Dijlah Journal of Engineering Science (DJES) Vol. 3, No. 2, June, 2026, pp. 28-38
ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 145

Adversarial Attacks on AI-Based Botnet Detection Systems in IoT: Key Threats and Countermeasures

Thurayya Breesam Kareem
University of Basrah, College of Fine Artst

Article Info

Article history:

Received Dec., 08, 2025
Revised Feb.,22, 2026
Accepted Mar.,18, 2026

Keywords:

Adversarial Machine Learning (AML), IoT Security, Botnet Detection, Federated Learning (FL), Adversarial Training

ABSTRACT

Artificial Intelligence (AI) models are critical for detecting advanced Internet of Things (IoT) botnets. However, these systems are highly vulnerable to Adversarial Machine Learning (AML), where malicious inputs are crafted to cause misclassification (e.g., identifying malicious traffic as benign), posing a systemic threat to IoT security. This systematic literature review (SLR) addresses the persistent "reality gap" between theoretical AML research, often derived from computer vision, and the practical, domain-specific constraints of network security.This paper synthesizes research from 2020–2025, providing comprehensive taxonomies of: (RQ1) targeted AI models, from traditional ML to modern Federated Learning (FL) frameworks; (RQ2) attack methodologies, highlighting the shift from feature-space (e.g., PGD) to realistic problem-space attacks (e.g., binary diversification, XAI-based attacks); (RQ3) proactive (e.g., Adversarial Training) and reactive defense strategies; and (RQ4) evaluation frameworks, critiquing the use of outdated datasets. Finally, (RQ5) we analyze open challenges, focusing on the IoT resource-constraint dilemma—where effective defenses like Adversarial Training are too computationally expensive for edge devices —and performance trade-offs. We conclude by outlining future directions, emphasizing the need for constraint-aware defenses, secure FL, and leveraging Generative AI.

Corresponding Author:

Thurayya Breesam Kareem
College of Fine Artst, University of Basrah
Basrah, Iraq
Email : thuraya.breesam@uobasrah.edu.iq

1. INTRODUCTION

The ubiquitous IoT is a corner stone of the Fourth Industrial Revolution (4IR) that supports shifting the paradigm of data generation and automation in various essential domains such as smart homes, healthcare, and Industrial- IoT (IIoT) [1]. However, such huge deployment opens a much larger and more heterogeneous attack surface. A significant number of IoT devices have inherent limitations, such as low computational power, memory, and storage, and may not be backed up by strong patchable security features making them the perfect candidates for cyber-attacks [2]. Almost certainly its ecosystems are a IoT botnets; which compromised devices are under the remote control of a "botmaster"[1]. These Botnets are the origin for notorious high-impact attacks like DDoS (Distributed Denial-of-Service) that can disable sensitive internet services including the famous Mirai botnet [1], [3]. Due to the constantly changing polymorphic nature of these threats, detection based on signatures is frequently ineffective [3], [4]. As a result, it has lead to more incorporation of Artificial Intelligence (AI) such as Machine Learning (ML) and Deep Learning (DL), in the development of modern Network Intrusion Detection Systems (NIDS). These AI-models are perfect for examining network behavior and identifying anomalies, providing a mechanism for catching new botnet activity that static signatures would never see [4], [5]. Despite detection performance, AI/ML models suffer from a serious system-level vulnerability: susceptibility to disruption through AML [6], [7]. An adversarial attack is a method in which an attacker maliciously designs some inputs, called "adversarial samples", by adding subtle perturbations to clean data, making the latter susceptible (sometimes completely) to chaotic phenomena [8].

