation (VAE and flows)	KDD99, X-IIoTID	High F1 under black-box attacks; OOD detection	Computationally expensive; non-real-time
Alasmari et al. [201]	CNN-LSTM, DistilBERT, SHAP for phishing detection	Web and Email datasets (>1.6 M samples)	Accuracy = 98.64%; 41% better transparency	Limited adaptation to new phishing strategies
Jin and Lee [30]	Antifragile governance model (conceptual)	Theoretical / literature synthesis	Advocates learning-based resilience and policy evolution	No empirical validation; conceptual only

Lessons learned: Trust, explainability, and governance are vital for resilient IoT ecosystems. XAI tools enhance transparency, trust-aware controllers boost resilience, and blockchain frameworks ensure accountability in distributed networks. However, significant challenges persist: robust calibration of trust metrics, scalability of blockchain infrastructures, interpretability in multimodal and dynamic IoT settings, and computational efficiency of explainable and latent-space defenses.

Recent advances add an important governance dimension. Digital twin frameworks offer continuous validation loops for secure testing and the creation of adversarial datasets, helping to future-proof

 doi:10.20944/preprints202512.1717.v1

AI models. They allow IoT systems to detect drift, retrain effectively, and comply with evolving ethical and regulatory standards. This convergence of explainability, auditability, and lifecycle governance points toward self-regulating IoT ecosystems in which transparency and resilience evolve jointly.

Future studies should focus on creating federated networks of reliable digital twins, along with ensuring privacy-preserving synchronization among institutions. Additionally, it is important to develop adaptive explainability metrics that can evolve with changing IoT data streams to ensure sustainable resilience that is both verifiable and comprehensible to humans.

6. Case Study: Adversarial Robustness on ToN-IoT

To demonstrate the practical implications of adversarial resilience in IoT learning pipelines, a controlled case study was conducted using the ToN-IoT dataset—one of the most comprehensive benchmarks for evaluating intrusion detection and network analytics in smart environments. The aim was to evaluate how subtle input variations affect classification integrity and to see if lightweight adversarial training can enhance robustness without sacrificing accuracy.

6.1. Dataset and Task

The ToN-IoT network-flow subset comprises 211,043 traffic records across 44 attributes, including transport- and application-layer indicators such as duration and connection state, as well as various HTTP response fields. Each instance is labeled as either benign or malicious, producing a binary classification problem with a moderately imbalanced distribution of $\{0 : 50{,}000,\ 1 : 161{,}043\}$. We standardized continuous features and one-hot encoded categorical attributes to ensure stable gradient propagation during training, effectively capturing the variability in IoT traffic across devices and protocols.

6.2. Model and Training Protocol

A lightweight feed-forward neural network was designed to predict flow-level security status, featuring two hidden layers with ReLU activations and dropout regularization, optimized using the Adam algorithm on stratified datasets. While such architectures are computationally efficient and thus appealing for edge deployment, their gradient sensitivity often makes them susceptible to adversarial manipulations. To evaluate this vulnerability, we generated adversarial samples using the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks under ℓ_∞ constraints with perturbation budgets $\epsilon \in \{0.00, 0.01, 0.02, 0.05, 0.10, 0.15\}$. We conducted single-step adversarial training (AT) using FGSM with $\epsilon = 0.10$ to enhance the robustness of a standard IoT classifier, highlighting the impact of adversarial regularization.

6.3. Baseline Performance

In a stable (unperturbed) environment, the baseline model achieved nearly perfect detection performance, achieving a test accuracy of 0.9999. Macro-averaged values for precision, recall, and F1-score were all greater than 0.9998, suggesting excellent generalization with very few false positives or negatives. Although these metrics may initially imply an almost perfect classifier, they obscure the sensitivity of the established decision boundaries—a weakness that becomes evident when adversarial stress testing is applied.

6.4. Adversarial Stress Testing

The test accuracy decline under various perturbation budgets for FGSM and PGD attacks is shown in Table 8. The baseline model maintained strong accuracy for minor perturbations (up to $\epsilon = 0.05$) but showed a significant drop in robustness beyond $\epsilon = 0.10$. For FGSM, accuracy dropped sharply from 0.9995 to 0.9450 at $\epsilon = 0.15$, indicating a 5.5% reduction from the clean baseline despite perturbations being visually or statistically minimal. PGD attacks, which are iterative and more potent, caused somewhat smaller yet still detectable decreases in accuracy, demonstrating that even classifiers with high confidence can be affected by structured adversarial noise. These results empirically support

 doi:10.20944/preprints202512.1717.v1

the theoretical assertion that conventional deep IoT models focus more on discriminative performance than on resilience.

Table 8. Accuracy under adversarial perturbations. Values represent test accuracy at different ℓ_∞ budgets ϵ.

			FGSM Accuracy			
Model	0.00	0.01	0.02	0.05	0.10	0.15
Baseline	0.9999	0.9999	0.9999	0.9995	0.9941	0.9450
Robust	0.9999	0.9999	0.9999	0.9999	0.9999	0.9986
			PGD Accuracy			
Model	0.00	0.01	0.02	0.05	0.10	0.15
Baseline	0.9997	0.9997	0.9997	0.9994	0.9977	0.9939
Robust	0.9997	0.9997	0.9997	0.9997	0.9997	0.9994

6.5. Adversarial Training and Robustness Gains

Adversarial training with FGSM perturbations at epsilon 0.10 significantly enhanced resilience with minimal impact on clean accuracy (0.9999 to 0.99994). With a stronger attack at epsilon 0.15, FGSM robustness rose from 0.9450 to 0.9986, and PGD performance improved from 0.9939 to 0.9994. This demonstrates that even a single-step adversarial training strategy can approximate the robustness levels achieved by more computationally expensive multi-step methods. The improvement shows how well simple gradient-based regularization works. It strengthens feature boundaries and helps the model focus on stable features instead of being misled by changes in the gradient.

Figure 13 shows the accuracy as a function of perturbation magnitude for FGSM and PGD attacks. The baseline curves reveal a notable drop in accuracy with increasing epsilon, especially for FGSM. In contrast, the adversarially trained model displays nearly flat curves, indicating a smoother decision boundary that is less affected by adversarial gradients. This indicates that training with targeted adversarial examples helps the network generalize beyond the clean data, making it more resilient to gradient-based attacks.

This experiment highlights key insights in IoT resilience research: relying solely on performance metrics from clean data can lead to misleading conclusions about real-world reliability. Models that do well in tests can still fail with minor interruptions. Lightweight adversarial training can strengthen their reliability at a low cost, making it a good choice for IoT devices with limited resources. It is also important to see robustness as a range, which emphasizes the need for adversarial evaluations to confirm how well IoT models work. This case study shows how regularization focused on resilience can change high-performing but vulnerable IoT models into dependable systems for edge deployment, connecting theoretical strength with practical reliability.

 doi:10.20944/preprints202512.1717.v1

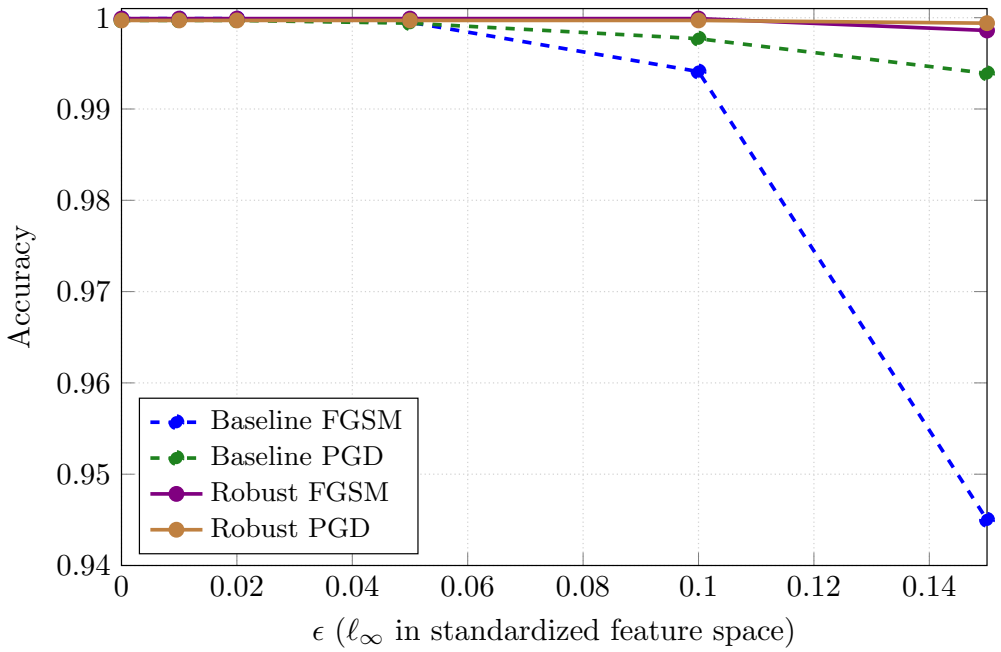


Figure 13. Accuracy versus perturbation budget ϵ for FGSM and PGD on the ToN-IoT dataset. Adversarial training with $\epsilon = 0.10$ preserves clean accuracy while markedly improving resilience under stronger perturbations.

7. Challenges and Future Directions

This survey shows a major transformation in IoT security research. Researchers are shifting the focus from specific defenses, like classification and secure routing. Instead, they are developing complete and flexible frameworks that combine learning, communication, and governance. Key improvements in IoT security include adversarial training, federated learning, hardware-level authentication, and explainable governance. However, building efficient, adaptable, and secure IoT systems remains a challenge. This section points out important gaps from current research and suggests future research directions for creating reliable, self-repairing, and trustworthy IoT systems. Figure 14 illustrates how IoT resilience has evolved from basic defenses to fully adaptive and explainable systems. Between 2022 and 2025, advancements have focused on hybrid and cross-layer approaches that integrate adversarial learning, explainability, and federated robustness. The next phase (2026–2030) is projected to focus on antifragile architectures that can self-heal, adapt to context, and enable real-time governance through decentralized trust mechanisms. This trajectory highlights a gradual merging of technical resilience and organizational accountability, positioning IoT systems as intelligent, transparent, and autonomously secure infrastructures.

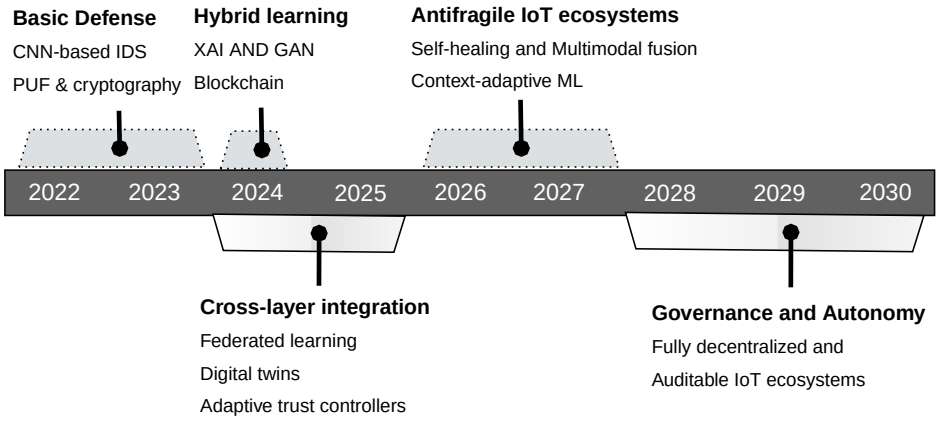


Figure 14. Timeline of IoT resilience research evolution and future forecast.

 doi:10.20944/preprints202512.1717.v1

7.1. Cross-Layer and Hybrid-Stressor Integration

Many studies have a common limitation: they focus on only one layer or type of threat. Most defenses are evaluated at the model level, such as adversarial robustness in CNNs and transformers, or within communication protocols, often ignoring the cascading interactions in real-world deployments. In practice, IoT disruptions are rarely confined to a single domain: packet loss often co-occurs with adversarial perturbations, and data drift typically accompanies sensor degradation. Future research should therefore develop cross-layer resilience frameworks that fuse hardware-level telemetry, network metrics, and learning confidence into a unified decision model. Additionally, it should construct hybrid-stressor benchmarks (e.g., PGD and 30% packet loss, or GAN-based poisoning under jamming) to quantify recovery dynamics under compound disturbances. Furthermore, future work could investigate multi-objective optimization approaches that balance latency, energy, and resilience metrics, ensuring that robustness improvements do not compromise real-time operation.

7.2. Scalability and Real-World Benchmarking

Despite impressive reported accuracies—often exceeding 95%—most IoT resilience frameworks remain confined to simulation or small-scale testbeds. Few studies test systems in heterogeneous, high-density environments where thermal drift, non-stationary data, and real wireless interference co-exist. Moreover, there is no standardized evaluation protocol analogous to ImageNet or GLUE in AI resilience research. Future efforts must establish open-source, cross-layer IoT resilience testbeds that incorporate heterogeneous hardware (e.g., Raspberry Pi, FPGA, ESP32) and diverse connectivity options (LoRa, Wi-Fi, 5G). Future research could design unified evaluation metrics—such as Area Under the Resilience Curve (AURC), time-to-recovery, and robustness-energy trade-off indices—to capture dynamic recovery behavior rather than static accuracy. Future research could promote real-time, longitudinal evaluation, including stress endurance testing to observe how systems degrade or adapt over weeks of operation.

7.3. Lightweight and Energy-Aware Robustness

Edge and microcontroller-class IoT devices are typically very limited in resources. Although advanced defenses like transformers or GAN-based augmentation are promising, they frequently come with high computational and energy requirements. Achieving hardware-efficient resilience remains a significant unresolved challenge. Upcoming research should concentrate on creating adversarial defenses suitable for TinyML that can condense robustness-enhancing components (such as adversarial training and latent detectors) so they fit within memory constraints of 100–200 kB. Future studies can leverage neuromorphic and analog circuitry to integrate robustness features directly into silicon, thereby facilitating physical resilience through mechanisms such as adaptive voltage, timing, or stochastic computing. Further research should create models for energy-resilience trade-offs to adjust defense complexity based on available power or network circumstances.

7.4. Federated and Decentralized Resilience

While federated and distributed learning reduce data exposure risks, they bring about new failure scenarios such as model poisoning, backdoor insertion, and fairness issues. Most existing defenses (like trimmed mean and GAN-based aggregation) operate under the assumption of stable participation and benign communication, conditions that are rarely met in real-world IoT environments. Future research should develop adaptive aggregation methods to filter out harmful updates in bandwidth-constrained environments. It should also integrate trust and reputation frameworks to improve client contributions and explore incentive-compatible federated frameworks that combine fairness with privacy measures. Lastly, efforts should focus on creating multi-agent reinforcement learning protocols that enhance collaborative resilience against distributed threats.

 doi:10.20944/preprints202512.1717.v1

7.5. Explainability, Governance, and Human Trust

As IoT systems take on more safety-critical functions—like healthcare monitoring, grid management, and autonomous driving—technical robustness alone is inadequate. Stakeholders need to be able to comprehend, assess, and verify the behavior of the system. Nonetheless, the challenge of incorporating interpretability and accountability while maintaining performance is significant. Key areas of focus include developing explainability-by-design models that generate interpretable intermediate representations (such as attention maps and trust scores) during training instead of depending on post-hoc evaluations. Future research could combine blockchain governance with explainable AI to generate verifiable audit trails, thereby aligning the transparency of machine learning with legal responsibility. This would involve creating trust-calibrated control loops that automatically adjust system autonomy based on operator confidence or contextual uncertainty. Investigating antifragility as a governance paradigm, where stress events trigger adaptive policy refinement instead of static recovery—transforming failures into structured learning opportunities.

7.6. Standardization, Ethics, and Societal Readiness

IoT resilience must align with ethical and legal standards. The absence of standardized privacy-resilience trade-offs and transparency in AI-driven security decisions compromises regulatory compliance and user acceptance. These challenges must be addressed to develop a more secure and trustworthy IoT ecosystem. To effectively tackle these issues, we need to establish international benchmarks and compliance frameworks for measurable resilience, similar to ISO standards for IoT antifragility. Furthermore, promoting open datasets and reproducible methodologies in adversarial IoT research is crucial for eliminating bias and ensuring fairness. Future work could promote human-in-the-loop governance models that integrate ethical reasoning, user consent, and human oversight into autonomous IoT decision systems.