These carefully crafted inputs are customized to not only exploit the learned decision boundaries of a model but also to result in a misclassification (e.g., making a botnet classifier classify malicious traffic as benign [7,8]). This is not a little bug, but systemic threat for the trust and reliability in AI based security systems [6], [9]. This risk is particularly high in the context of IoT for two main reasons. First, the consequences of a successful attack are typically not restricted to the digital domain but may propagate in physical form (e.g., causing interruptions in industrial activities or disrupting health monitoring) [5]. Moreover, the ”insider threat” scenario (an attacker controls only one device in the network [10], [6]) can pragmatically serve as a vector to launch advanced white-box or gray-box attacks against all the security defenses. The agreement with the assumption that AML is a well-investigated domain, it means; most of the early stage works were dedicated to non-structured data domains like computer vision [11]. Its evolution in a well-structured, time-series and semantically rigid domain like the network security (more specifically for IoT botnet detection) is less mature than its usage in traditional domains where it brings particular challenges [12]. However, there still remains a considerable “reality gap” where a many of the theoretically derived attacks are not actually practical since they do not respect operational constraints and interdependencies in network protocols [13], [14]. , it is of great importance to conduct a systematic review to summarize and categorize the adversarial attacks and defenses as well as make a critically analysis towards the existing works on adversarial attacking against AI-empowered IoT botnet detectors. Reading through the fragmented literature, to map this landscape, identify the “realistic” threat vectors and assessing whether proposed defences work [15] has motivated this research as well. This survey provides a basis for researchers to find some promising future directions and for practitioners to understand more advanced next-generation security solutions that can better address increasingly sophisticated adversaries. In this regard, our paper makes a contribution by conducting a systematic literature review (SLR) to identify the state-of-the-art taxonomy of adversarial attacks and defenses on AI-based IoT botnet detection systems. The structure of this review is designed to address these five core research questions methodologically:

1. What are the most seen AI/ML models (e.g., DNNs, SVM, Random Forest) that can be subject to adversarial attacks in IoT botnet detection?
2. What type of adversarial attacks can we launch on botnet detectors (Taxonomy)?
3. What are the defensive techniques (Adversarial Training, Defensive Distillation) recommended to enhance the resistance of botnet detectors against these attacks?
4. What are the metrics and datasets used to quantify the effectiveness of attacks and robustness of defense? What are the research gaps, open issues directions for future work in LES?

2. LITERATURE REVIEW: TAXONOMIES AND FRAMEWORKS

In this section, we report our main discoveries extracted from the systematic review and organize the synthesized papers in a structured taxonomy. It addresses explicitly the first four research questions (RQ1-RQ4), by presenting a full taxonomy of: targeted AI/ML models, adversarial attack methods used against them, defense techniques proposed and evaluation approaches applied to measure robustness.

2.1 Taxonomy of Targeted Models (RQ1) Task model

Process models are for organizing activities, when the main task to be accomplished is already defined at a high level and not further specifications at lower levels need to be described by the process model; goal models, conversely, represent from a more abstract level down to more specific details; target-oriented design process: aimed at one or several levels unspecified in advance but according to the type of task could specify low-level activities commands:

- The attack surface for AML in IoT botnet detection domain is broad and embraces multiple AI/ML models. Adversary’s selecting the target model is commonly dependent on how popular this model is in security systems, and how its design and internal architecture make it susceptive to attacks. According to the colleting literature, we categorize the main types of models attacked in three categories (including transferring attack) and one benchmark, as presented in Table I.

- Typical Machine Learning Classifiers: This class of classifiers comprises the early models often used for NIDS and are still current as crucial benchmark models against which robustness has been tested. Such models are focused by their opponents as they have an easier decision boundary. A thorough study [17] analyzing a wide range of nine classical models showed the strength weakness of them. The most common algorithms covered in the literature are: SVM, RF, DT, KNN and NB etc [17].

  • Deep Learning (DL) Models: With the advancement of detection systems to cope with the high volume and high velocity nature of IoT data, Deep Learning (DL) models are now the focus of both work on deployment as well as work on adversarial. The complexity, although results in high accuracy, forms a large and non-linear attack surface [18], [19]. This encompasses Deep Neural Networks (DNNs),

Multilayer Perceptrons (MLPs) [18] and sequential models as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTMs) [19].

- Advanced and Hybrid Architectures: When previous modes of detection mature, adversaries shift their focus on resource-intensive specialized next-generation models. These range from hybrid architectures (LSTM-CNN) [20], GNNs for network graph analysis [21] and variations in architecture (SNN) [22].

Distributed and Federated Approaches: The distributed type of IoT has given rise to the use of FL [23]. Under this paradigm, models are trained directly on edge devices, and only model updates are transmitted to a central server for aggregation [24]. This architectural design brings to the scene a new attack surface, and consequently FL models have become one of the main emerging targets, specially for poisoning attacks [23], [24].

Table 1; Taxonomy Of Targeted Ai/Ml Models In Iot Botnet Detection (Rq1).

Category	Key Models	Purpose in IoT Detection	Key Vulnerability / Attack Vector
Traditional ML	SVM, RF, DT, KNN	Baseline detection of anomalies/attacks based on statistical traffic features.	Exploitation of simpler, often linear or rule-based, decision boundaries.
Deep Learning (DL)	DNN/MLP, LSTM, RNN	Learning complex, non-linear, and temporal patterns in high-volume traffic.	High sensitivity to small, gradient-based perturbations (brittleness).
Advanced/Hybrid	GNN, LSTM-CNN, SNN	Specialized analysis of graph structures (GNN) or hardware fingerprints (LSTM-CNN).	Exploitation of novel data representations (e.g., graph structure) and architectural features.
Federated Learning (FL)	Distributed DNNs/CNNs	Privacy-preserving, decentralized model training on edge devices.	Manipulation of the training process via poisoning attacks (corrupted model updates) from malicious clients.

2.2 Taxonomy of Adversarial Attacks (RQ2)

A taxonomy is required for maintaining the categorization of various threats. We categorize attacks along three axes according to the synthesized literature, which is consistent with established reports including NIST [25].

2.2.1 Classification by Attacker's Knowledge

- White-Box: The attacker has full knowledge of the target model and its architecture, parameters and gradients [25]. This is the most powerful setting, frequently employed for measuring robustness [26], and models realistic threat from an insider [6], [27].
- Black-Box: The adversary has no information of the detail inside and communication by query (oracle) [25]. This is the most probable case of external attackers [26].
- Gray-Box: Adversary has incomplete knowledge, like the set of features employed but no the exact values for the model parameters (e.g., [6]).

2.2.2 Classification by Attack Phase (Lifecycle Stage)

2.3 Evasion (Inference-Time) Attacks

The most common type of attacks in the context of IoT botnet detection and hence also studied quite comprehensively [7], [28], [29]. The attack is launched after the model has been trained and deployed. The adversary aims to synthesize a malicious sample (e.g., network flow) so that the detector wrongfully classifies it as BENIGN [7], [29]. These are attacks that manipulate the integrity of the training process [7], [28], [29]. Adversary poisons the training set with malign samples in advance, before deploying a model in order to compromise the learning process or implant a backdoor [7]. This threat is of particular concern in distributed systems such as Federated Learning [24].Targeted: The goal of the adversary is to misclassify an input for the model into a desired wrong class[29]. Non-Targeted: (Indiscriminate) - The adversary intends the decision on X to be an incorrect answer of any class rather than its original one [29].

The main evasion mechanisms reviewed in the literature are summarised in Table II and we consider these primary against operational botnet detectors. These methods have shifted from gradientbased techniques, which have often been borrowed directly from computer vision [30] to more complex "realistic" attacks constructed to operate within the constrained environment of network security [13], [14].