7.7. Toward Antifragile IoT Ecosystems

Ultimately, the next frontier of IoT resilience lies in transcending robustness and resilience toward antifragility—systems that improve under stress rather than merely recover from it. This transition demands unifying the hardware, software, and governance dimensions into a cohesive learning ecosystem that self-evolves through exposure to perturbations. Achieving this vision will require continuous self-evaluation pipelines that inject and analyze stressors autonomously. Future research should integrate digital twins for iterative reinforcement, allowing virtualized IoT replicas to simulate, fail, and learn without disrupting the real world. Future work should incorporate cross-domain collaboration between AI, control theory, and cybersecurity communities to formalize antifragility metrics and validation procedures.

In summary, the evolution of IoT resilience research reflects a shift from post-hoc defense toward proactive, adaptive learning. Bridging adversarial robustness, federated cooperation, hardware-rooted trust, and explainable governance will pave the way for self-defending and self-improving IoT systems capable of sustaining critical operations in an increasingly adversarial and dynamic digital world.

8. Conclusions

This review provides a thorough analysis of resilience in the IoT, focusing on adversarial learning, intrusion detection, hardware safeguards, federated learning, and governance mechanisms. The resilience in IoT is not a single capability but an emergent property arising from the coordinated interaction of physical, cyber, and cognitive defenses. Across the literature, several clear patterns emerge. Deep learning models—while central to intelligent IoT analytics—remain highly susceptible to adversarial and environmental stress. Hybrid training, ensemble modeling, and generative data augmentation show promise in narrowing this vulnerability gap. Innovations in hardware and communication, such as PUFs and lightweight encryption, strengthen device integrity and trust. Federated and distributed learning enhance collaboration and privacy but pose risks like data poisoning

 doi:10.20944/preprints202512.1717.v1

and backdoor attacks. Explainable AI, blockchain auditability, and adaptive governance promote transparency and accountability. However, current efforts often lack effective cross-layer coordination and large-scale evaluations, as defenses are typically tested in simplified scenarios that don't reflect the complexity of IoT ecosystems. Looking forward, the evolution of IoT resilience will depend on unifying these fragmented approaches into adaptive, explainable, and antifragile architectures—systems capable not only of recovery but of improvement through exposure to disruption. This will require tighter integration across sensing, communication, and intelligence layers to achieve actual self-healing behavior. Federated learning frameworks need to ensure fairness and trust in challenging environments while developing energy-efficient and interpretable models. Research on IoT resilience is crucial, as the combination of AI, blockchain, and adaptive control offers a chance to improve the reliability of connected systems. By using antifragile design principles, future IoT ecosystems can evolve from vulnerable infrastructures into intelligent, self-sustaining networks that adapt and learn from their surroundings.

Funding: This research received no external funding.

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

References

1. Abikoye, O.C.; Bajeh, A.O.; Awotunde, J.B.; Ameen, A.O.; Mojeed, H.A.; Abdulraheem, M.; Oladipo, I.D.; Salihu, S.A. Application of internet of thing and cyber physical system in Industry 4.0 smart manufacturing. In *Emergence of Cyber Physical System and IoT in Smart Automation and Robotics: Computer Engineering in Automation*; Springer International Publishing: Cham, Switzerland, 2021; pp. 203–217.
2. Zeadally, S.; Bello, O. Harnessing the power of Internet of Things based connectivity to improve healthcare. *Internet Things* **2021**, *14*, 100074.
3. Djenna, A.; Harous, S.; Saidouni, D.E. Internet of things meet internet of threats: New concern cyber security issues of critical cyber infrastructure. *Appl. Sci.* **2021**, *11*, 4580.
4. Mishra, P.; Singh, G. Internet of Vehicles for Sustainable Smart Cities: Opportunities, Issues, and Challenges. *Smart Cities* **2025**, *8*, 93.
5. Hussain, M.Z.; Hanapi, Z.M. Efficient secure routing mechanisms for the low-powered IoT network: A literature review. *Electronics* **2023**, *12*, 482.
6. Tsiknas, K.; Taketzis, D.; Demertzis, K.; Skianis, C. Cyber threats to industrial IoT: A survey on attacks and countermeasures. *IoT* **2021**, *2*, 163–186.
7. Ntafloukas, K.; McCrum, D.P.; Pasquale, L. A cyber-physical risk assessment approach for internet of things enabled transportation infrastructure. *Appl. Sci.* **2022**, *12*, 9241.
8. Singh, K.; Yadav, M.; Singh, Y.; Moreira, F. Techniques in reliability of internet of things (IoT). *Procedia Comput. Sci.* **2025**, *256*, 55–62.
9. Nan, C.; Sansavini, G.; Kröger, W. Building an integrated metric for quantifying the resilience of interdependent infrastructure systems. In *International Conference on Critical Information Infrastructures Security*; Springer International Publishing: Cham, Switzerland, 2014; pp. 159–171.
10. Jin, X.; Gu, X. Option-based design for resilient manufacturing systems. *IFAC-Pap.* **2016**, *49*, 1602–1607.
11. Hosseini, S.; Barker, K. Modeling infrastructure resilience using Bayesian networks: A case study of inland waterway ports. *Comput. & Ind. Eng.* **2016**, *93*, 252–266.
12. Mottahedi, A.; Sereshki, F.; Ataei, M.; Nouri Qarahasanlou, A.; Barabadi, A. The resilience of critical infrastructure systems: A systematic literature review. *Energies* **2021**, *14*, 1571.
13. Rekeraho, A.; Cotfas, D.T.; Balan, T.C.; Cotfas, P.A.; Acheampong, R.; Tuyishime, E. Cybersecurity Threat Modeling for IoT-Integrated Smart Solar Energy Systems: Strengthening Resilience for Global Energy Sustainability. *Sustainability* **2025**, *17*, 2386.
14. Panda, D.; Padhy, N.; Sharma, K. Strengthening IoT Resilience: A Study on Backdoor Malware and DNS Spoofing Detection Methods. In *2025 International Conference on Emerging Systems and Intelligent Computing (ESIC)*; IEEE: Piscataway, NJ, USA, 2025; pp. 795–800.
15. Zhou, S.; Ye, D.; Zhu, T.; Zhou, W. (2025). Defending Against Neural Network Model Inversion Attacks via Data Poisoning. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 16324–16338.