Table 2; Classification Of Key Evasion Attack Techniques (Rq2)

Attack Technique	Knowledge	Principle/Methodology	Key Characteristic/Efficacy
FGSM (Fast Gradient Sign)	White-Box	Gradient-Based: A fast, one-shot attack that adds a perturbation in the direction of the loss function's gradient sign.	Computationally fast; effective against simple models but less so against robust defenses.
BIM/PGD	White-Box	Gradient-Based: Iterative versions of FGSM that apply smaller perturbations multiple times. PGD adds a random start point.	More potent than FGSM: PGD is a standard for robustness evaluation.
C&W (Carlini & Wagner)	White-Box	Optimization-Based: An iterative attack that uses optimization to find the minimum perturbation required for misclassification.	Highly potent and effective, even against defenses like defensive distillation, but computationally very expensive.
JSMA	White-Box	Gradient-Based (Saliency): Calculates the Jacobian matrix to identify the most salient features and perturbs only those.	Creates sparse and subtle attacks by modifying minimal features.
Binary Diversification	Black-Box	Realistic (Problem-Space): Evades malware detectors by modifying the binary file (e.g., Chunker, Disguiser) to alter its static features/hash while preserving functionality.	High practical relevance; evades static analysis without needing gradients.
Explainability Attack	Black-Box	Realistic (Query-Based): Uses XAI tools (e.g., SHAP) to query the model and identify the most influential feature (e.g., frame.len), then modifies only that feature.	Highly effective (e.g., 100% ASR), gradient-free, and requires very few queries.

2.4 Taxonomy of Defense Strategies (RQ3)

Various defensive strategies have been suggested to counter the identified threats. We categorize these defenses according to their main mode of operation, in line with prior work [25], [26].

2.4.1 Proactive Defenses (Model Hardening):

These methods look to create inherently robust models before attacks have been observed [28], [29].

- Adversarial Training (AT): AT is the most prevailing and empirically strong defense mechanism [26], [28], [34], [37]. The idea is to "vaccinate" the model by adding examples to training dataset during training that are adversarial (often produced using some form of PGD). This makes the model to generalize more and learn features which are robust.forces 21] adversarial perturbations, etc, restricting the complexity of the join space as a regularization strategy [29].
- GANs-based Augmentation: GANs can be employed to produce realistic synthetic botnet flows in order to enhance AT, and help mitigate low data volume and class imbalance. This concept, named Botshot, has the ability to proactively harden the detector [38].
- Model Changes: These may be architectural by nature. Defensive Distillation: At the core of Defensive Filtration is a defense mechanism called Defensive Distillation, which tries to smooth out the decision boundary the model has learned by training a "student" model on probabilities ("soft labels") outputted by a "teacher" model [28], [29]. Strong Architectures, such as SNN [22], also exhibit a higher intrinsic resistance to gradient-based adversaries compared to standard FNN

2.4.2 Reactive Defenses (Detection at Inference)

These techniques serve as a spam filter at deployment-time, to try to reject adversarial inputs before they are fed through the classifier [28], [29].

- Input Preprocessing: Low cost techniques to "denoise" or "purify" inputs by removing adversarial perturbations. Typical techniques include Feature Squeezing (bit-depth reduction) and randomization [28], [29].
- Adversarial Detection: Adversarial inputs can be detected by applying a separate anomaly detector or secondary classifier to discriminate between benign and adversarial inputs [28], [29].

2.4.3 Security and Privacy in FL

To defend against poisoning attacks in the FL setting, protection mechanisms secure the aggregation on the server-side. The first line of defense is Robust Aggregation where statistical algorithms (e.g., Trimmed Mean, Median) are employed instead of average for the detection and removal of malicious model updates contributed by compromised IoT clients [23], [24].

2.4.4 XAI as a Defence

Novel XAI tools (e.g., LIME, SHAP) are being recognized as a defense element. For making the model's decision transparent, XAI can be employed defensively in pinpointing why a certain model is sensitive (i.e., it depends on

non-robust, spurious features), thus enabling developers to deploy and validate more robust models [40]. Table III presents comparison on these defense categories concerning the aims, merits, and demerits as surveyed from the literature.

Table 3; Comparative Analysis of Defense Strategies (Rq3)

Defense Category	Key Technique	Primary Objective	Key Advantage	Key Limitation / Weakness
Proactive	**Adversarial Training**	Harden the model by training on adversarial examples.	Most empirically effective defense for increasing intrinsic robustness.	Computationally expensive; can reduce accuracy on clean data; resource-heavy for IoT.
Proactive	**Defensive Distillation**	Smooth the model's decision boundary to hide gradients.	Effective against simple gradient-based attacks.	Ineffective against strong optimization attacks (e.g., C&W); can cause training instability.
Reactive	**Input Preprocessing**	"Denoise" or "purify" the input before classification.	Lightweight, fast, and model-agnostic; suitable for resource-constrained devices.	Less robust; can be easily bypassed by an adaptive attacker.
Reactive	**Adversarial Detection**	Identify and reject adversarial inputs as anomalies.	Can potentially detect unknown (zero-day) attack patterns.	The detector itself can be attacked; may suffer from high false positive rates.
FL-Specific	**Robust Aggregation**	Secure the global model from malicious client updates.	Directly mitigates poisoning attacks in a distributed FL environment.	Can slow model convergence; may filter legitimate atypical updates.
Meta-Defense	**XAI for Defense**	Understand why a model is vulnerable to build better models.	Provides deep insight into model flaws (e.g., spurious correlations).	"Double-edged sword"; can also be weaponized by attackers to find vulnerabilities [36], [40].

2.5 Evaluation Frameworks Analysis (RQ4)

For analysis of attacks and defenses, the methods carry outare crucial to the validity of results.

2.5.1 Evaluation Metrics

Common classification metrics are adopted for a baseline and performance degradation: Accuracy, Precision, Recall (TPR), F1-Score [17], [18], [19], [41]. But in order to measure adversarial impact directly, there is one most popular metric – Attack Success Rate (ASR), which represents the proportion of adversarial samples that successfully attack the model [17], [18], [42].

2.5.2 Datasets

- The dataset selection is also a controversial issue in the literature, as summarized in Table IV.
- General NIDS Datasets: A good deal of research continues to use general, outdated and nondomain specific NIDS datasets such as NSL-KDD (from KDD Cup 99), UNSW-NB15, and CIC- IDS-2017/2018 [42], [43]. The application of these older datasets such as NSL-KDD is questioned, for the reason that it does not represent real modern IoT traffic or botnet characteristics [17], [32].
- IoT-Specific Datasets - In the pursuit of realism, recently works are adopting IoT-specific datasets like BoT-IoT, N-BaIoT and IoT-23 [44].
- Specialized Datasets: Some researchers use customize datasets for particular purposes, including MQTT dataset for protocol specific attack [45], malware binary datasets like CUBE-MALIOT [34] and hardware fingerprint dataset such as LwH Bench.

Table 4; Analysis Of Common Datasets Used In Aml For Iot Research (Rq4)

Dataset Category	Representative Datasets	Typical Use Case	Analysis of Relevance / Limitations
General NIDS	NSL-KDD, UNSW-NB15, CIC-IDS-2017/2018	Baseline NIDS evaluation.	**Criticism:** Often outdated (NSL-KDD) and not specific to IoT traffic protocols or device behaviors.
IoT-Specific	BoT-IoT, N-BaIoT, IoT-23	IoT botnet/anomaly detection.	**High Relevance:** Contains realistic traffic from actual IoT devices. Preferred for modern, context-aware analysis.
Specialized	CUBE-MALIOT, LwHBench, MQTT Datasets	IoT Malware Analysis, Hardware Fingerprinting, Protocol-Specific Analysis.	**High Specificity:** Required for non-NIDS detection vectors (e.g., static malware analysis, hardware-level attacks).