 doi:10.20944/preprints202512.1717.v1

16. Fares, S.; Nandakumar, K. Attack to defend: Exploiting adversarial attacks for detecting poisoned models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 24726–24735.
17. Jarwan, A.; Sabbah, A.; Ibnkahla, M. Information-oriented traffic management for energy-efficient and loss-resilient IoT systems. *IEEE Internet Things J.* **2021**, *9*, 7388–7403.
18. Fang, X.; Zheng, L.; Fang, X.; Chen, W.; Fang, K.; Yin, L.; Zhu, H. Pioneering advanced security solutions for reinforcement learning-based adaptive key rotation in Zigbee networks. *Sci. Rep.* **2024**, *14*, 13931.
19. Cirne, A.; Sousa, P.R.; Resende, J.S.; Antunes, L. Hardware security for internet of things identity assurance. *IEEE Commun. Surv. Tutor.* **2024**, *26*, 1041–1079.
20. Delvaux, J.; Peeters, R.; Gu, D.; Verbauwhede, I. A survey on lightweight entity authentication with strong PUFs. *ACM Comput. Surv.* **2015**, *48*, 1–42.
21. Li, K.; Li, C.; Yuan, X.; Li, S.; Zou, S.; Ahmed, S.S.; Ni, W.; Niyato, D.; Jamalipour, A.; Dressler, F.; 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*, 46269–46293.
22. Chakraborty, A.; Alam, M.; Dey, V.; Chattopadhyay, A.; Mukhopadhyay, D. A survey on adversarial attacks and defences. *CAAI Trans. Intell. Technol.* **2021**, *6*, 25–45.
23. Goyal, S.; Doddapaneni, S.; Khapra, M.M.; Ravindran, B. A survey of adversarial defenses and robustness in nlp. *ACM Comput. Surv.* **2023**, *55*, 1–39.
24. Aaqib, M.; Ali, A.; Chen, L.; Nibouche, O. IoT trust and reputation: A survey and taxonomy. *J. Cloud Comput.* **2023**, *12*, 42.
25. Segovia-Ferreira, M.; Rubio-Hernan, J.; Cavalli, A.; Garcia-Alfaro, J. A survey on cyber-resilience approaches for cyber-physical systems. *ACM Comput. Surv.* **2024**, *56*, 1–37.
26. Khaloopour, L.; Su, Y.; Raskob, F.; Meuser, T.; Bless, R.; Janzen, L.; Abedi, K.; Andjelkovic, M.; Chaari, H.; Chakraborty, P.; et al. Resilience-by-design in 6G networks: Literature review and novel enabling concepts. *IEEE Access* **2024**, *12*, 155666–155695.
27. Alrumaih, T.N.; Alenazi, M.J.; AlSowaygh, N.A.; Humayed, A.A.; Alablani, I.A. Cyber resilience in industrial networks: A state of the art, challenges, and future directions. *J. King Saud Univ. Comput. Inf. Sci.* **2023**, *35*, 101781.
28. Berger, C.; Eichhammer, P.; Reiser, H.P.; Domaschka, J.; Hauck, F.J.; Habiger, G. A survey on resilience in the iot: Taxonomy, classification, and discussion of resilience mechanisms. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–39.
29. Grassi, V.; Mirandola, R.; Perez-Palacin, D. Towards a conceptual characterization of antifragile systems. In Proceedings of the 2023 IEEE 20th International Conference on Software Architecture Companion (ICSA-C), L'Aquila, Italy, 13–17 March 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 121–125.
30. Jin, M.; Lee, H. Position: Ai safety must embrace an antifragile perspective. *arXiv* **2025**, arXiv:2509.13339.
31. Koç, Y.; Warnier, M.; Van Mieghem, P.; Kooij, R.E.; Brazier, F.M. The impact of the topology on cascading failures in a power grid model. *Phys. A: Stat. Mech. Its Appl.* **2014**, *402*, 169–179.
32. Koç, Y.; Raman, A.; Warnier, M.; Kumar, T. Structural vulnerability analysis of electric power distribution grids. *Int. J. Crit. Infrastruct.* **2016**, *12*, 311–330.
33. Beyza, J.; Yusta, J.M. Characterising the security of power system topologies through a combined assessment of reliability, robustness, and resilience. *Energy Strategy Rev.* **2022**, *43*, 100944.
34. Ackermann, J. Parameter space design of robust control systems. *IEEE Trans. Autom. Control* **2003**, *25*, 1058–1072.
35. Alibašić, H. *Strategic Resilience and Sustainability Planning: Management Strategies for Sustainable and Climate-Resilient Communities and Organizations*; Springer: Berlin/Heidelberg, Germany, 2022.
36. Alibašić, H. Hyper-engaged citizenry, negative governance and resilience: Impediments to sustainable energy projects in the United States. *Energy Res. Soc. Sci.* **2023**, *100*, 103072.
37. Alibašić, H. Advancing disaster resilience: The ethical dimensions of adaptability and adaptive leadership in public service organizations. *Public Integr.* **2025**, *27*, 209–221.
38. Arghandeh, R.; Von Meier, A.; Mehrmanesh, L.; Mili, L. On the definition of cyber-physical resilience in power systems. *Renew. Sustain. Energy Rev.* **2016**, *58*, 1060–1069.
39. Zhou, Y.; Wang, J.; Yang, H. Resilience of transportation systems: Concepts and comprehensive review. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 4262–4276.
40. Taleb, N.N. *Antifragile: Things that Gain from Disorder*; Penguin UK, 2012

 doi:10.20944/preprints202512.1717.v1

41. Pravin, C.; Martino, I.; Nicosia, G.; Ojha, V. Fragility, robustness and antifragility in deep learning. *Artif. Intell.* **2024**, *327*, 104060.
42. Simpson, J.; Oosthuizen, R.; Sawah, S.E.; Abbass, H. Agile, antifragile, artificial-intelligence-enabled, command and control. *arXiv* **2021**, arXiv:2109.06874.
43. Scotti, V.; Perez-Palacin, D.; Brauzi, V.; Grassi, V.; Mirandola, R. *Antifragility via Online Learning and Monitoring: An IoT Case Study*; Karlsruher Institut für Technologie: Karlsruhe, Germany, 2025.
44. Jones, K.H. Engineering antifragile systems: A change in design philosophy. *Procedia Comput. Sci.* **2014**, *32*, 870–875.
45. Hillson, D. Beyond resilience: Towards antifragility? *Contin. Resil. Rev.* **2023**, *5*, 210–226.
46. Menon, D.; Anand, B.; Chowdhary, C.L. Digital twin: Exploring the intersection of virtual and physical worlds. *IEEE Access* **2023**, *11*, 75152–75172.
47. Stellios, I.; Kotzanikolaou, P.; Psarakis, M.; Alcaraz, C.; Lopez, J. A survey of iot-enabled cyberattacks: Assessing attack paths to critical infrastructures and services. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 3453–3495.
48. Raymond, D.R.; Midkiff, S.F. Denial-of-service in wireless sensor networks: Attacks and defenses. *IEEE Pervasive Comput.* **2008**, *7*, 74–81.
49. Ahmed, K.M.; Shams, R.; Khan, F.H.; Luque-Nieto, M.A. Securing underwater wireless sensor networks: A review of attacks and mitigation techniques. *IEEE Access* **2024**, *12*, 161096–161133.
50. Aslan, Ö.; Aktuğ, S.S.; Ozkan-Okay, M.; Yilmaz, A.A.; Akin, E. A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions. *Electronics* **2023**, *12*, 1333.
51. Sun, G.; Cong, Y.; Dong, J.; Wang, Q.; Lyu, L.; Liu, J. Data poisoning attacks on federated machine learning. *IEEE Internet Things J.* **2021**, *9*, 11365–11375.
52. Xia, G.; Chen, J.; Yu, C.; Ma, J. Poisoning attacks in federated learning: A survey. *IEEE Access* **2023**, *11*, 10708–10722.
53. Abroshan, H. AI to protect AI: A modular pipeline for detecting label-flipping poisoning attacks. *Mach. Learn. Appl.* **2025**, *22*, 100768.
54. Zeng, Y.; Pan, M.; Just, H.A.; Lyu, L.; Qiu, M.; Jia, R. Narcissus: A practical clean-label backdoor attack with limited information. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, Copenhagen, Denmark, 26–30 November 2023; pp. 771–785.
55. Hanif, M.A.; Chattopadhyay, N.; Ouni, B.; Shafique, M. Survey on Backdoor Attacks on Deep Learning: Current Trends, Categorization, Applications, Research Challenges, and Future Prospects. *IEEE Access* **2025**, *13*, 93190–93221.
56. Liang, Y.; He, D.; Chen, D. Poisoning attack on load forecasting. In Proceedings of the 2019 IEEE innovative smart grid technologies-Asia (ISGT Asia), Chengdu, China, 21–24 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1230–1235
57. Alotaibi, A.; Rassam, M.A. Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense. *Future Internet* **2023**, *15*, 62.
58. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
59. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*; Chapman and Hall/CRC: New York, NY, USA, 2018; pp. 99–112.
60. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv* **2017**, arXiv:1706.06083.
61. Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; Li, J. Boosting adversarial attacks with momentum. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9185–9193.
62. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. Deepfool: A simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.
63. Liu, J.; Li, Y.; Guo, Y.; Liu, Y.; Tang, J.; Nie, Y. Generation and Countermeasures of adversarial examples on vision: A survey. *Artif. Intell. Rev.* **2024**, *57*, 199.
64. Zheng, S.; Han, D.; Lu, C.; Hou, C.; Han, Y.; Hao, X.; Zhang, C. Transferable Targeted Adversarial Attack on Synthetic Aperture Radar (SAR) Image Recognition. *Remote Sens.* **2025**, *17*, 146.
65. Li, J.; Xu, Y.; Hu, Y.; Ma, Y.; Yin, X. You Only Attack Once: Single-Step DeepFool Algorithm. *Appl. Sci.* **2024**, *15*, 302.