3. RELATED WORKS

The space of AML in the context of IoT botnet detection stands out as a crossroads between several mature research domains. It is important to locate the contribution of this review into a map that includes studies and seminal on which it is based, clarify its focus and highlight what research gap found in this paper was addressed. The relevant literature can be generally classified as:

1. General articles about AI based IoT botnet detection.

2. Overview of AML attacks and defenses.
3. Targeted surveys on the junction of AML and cyber security.

3.1 Analysis of Existing Surveys

3.1.1 Overview of AI-based IoT Botnet Identification in General

There exist a rich literature of papers on AI/ML models' review for IoT security, yet the focus has always been, detection performance and accuracy; not adversarial robustness. Surveys by Lefoane et al. [1], Szymoniak et al. [2], and Kumar et al. [46] present an in-depth taxonomy of threats for IoT and how AI can be used for finding them. Also, the researches of Mohammed & Alothman [4] and Ahmed & Abdullah [5] review the performance of several ML algorithms in NIDS per se. Although these papers have taken adversarial attacks as a new challenge, AML isn't the main research topic of them. They are the benchmark for what models are subject to detection, but they do not analyze in a systematic way how those models specifically break under adversarial scenarios.

3.1.2 General Surveys on Adversarial Machine Learning

A second type of the literature reviews the AML domain in general, presenting high-level taxonomies on attacks and defenses. Works by Chakraborty et al. [28], Abomakhelb et al. [29], and Malik et al. [32] present an extensive taxonomy of attack vectors (e.g., Evasion, Poisoning) and defence strategies (e.g., Adversarial Training, Defensive Distillation). But the major shortcoming of these overall surveys with respect to this paper is that they depend heavily on computer vision (CV) [30]. The concepts of attacks such as FGSM, PGD, etc. are typically introduced with an example based on image data. While being of foundational importance, they have not considered the domain-specific constraints and interdependencies at semantic level especially for traffic data that has structure such as network communication data as obviously highlighted in (Section 2) but is important for realistic threat assessment in NIDS [13], [14].**]**.

3.1.3 Specialized AML Surveys for Cybersecurity

Review Then, more recently, reviews have started to close this gap by concentrating specifically on AML in the cybersecurity field. Hasan et al. [33] presented a taxonomy of AML attacks on DL-based NIDS and Kumar et al. noted recent attacks on AI-equipped IoT malware detectors. These works are particularly relevant in the investigation of AML applied to the cybersecurity context. Concurrently, Aloraini et al. [6] Taking a more novel approach to this perspective, DarkMatter proposed analyzing AML in IoT considering an "insider threat" (white-box/gray-box) stance.

3.2 Identifying the Research Gap

The most closely related work are also the following systematic reviews [9–11] by Al-Masqari and coworkers. [43] concerning AML in IoT IDS' and the 2024 survey of Al-Imam et al. [48] in AML for ML-driven botnet detectors. These are the essential base of these works." Nevertheless, with the field progressing at a very fast pace, such a review is required not only classifying these building blocks of NSs but also critically digesting their findings based on the most recent developments and urgent cross-domain challenges. This systematic review meets a clear and urgent need as it delivers an integrated (i.e. 2020-25) synthesis on the whole of the following that compares to no other:

1. General And Specific Focus: It studies general NIDS [33],[43] and host-based malware detection[34] (of IoT) attacks facilitating both these categories, that are mainly addressed separately.
2. Attack Realism for Emphasis: It is centrally focused on the "Reality Gap" [13], [14], [37] by considering sober problem-space attacks (e.g., binary diversification [34], 35], XAI-based attacks [36]) instead of candid fun feature-space attacks.
3. Inclusion of Distributed Models: It covers the novel Apocalyptic (ni-ApoL) and Emerging threat vector which is FL attacks and defenses in distributed IoT environment [23], [24].
4. The Resource-Constraint Challenge: It shifts the focus of all discussions to the IoT's resource issue, and critically compares practicability regarding computationally expensive security measures (Adversarial Training being an example) with lightweight substitution options [1], [2], [41]. In summary, whilst existing works (Table V) have identified the building blocks, we take a comprehensive view on why and how these components are interconnected in the concrete, confined setting of IoT botnet detection, hence answering all five RQs in an integrated manner