 doi:10.20944/preprints202512.1717.v1

66. Yang, W.; Wang, S.; Wu, D.; Cai, T.; Zhu, Y.; Wei, S.; Zhang, Y.; Yang, X.; Tang, Z.; Li, Y. Deep learning model inversion attacks and defenses: A comprehensive survey. *Artif. Intell. Rev.* **2025**, *58*, 242.
67. Zhao, K.; Li, L.; Ding, K.; Gong, N.Z.; Zhao, Y.; Dong, Y. A Systematic Survey of Model Extraction Attacks and Defenses: State-of-the-Art and Perspectives. *arXiv* **2025**, arXiv:2508.15031.
68. Fang, H.; Qiu, Y.; Yu, H.; Yu, W.; Kong, J.; Chong, B.; Chen, B.; Wang, X.; Xia, Sh.; Xu, K. (2024). Privacy leakage on dnns: A survey of model inversion attacks and defenses. *arXiv* **2024**, arXiv:2402.04013.
69. Altaweel, A.; Mukkath, H.; Kamel, I. Gps spoofing attacks in fanets: A systematic literature review. *IEEE Access* **2023**, *11*, 55233–55280.
70. Parameswarath, R.P.; Abhishek, N.V.; Sikdar, B. A quantum safe authentication protocol for remote keyless entry systems in cars. In Proceedings of the 2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall), Hong Kong, 10–13 October 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–7.
71. Zhang, J.; Shen, G.; Saad, W.; Chowdhury, K. Radio frequency fingerprint identification for device authentication in the internet of things. *IEEE Commun. Mag.* **2023**, *61*, 110–115.
72. Coston, I.; Plotnizky, E.; Nojoumian, M. Comprehensive Study of IoT Vulnerabilities and Countermeasures. *Appl. Sci.* **2025**, *15*, 3036.
73. Zhang, J.; Ardizzon, F.; Piana, M.; Shen, G.; Tomasin, S. Physical Layer-Based Device Fingerprinting For Wireless Security: From Theory To Practice. *IEEE Trans. Inf. Forensics Secur.* **2025**, *20*, 5296–5325.
74. Pérez-Resa, A.; Garcia-Bosque, M.; Sánchez-Azqueta, C.; Celma, S. Self-synchronized encryption for physical layer in 10gbps optical links. *IEEE Trans. Comput.* **2019**, *68*, 899–911.
75. Kponyo, J.J.; Agyemang, J.O.; Klogo, G.S.; Boateng, J.O. Lightweight and host-based denial of service (DoS) detection and defense mechanism for resource-constrained IoT devices. *Internet Things* **2020**, *12*, 100319.
76. Pu, C. Energy depletion attack against routing protocol in the Internet of Things. In Proceedings of the 2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC), Las Vegas, NV, USA, 11–14 January 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–4.
77. Alamri, H.A.; Thayananthan, V. Bandwidth control mechanism and extreme gradient boosting algorithm for protecting software-defined networks against DDoS attacks. *IEEE Access* **2020**, *8*, 194269–194288.
78. Tang, D.; Dai, R.; Zuo, C.; Chen, J.; Li, K.; Qin, Z. A Low-rate DoS Attack Mitigation Scheme Based on Port and Traffic State in SDN. *IEEE Trans. Comput.* **2025**, *74*, 1758–1770.
79. Raikwar, M.; Gligoroski, D. Non-interactive vdf client puzzle for dos mitigation. In Proceedings of the 2021 European Interdisciplinary Cybersecurity Conference, Virtual, 10–11 November 2021; pp. 32–38.
80. Arabas, P.; Dawidiuk, M. Filter aggregation for DDoS prevention systems: Hardware perspective. *Int. J. Inf. Secur.* **2025**, *24*, 1–18.
81. Malhotra, P.; Singh, Y.; Anand, P.; Bangotra, D.K.; Singh, P.K.; Hong, W.C. Internet of things: Evolution, concerns and security challenges. *Sensors* **2021**, *21*, 1809.
82. Adelantado, F.; Vilajosana, X.; Tuset-Peiro, P.; Martinez, B.; Melia-Segui, J.; Watteyne, T. Understanding the limits of LoRaWAN. *IEEE Commun. Mag.* **2017**, *55*, 34–40.
83. Abdallah, B.; Khriji, S.; Chéour, R.; Lahoud, C.; Moessner, K.; Kanoun, O. Improving the reliability of long-range communication against interference for non-line-of-sight conditions in industrial Internet of Things applications. *Appl. Sci.* **2024**, *14*, 868.
84. Ghanem, A. Security Analysis of Rolling Code-based Remote Keyless Entry Systems. Ph.D. Thesis, 2022.
85. Csikor, L.; Lim, H.W.; Wong, J.W.; Ramesh, S.; Parameswarath, R.P.; Chan, M.C. Rollback: A new time-agnostic replay attack against the automotive remote keyless entry systems. *ACM Trans. Cyber-Phys. Syst.* **2024**, *8*, 1–25.
86. Tsunoda, Y.; Fujiwara, Y. The Asymptotics of Difference Systems of Sets for Synchronization and Phase Detection. In Proceedings of the 2023 IEEE International Symposium on Information Theory (ISIT), Taipei, Taiwan, 25–30 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 678–683.
87. Huang, P.; Chen, G.; Zhang, X.; Liu, C.; Wang, H.; Shen, H.; Bian, Y.; Lu, Y.; Ruan, Z.; Li, B.; et al. Fast and Scalable Selective Retransmission for RDMA. In Proceedings of the IEEE INFOCOM 2025-IEEE Conference on Computer Communications, London, UK, 19–22 May 2025; IEEE: Piscataway, NJ, USA, 2025; pp. 1–10.
88. Wu, Z.; Zhang, Y.; Tian, F.; Wu, M.; Zhai, A.; Zhang, Z.L. Interleaved Function Stream Execution Model for Cache-Aware High-Speed Stateful Packet Processing. In Proceedings of the 2024 IEEE 44th International Conference on Distributed Computing Systems (ICDCS), Jersey City, NJ, USA, 23–26 July 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 531–542.
89. Qadri, Y.A.; Jung, H.; Niyato, D. Toward the Internet of Medical Things for Real-Time Health Monitoring Over Wi-Fi. *IEEE Netw.* **2024**, *38*, 229–237.

 doi:10.20944/preprints202512.1717.v1

90. Gautam, M.K.; Pati, A.; Mishra, S.K.; Appasani, B.; Kabalci, E.; Bizon, N.; Thounthong, P. A comprehensive review of the evolution of networked control system technology and its future potentials. *Sustainability* **2021**, *13*, 2962.
91. Li, C.; Qi, P.; Wang, D.; Li, Z. On the anti-interference tolerance of cognitive frequency hopping communication systems. *IEEE Trans. Reliab.* **2020**, *69*, 1453–1464.
92. Wang, S.; Dai, W.; Sun, J.; Xu, Z.; Li, G.Y. Uncertainty awareness in wireless communications and sensing. In *IEEE Communications Magazine*; IEEE: Piscataway, NJ, USA, 2025.
93. Reinfurt, L.; Breitenbücher, U.; Falkenthal, M.; Leymann, F.; Riegg, A. Internet of things patterns. In Proceedings of the 21st European Conference on Pattern Languages of programs, Irsee, Germany, 6–10 July 2016; pp. 1–21.
94. Kumar, R.; Agrawal, N. EDMA-RM: An Event-Driven and Mobility-Aware Resource Management Framework for Green IoT-Edge-Fog-Cloud Networks. *IEEE Sens. J.* **2024**, *24*, 23004–23012.
95. Al-Kadhim, H.M.; Al-Raweshidy, H.S. Energy efficient data compression in cloud based IoT. *IEEE Sens. J.* **2021**, *21*, 12212–12219.
96. Sabovic, A.; Aernouts, M.; Subotic, D.; Fontaine, J.; De Poorter, E.; Famaey, J. Towards energy-aware tinyML on battery-less IoT devices. *Internet Things* **2023**, *22*, 100736.
97. Tekin, N.; Acar, A.; Aris, A.; Uluagac, A.S.; Gungor, V.C. Energy consumption of on-device machine learning models for IoT intrusion detection. *Internet Things* **2023**, *21*, 100670.
98. MeasureX. Understanding Pressure Sensor Drift: Causes, Effects & How to Prevent It. 25 August 2025. Available online: https://www.measurex.com.au (accessed on: 12 November 2025).
99. Sutar, S.; Raha, A.; Raghunathan, V. Memory-based combination PUFs for device authentication in embedded systems. *IEEE Trans. Multi-Scale Comput. Syst.* **2018**, *4*, 793–810.
100. Kannan, R.; Jain, S. Adaptive recalibration algorithm for removing sensor errors and its applications in motion tracking. *IEEE Sens. J.* **2018**, *18*, 2916–2924.
101. Kusters, C.J. *Helper Data Schemes for Secret-Key Generation Based on Sram Pufs: Bias & Multiple Observations*; Technische Universiteit Eindhoven: Eindhoven, The Netherlands, 2020.
102. Karpinskyy, B.; Lee, Y.; Choi, Y.; Kim, Y.; Noh, M.; Lee, S. 8.7 Physically unclonable function for secure key generation with a key error rate of 2E-38 in 45nm smart-card chips. In Proceedings of the 2016 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 31 January–4 February 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 158–160.
103. Yang, L.; Shami, A. IoT data analytics in dynamic environments: From an automated machine learning perspective. *Eng. Appl. Artif. Intell.* **2022**, *116*, 105366.
104. Alqahtani, M. Nonlinear autoregressive prediction model for VAWT power supply network energy management. *Energy Rep.* **2025**, *13*, 5446–5462.
105. Yang, L.; Shami, A. A lightweight concept drift detection and adaptation framework for IoT data streams. *IEEE Internet Things Mag.* **2021**, *4*, 96–101.
106. Yang, L.; Manias, D.M.; Shami, A. Pwpae: An ensemble framework for concept drift adaptation in iot data streams. In Proceedings of the 2021 IEEE Global Communications Conference (Globecom), Madrid, Spain, 7–11 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
107. Gupta, B.B.; Quamara, M. An overview of Internet of Things (IoT): Architectural aspects, challenges, and protocols. *Concurr. Comput. Pract. Exp.* **2020**, *32*, e4946.
108. Cirillo, F.; Esposito, C. Efficient PUF-Based IoT Authentication Framework without Fuzzy Extractor. In Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing, Sicily, Italy, 31 March–4 April 2025; pp. 695–704.
109. Al-Qaisi, A.; Aldahdouh, K.; Al-Sit, W.T.; Olaimat, A.A.; Alouneh, S.; Darabkh, K.A. Low Power Wide Area Network (LPWAN) Protocols: Enablers for Future Wireless Networks. *Results Eng.* **2025**, *27*, 105866.
110. Araujo, R.; da Silva, L.; Santos, W.; Souza, M. Cognitive Radio Strategy Combined with MODCOD Technique to Mitigate Interference on Low-Orbit Satellite Downlinks. *Sensors* **2023**, *23*, 7234.
111. Burhan, M.; Rehman, R.A.; Khan, B.; Kim, B.S. IoT elements, layered architectures and security issues: A comprehensive survey. *sensors* **2018**, *18*, 2796.
112. Awad, Z.; Zakaria, M.; Hassan, R. An enhanced ensemble defense framework for boosting adversarial robustness of intrusion detection systems. *Sci. Rep.* **2025**, *15*, 14177.
113. Lim, W.; Yong, K.S.C.; Lau, B.T.; Tan, C.C.L. Future of generative adversarial networks (GAN) for anomaly detection in network security: A review. *Comput. Secur.* **2024**, *139*, 103733.

 doi:10.20944/preprints202512.1717.v1

114. Yan, H.; Lin, X.; Li, S.; Peng, H.; Zhang, B. Global or local adaptation? Client-sampled federated meta-learning for personalized IoT intrusion detection. *IEEE Trans. Inf. Forensics Secur.* **2024**, *20*, 279–293.
115. Peng, J.; Li, W.; Vlaski, S.; Ling, Q. Mean aggregator is more robust than robust aggregators under label poisoning attacks on distributed heterogeneous data. *J. Mach. Learn. Res.* **2025**, *26*, 1–51.
116. Erazo-Garzón, L.; Cedillo, P.; Rossi, G.; Moyano, J. A domain-specific language for modeling IoT system architectures that support monitoring. *IEEE Access* **2022**, *10*, 61639–61665.
117. Mikołajewska, E.; Mikołajewski, D.; Mikołajczyk, T.; Paczkowski, T. Generative AI in AI-based digital twins for fault diagnosis for predictive maintenance in Industry 4.0/5.0. *Appl. Sci.* **2025**, *15*, 3166.
118. Kiasari, M.; Ghaffari, M.; Aly, H.H. A comprehensive review of the current status of smart grid technologies for renewable energies integration and future trends: The role of machine learning and energy storage systems. *Energies* **2024**, *17*, 4128.
119. Kulothungan, V. Using Blockchain Ledgers to Record AI Decisions in IoT. *IoT* **2025**, *6*, 37.
120. Hermosilla, P.; Berríos, S.; Allende-Cid, H. Explainable AI for Forensic Analysis: A Comparative Study of SHAP and LIME in Intrusion Detection Models. *Appl. Sci.* **2025**, *15*, 7329.
121. Naik, N.; Surendranath, N.; Raju, S.A.B.; Madduri, C.; Dasari, N.; Shukla, V.K.; Patil, V. Hybrid deep learning-enabled framework for enhancing security, data integrity, and operational performance in Healthcare Internet of Things (H-IoT) environments. *Sci. Rep.* **2025**, *15*, 31039.
122. Saranya, A.; Subhashini, R. A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decis. Anal. J.* **2023**, *7*, 100230.
123. Antonelli, F.; Yang, M.; Cozzani, V. Enhancing pipeline system resilience: A reliability-centric approach. *J. Pipeline Sci. Eng.* **2024**, *5*, 100252.
124. Abdulhussain, S.H.; Mahmmod, B.M.; Alwhelat, A.; Shehada, D.; Shihab, Z.I.; Mohammed, H.J.; Abdulameer, T.H.; Alsabah, M.; Fadel, M.H.; Ali, S.K.; et al. A comprehensive review of sensor technologies in IOT: Technical aspects, challenges, and future directions. *Computers* **2025**, *14*, 342.
125. Wang, Z.; Yu, J.; Gao, M.; Yuan, W.; Ye, G.; Sadiq, S.; Yin, H. Poisoning attacks and defenses in recommender systems: A survey. *arXiv* **2024**, arXiv:2406.01022.
126. Shi, Y.; Sagduyu, Y.E. Evasion and causative attacks with adversarial deep learning. In Proceedings of the MILCOM 2017—2017 IEEE Military Communications Conference (MILCOM), Baltimore, MA, USA, 23–25 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 243–248.
127. An, S.; Jang, D.J.; Lee, E.K. Adversarial Evasion Attacks on SVM-Based GPS Spoofing Detection Systems. *Sensors* **2025**, *25*, 6062.
128. Rahman, S.; Pal, S.; Fallah, A.; Doss, R.; Karmakar, C. RAD-IoMT: Robust adversarial defense mechanisms for IoMT medical image analysis. *Ad. Hoc. Netw.* **2025**, *178*, 103935.
129. Kermany, D. Labeled Optical Coherence Tomography (oct) and Chest X-Ray Images for Classification. Mendeley Data. 2018. Available online: https://cir.nii.ac.jp (accessed on: 12 November 2025).
130. Gungor, O.; Rosing, T.; Aksanli, B. Stewart: Stacking ensemble for white-box adversarial attacks towards more resilient data-driven predictive maintenance. *Comput. Ind.* **2022**, *140*, 103660.
131. Bosello, M. UNIBO Powertools Dataset. 2021. Available online: https://cris.unibo.it (accessed on: 12 November 2025).
132. Zhang, L.; Lambotharan, S.; Zheng, G.; Liao, G.; Liu, X.; Roli, F.; Maple, C. Vision Transformer with Adversarial Indicator Token against Adversarial Attacks in Radio Signal Classifications. *IEEE Internet Things J.* **2025**, *12*, 35367–35379.
133. Zyane, A.; Jamiri, H. Securing IoT Networks with Adversarial Learning: A Defense Framework Against Cyber Threats. In Proceedings of the 2025 5th International Conference on Innovative Research in Applied Science, Engineering and Technology (IRASET), Dhar El Mahraz Fez, Morocco, 15–16 May 2025; IEEE: Piscataway, NJ, USA, 2025; pp. 1–7.
134. Neto, E.C.P.; Dadkhah, S.; Ferreira, R.; Zohourian, A.; Lu, R.; Ghorbani, A.A. CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment. *Sensors* **2023**, *23*, 5941.
135. Efatinasab, E.; Brighente, A.; Donadel, D.; Conti, M.; Rampazzo, M. Towards robust stability prediction in smart grids: GAN-based approach under data constraints and adversarial challenges. *Internet Things* **2025**, *33*, 101662.
136. Arzamasov, V. Electrical grid stability simulated data. *UCI Mach. Learn. Repos.* **2018**, *10*, C5PG66.
137. Javed, H.; El-Sappagh, S.; Abuhmed, T. Robustness in deep learning models for medical diagnostics: Security and adversarial challenges towards robust AI applications. *Artif. Intell. Rev.* **2024**, *58*, 12.