Table 5; Comparison of Related Surveys and Identification of Research Gap

Representative Survey(s)	Primary Focus	Scope	Key Limitation / Gap Addressed by This Review
Lefoane et al. [1] Kumar et al. [46]	AI for Botnet Detection	IoT Security	**Lacks AML Depth:** Focuses on detection accuracy; adversarial attacks are mentioned as a challenge but not systematically analyzed.
Chakraborty et al. [28] Abomakhelb et al. [29]	General AML Taxonomies	General AI / Computer Vision	**Lacks Domain Specificity:** Does not address the "Reality Gap" or unique constraints (problem-space) of IoT network data.
Hasan et al. [33] Kumar et al. [47]	AML in NIDS / Malware	Cybersecurity (General)	**Scope is Partial:** Addresses NIDS or Malware, but not a synthesized view. Does not deeply integrate IoT resource constraints or FL.
Al-Masqari et al. [43] Al-Imam et al. [48]	AML in IoT NIDS / Botnets	IoT Security (Botnets)	**Foundational Survey:** Provides a necessary baseline.
This Systematic Review	**AML in IoT Botnet Detectors**	**IoT Security (Botnets)**	**Fills the Gap by:** Synthesizing NIDS and Malware attacks, emphasizing Attack Realism [13], [14], [37], analyzing FL vulnerabilities [23], [24], and centering the entire discussion on the Resource-Constraint Dilemma [1], [2], [41].

4. DISCUSSION AND ANALYSIS (RQ5)

As such, to the best of our knowledge, this systematic review has successfully aggregated and categorized the research field (RQ1-RQ4) by recognizing which models are targeted and which attack vectors, defense strategies and evaluation frameworks have been considered. To answer RQ5, this section offers a discussion and critique of these findings, questioning the major challenges that shape this area and identifying the key future research directions that must be addressed to create reliable, scalable, and secure AI-powered IoT botnet detectors.

4.1 Synthesis of the State-of-the-Art

One emergent theme in the synthesized literature was a discernible and continual "arms race" dynamic [28]. The hierarchy in research has developed in a cycle:

1. Defenders make better AI/ML models (e.g., DNNs, LSTMs) for addressing the shortcomings of static signatures [1, 4, 19].
2. Adversaries fight back by showing vulnerabilities of these models with gradient-based attacks (PGD, C&W) [18, 26].
3. Defenders on the other hand defend with proactive defenses such as Adversarial Training (AT) [26, 37].
4. Attackers, on the other hand, advance second-order attacks through more advanced black-box and realistic problem-space attacks (e.g., binary diversification [36], XAI-based attack [37]) to evade such defense mechanisms [34].

This development reflects the fact that AML is not something that is static, but an increasing and ongoing threat. Although the area has been growing, its application to IoT botnet detection is rather in early stage [1], [46]. Recommended Reading: After carefully reviewing literature summarized above, We are of the view that the field has a number of unsolved, fundamental subproblems that distinguish (theoretical) research from deployable security.

4.2 Open Challenges and Research Directions

The "Reality Gap" -Problem-Space vs.Feature-Space Attacks Optional subsection inch of left column optional inchesThe central issue that causes the previous to traditional adversaries gap(\"""\)) exist is denoted as thefeature-space attack.This occurs whenever communication and computation deviates from themidon-targetdomain. The major obstacle of these method is the reality gap [13], [14], [37]. Many influential AML contributions, which are based on computer vision task [30], perform attacks in the feature space (i.e., perturbing the numerical vector that is given as input to the model). But, in the IoT network security domain, data is well structured and controlled by strict domain constraints [13]. An attack that modifies a feature-space vector by using 6.5 instead of 6 (the protocol number for TCP) is semantically senseless and would have been discarded well before reaching a detector [13], [37]. These “unrealistic” attacks only offer a poor measure of robustness. Recent work highlights that an attack must operate in problem-space (e.g., the real packet or binary file) and satisfy all semantics, timing, and structure requirements for it to be valid [13], [14], [37], [42].