 doi:10.20944/preprints202512.1717.v1

138. Goodfellow, I.; Papernot, N.; McDaniel, P.D.; Feinman, R.; Faghri, F.; Matyasko, A.; ... & Sheatsley, R. Cleverhans v0.1: An adversarial machine learning library. *arXiv* **2016**, arXiv:1610.0076817.
139. Papernot, N.; Goodfellow, I.; Sheatsley, R.; Feinman, R.; McDaniel, P. cleverhans v2.0.0: An adversarial machine learning library. *arXiv* **2016**, arXiv:1610.0076810.
140. Moghaddam, P.S.; Vaziri, A.; Khatami, S.S.; Hernando-Gallego, F.; Martín, D. Generative Adversarial and Transformer Network Synergy for Robust Intrusion Detection in IoT Environments. *Future Internet* **2025**, *17*, 258.
141. Alsaedi, A.; Moustafa, N.; Tari, Z.; Mahmood, A.; Anwar, A. TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems. *IEEE Access* **2020**, *8*, 165130–165150.
142. Tian, J.; Wang, B.; Li, J.; Konstantinou, C. Adversarial attack and defense methods for neural network based state estimation in smart grid. *IET Renew. Power Gener.* **2022**, *16*, 3507–3518.
143. Hong, T.; Pinson, P.; Fan, S. Global energy forecasting competition 2012. *Int. J. Forecast.* **2014**, *30*, 357–363.
144. Tusher, A.S.; Rahman, M.A.; Islam, M.R.; Bosak, S.; Hossain, M.J. FEMUS-Nowcast: A Robust Deep Learning Model for Sky Image–Based Short-Term Solar Forecasting Under Adversarial Attacks. *Int. J. Energy Res.* **2025**, *2025*, 8286945.
145. Nie, Y.; Li, X.; Scott, A.; Sun, Y.; Venugopal, V.; Brandt, A. SKIPP'D: A SKy Images and Photovoltaic Power Generation Dataset for short-term solar forecasting. *Sol. Energy* **2023**, *255*, 171–179.
146. Alsubai, S.; Karovič, V.; Almadhor, A.; Hejaili, A.A.; Juanatas, R.A.; Sampedro, G.A. Future-Proofing AI Models in IoMT Environment: Adversarial Dataset Generation and Defense Strategies. *Digit. Twins Appl.* **2025**, *2*, e70010.
147. Hady, A.A.; Ghubaish, A.; Salman, T.; Unal, D.; Jain, R. Intrusion detection system for healthcare systems using medical and network data: A comparison study. *IEEE Access* **2020**, *8*, 106576–106584.
148. O'shea, T.J.; West, N. Radio machine learning dataset generation with gnu radio. *GNU Radio Conf.* **2016**, *1*, 1.
149. Luan, S.; Gao, Y.; Liu, T.; Li, J.; Zhang, Z. Automatic modulation classification: Cauchy-Score-function-based cyclic correlation spectrum and FC-MLP under mixed noise and fading channels. *Digit. Signal Process.* **2022**, *126*, 103476.
150. Son, N.K.; Sangaiah, A.K.; Medhane, D.V.; Alenazi, M.J.; Aborokbah, M. Enhancing Resilience in Edge IoT Devices Against Adversarial Attacks. *IEEE Consum. Electron. Mag.* **2024**, *14*, 48–56.
151. Moustafa, N.; Slay, J. UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In Proceedings of the 2015 Military Communications and Information Systems Conference (MilCIS), Canberra, Australia, 10–12 November 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1–6.
152. Moustafa, N. New generations of internet of things datasets for cybersecurity applications based machine learning: TON_IoT datasets. In Proceedings of the eResearch Australasia Conference, Brisbane, Australia, 22–24 October 2019; pp. 21–25.
153. Hink, R.C.B.; Beaver, J.M.; Buckner, M.A.; Morris, T.; Adhikari, U.; Pan, S. Machine learning for power system disturbance and cyber-attack discrimination. In Proceedings of the 2014 7th International Symposium on Resilient Control Systems (ISRCS), Denver, CO, USA, 19–21 August 2014; IEEE: Piscataway, NJ, USA, 2019; pp. 1–8
154. Khatami, S.S.; Shoeibi, M.; Oskouei, A.E.; Martin, D.; Dashliboroun, M.K. 5DGWO-GAN: A Novel Five-Dimensional Gray Wolf Optimizer for Generative Adversarial Network-Enabled Intrusion Detection in IoT Systems. *Computers. Mater. Contin.* **2025**, *82*, 881–911.
155. Alwaisi, Z. Memory-efficient and robust detection of Mirai botnet for future 6G-enabled IoT networks. *Internet Things* **2025**, *32*, 101621.
156. Vajrobol, V.; Gupta, B.B.; Gaurav, A.; Chuang, H.M. Adversarial learning for Mirai botnet detection based on long short-term memory and XGBoost. *Int. J. Cogn. Comput. Eng.* **2024**, *5*, 153–160.
157. Alajaji, A. FortiNIDS: Defending Smart City IoT Infrastructures Against Transferable Adversarial Poisoning in Machine Learning-Based Intrusion Detection Systems. *Sensors* **2025**, *25*, 6056.
158. Sharafaldin, I.; Lashkari, A.H.; Hakak, S.; Ghorbani, A.A. Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy. In Proceedings of the 2019 International Carnahan Conference on Security Technology (ICCST), Chennai, India, 1–3 October 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–8.
159. Omara, A.; Kantarci, B. An AI-driven solution to prevent adversarial attacks on mobile Vehicle-to-Microgrid services. *Simul. Model. Pract. Theory* **2024**, *137*, 103016.
160. Huber, P.; Ott, M.; Friedli, M.; Rumsch, A.; Paice, A. Residential power traces for five houses: The ihomelab rapt dataset. *Data* **2020**, *5*, 17.

 doi:10.20944/preprints202512.1717.v1

161. Morshedi, R.; Matinkhah, S.M. Combining Generative Adversarial Networks (GANs) With Gaussian Noise for Anomaly Detection in Internet of Things (IoT) Traffic. *Eng. Rep.* **2025**, *7*, e70205.
162. Sharafaldin, I.; Lashkari, A.H.; Ghorbani, A.A. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp* **2018**, *1*, 108–116.
163. Lakshminarayana, S.; Sthapit, S.; Jahangir, H.; Maple, C.; Poor, H.V. Data-driven detection and identification of IoT-enabled load-altering attacks in power grids. *IET Smart Grid* **2022**, *5*, 203–218.
164. Bao, Z.; Lin, Y.; Zhang, S.; Li, Z.; Mao, S. Threat of adversarial attacks on DL-based IoT device identification. *IEEE Internet Things J.* **2021**, *9*, 9012–9024.
165. Sánchez, P.M.S.; Celdrán, A.H.; Bovet, G.; Pérez, G.M. Adversarial attacks and defenses on ML-and hardware-based IoT device fingerprinting and identification. *Future Gener. Comput. Syst.* **2024**, *152*, 30–42.
166. Sánchez, P.M.S.; Valero, J.M.J.; Celdrán, A.H.; Bovet, G.; Pérez, M.G.; Pérez, G.M. LwHBench: A low-level hardware component benchmark and dataset for Single Board Computers. *Internet Things* **2023**, *22*, 100764.
167. Cao, Y.; Kou, M.; Lai, Y.; Mei, Z. S²-Code: A Resilient and Lightweight Self-Synchronizing Authentication Protocol for Unreliable IoT Networks. *IEEE Access* **2025**, *13*, 156153–156169.
168. Blanchet, B.; Smyth, B.; Cheval, V.; Sylvestre, M. ProVerif 2.00: Automatic cryptographic protocol verifier, user manual and tutorial. *Version* **2018**, *16*, 5–16.
169. Hemavathy, S.; Bhaaskaran, V.K. Arbiter PUF—A review of design, composition, and security aspects. *IEEE Access* **2023**, *11*, 33979–34004.
170. Aribilola, I.; Lee, B.; Asghar, M.N. Möbius transformation and permutation based S-box to enhance IOT multimedia security. *IEEE Access* **2024**, *12*, 140792–140808.
171. Elhajj, M.; Attar, A.E.; Mikati, A. Integrating IoT and blockchain for smart urban energy management: Enhancing sustainability through real-time monitoring and optimization. *Clust. Comput.* **2025**, *28*, 960.
172. Kelly, J.; Knottenbelt, W. The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. *Sci. Data* **2015**, *2*, 1–14.
173. Borst, T.W. Aggregation of energy consumption forecasts across spatial levels. 2023.
174. Alnfiai, M.M. AI-powered cyber resilience: A reinforcement learning approach for automated threat hunting in 5G networks. *EURASIP J. Wirel. Commun. Netw.* **2025**, *2025*, 68.
175. Dong, H.; Wei, Z.; Peiyi, C.; Yiqing, L.; Hua, H. Multi-Layered Optimization for Adaptive Decoy Placement in Cyber-Resilient Power Systems Under Uncertain Attack Scenarios. *IET Renew. Power Gener.* **2025**, *19*, e70078.
176. Reis, M.J. Edge-FLGuard: A Federated Learning Framework for Real-Time Anomaly Detection in 5G-Enabled IoT Ecosystems. *Appl. Sci.* **2025**, *15*, 6452.
177. Albanbay, N.; Tursynbek, Y.; Graffi, K.; Uskenbayeva, R.; Kalpeyeva, Z.; Abilkaiyr, Z.; Ayapov, Y. Federated learning-based intrusion detection in IoT networks: Performance evaluation and data scaling study. *J. Sens. Actuator Netw.* **2025**, *14*, 78.
178. Shabbir, A.; Manzoor, H.U.; Manzoor, M.N.; Hussain, S.; Zoha, A. Robustness against data integrity attacks in decentralized federated load forecasting. *Electronics* **2024**, *13*, 4803.
179. Mulla, R. Hourly Energy Consumption. Available online: https://www.kaggle.com (accessed on 12 November 2025).
180. Haghbin, Y.; Badiei, M.H.; Tran, N.H.; Piran, M.J. Resilient Federated Adversarial Learning With Auxiliary-Classifier GANs and Probabilistic Synthesis for Heterogeneous Environments. *IEEE Trans. Netw. Serv. Manag.* **2025**, *22*, 4998–5014.
181. Deng, L. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Process. Mag.* **2012**, *29*, 141–142.
182. Cohen, G.; Afshar, S.; Tapson, J.; Van Schaik, A. (2017, May). EMNIST: Extending MNIST to handwritten letters. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 2921–2926.
183. Al Dalaien, M.; Ullah, R.; Al-Haija, Q.A. A Dual-Aggregation Approach to Fortify Federated Learning Against Poisoning Attacks in IoTs. *Array* **2025**, 100520.
184. Mukisa, K.J.; Ahakonye, L.A.C.; Kim, D.S.; Lee, J.M. Blockchain-Augmented FL IDS for Non-IID Edge-IoT Data Using Trimmed Mean Aggregation. *IEEE Internet Things J.* **2025**, *12*, 45150–45159.
185. Neto, E.C.P.; Taslimasa, H.; Dadkhah, S.; Iqbal, S.; Xiong, P.; Rahman, T.;Ali A. Ghorbani. CICIoV2024: Advancing realistic IDS approaches against DoS and spoofing attack in IoV CAN bus. *Internet Things* **2024**, *26*, 101209.

 doi:10.20944/preprints202512.1717.v1

186. Ferrag, M.A.; Friha, O.; Hamouda, D.; Maglaras, L.; Janicke, H. Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning. *IEEe Access* **2022**, *10*, 40281–40306.
187. Kumar, P.; Mullick, S.; Das, R.; Nandi, A.; Banerjee, I. IoTForge Pro: A security testbed for generating intrusion dataset for industrial IoT. *IEEE Internet Things J.* **2024**, *12*, 8453–8460.
188. J. Vinita, An incentive-aware federated bargaining approach for client selection in decentralized federated learning for IoT smart homes. *Sci. Rep.* **2025**, *15*, 34412.
189. Prasad, K.S.; Udayakumar, P.; Laxmi Lydia, E.; Ahmed, M.A.; Ishak, M.K.; Karim, F.K.; Mostafa, S.M. A two-tier optimization strategy for feature selection in robust adversarial attack mitigation on internet of things network security. *Sci. Rep.* **2025**, *15*, 2235.
190. ALFahad, S.; Parambath, S.P.; Anagnostopoulos, C.; Kolomvatsos, K. Node selection using adversarial expert-based multi-armed bandits in distributed computing. *Computing* **2025**, *107*, 85.
191. Nugraha, B.; Jnanashree, A.V.; Bauschert, T. A versatile XAI-based framework for efficient and explainable intrusion detection systems. *Ann. Telecommun.* **2025**, 1–26.
192. St, L.; Wold, S. Analysis of variance (ANOVA). *Chemom. Intell. Lab. Syst.* **1989**, *6*, 259–272.
193. Amponis, G.; Radoglou-Grammatikis, P.; Nakas, G.; Goudos, S.; Argyriou, V.; Lagkas, T.; Sarigiannidis, P. 5G core PFCP intrusion detection dataset. In Proceedings of the 2023 12th International Conference on Modern Circuits and Systems Technologies (MOCAST), Athens, Greece, 28–30 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–4.
194. Johnson, A.E.; Pollard, T.J.; Shen, L.; Lehman, L.W.H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L.A.; Mark, R.G. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016**, *3*, 1–9.
195. Moody, G.B.; Mark, R.G. The impact of the MIT-BIH arrhythmia database. *IEEE Eng. Med. Biol. Mag.* **2001**, *20*, 45–50.
196. Majumdar, S.; Awasthi, A. From vulnerability to resilience: Securing public safety GPS and location services with smart radio, blockchain, and AI-driven adaptability. *Electronics* **2025**, *14*, 1207.
197. He, K.; Kim, D.D.; Asghar, M.R. NIDS-Vis: Improving the generalized adversarial robustness of network intrusion detection system. *Comput. Secur.* **2024**, *145*, 104028.
198. He, K.; Kim, D.; Zhang, Z.; Ge, M.; Lam, U.; Yu, J. *UQ IoT IDS Dataset 2021*; The University of Queensland: Brisbane City, Australia, 2022.
199. Al-Fawa'reh, M.; Abu-Khalaf, J.; Janjua, N.; Szewczyk, P. Detection of on-manifold adversarial attacks via latent space transformation. *Comput. Secur.* **2025**, *154*, 104431.
200. Al-Hawawreh, M.; Sitnikova, E.; Aboutorab, N. X-IIoTID: A connectivity-agnostic and device-agnostic intrusion data set for industrial Internet of Things. *IEEE Internet Things J.* **2021**, *9*, 3962–3977.
201. Alasmari, S.M.; Sakly, H.; Kraiem, N.; Algarni, A. Phishing detection in IoT: An integrated CNN-LSTM framework with explainable AI and LLM-enhanced analysis. *Discov. Internet Things* **2025**, *5*, 102.
202. Chiew, K.L.; Tan, C.L.; Wong, K.; Yong, K.S.; Tiong, W.K. A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. *Inf. Sci.* **2019**, *484*, 153–166.
203. Opara, C.; Chen, Y.; Wei, B. Look before you leap: Detecting phishing web pages by exploiting raw URL and HTML characteristics. *Expert Syst. Appl.* **2024**, *236*, 121183.
204. Singh, A.K. Malicious and benign webpages dataset. *Data Brief* **2020**, *32*, 106304.
205. Hannousse, A.; Yahiouche, S. Towards benchmark datasets for machine learning based website phishing detection: An experimental study. *Eng. Appl. Artif. Intell.* **2021**, *104*, 104347.