4.2.1 The IoT Resource-Constraint Dilemma

Fundamentally, the strongest countermeasures are inherently in conflict with the IoT reality. The most effective adversarial defence known so far is Adversarial Training (AT), but it is expensive due to the computation and training of complex models against large-scale adversaries [26], [41]. This prevents it from being used on the device level in power-constrained IoT environments (e.g., constrained CPU, memory, and power) [1], [2], [41]. This challenge often leads to having to fall back on weaker, more reactive sdefenses (e.g., input pre-processing) [28] or placing all the defensive eggs in one basket by requirements concentrated gateways that can also become a crowd-ed bottleneck.

4.2.2 Performance Trade-offs

The pros and cons of robustness Robustness is not a free lunch. The trade-offs that are commonly mentioned in the literature:

1. Accuracy vs. Robustness: Adversarial Training improves robustness against attacks, at the cost of a loss in accuracy on clean benign data [26], [28], [41]. This "robustness-accuracy trade-off" imposes a challenging decision on system developers.
2. Normalization versus Robustness: A key insight from a study [22] was the fact that contrary to the conventional relationship between feature normalization (a standard data preprocessing step used for increasing accuracy), and adversarial robustness It turned out, normalization could exacerbate vulnerability by smoothing of the models loss surface making it easier for an adversary to find an explorable gradient..

4.2.3 Robustness Evaluation Challenges

The robustness is however difficult to judge.

- Outdated Datasets: This domain is largely dependent on deprecated (e.g., NSL-KDD) and non-IoT-specific (e. g., CIC-IDS) datasets [17], [32] which is impractical due to the fact that today's botnet tactics have evolved, not to mention IoT-specific protocols [45].
- Adaptive Attackers: Unlike most defenses which are analyzed only against static, known attacks. They are not evaluated against adaptive attackers who know the defense and design their attack to sidestep it [18], [30]. This gives you a false impression of security.

4.3 Future Research Directions

These issues refer to some key trends of the future efforts that are needed.

4.3.1 Secure and Resilient FL

FL is a natural architectural choice for IoT, maintaining data privacy and decentralising the computation burden [23]. However it is particularly vulnerable to poisoning attacks [24]. The ongoing research would then undertake to design computationally efficient and secure robust aggregation algorithms [24] along with data validation scheme on the device side (on-device data validation) in order to construct practical FL frameworks that can be deployed securely over less powered IoT devices [23].

4.3.2 The Two Sides of Explainable AI (XAI)

XAI brings a `double-edged sword' [4].

- XAI for Defense: In future, one should utilize XAI tools (e.g., LIME, SHAP) not only as transparency tools but also as a debugging tool to understand in failing cases why models fail. This enables researchers to spot spurious correlations (non-robust features) and construct inherently more robust models [40].
- textbf{XAI-Aware Defenses:} On the other hand, researchers should also anticipate that adversaries are going to weaponize XAI and explainable models (as whowns in Explainability-Based black box Attacks [36]). Future defenses must be ``explanation-aware": not just protect their predictions but also (all-or-nothing) defend explanations themselves.

4.3.3 Proactive Defense with Generative AI (GenAI)

The current defense model is mainly passive. And generative AI (such as GANs) may enable proactive defense [39]. Rather than learn only from previous attacks, future systems could use GANs to augment training data with new unseen attack scenarios. Such a practice, and in this case the method of "Botshot" [38] for data augmentation, can make detectors more robust to zero-day adversarial threats and future ones.

4.3.4 Realistic, Constrained-Aware AML

At the end, we need the domain to quickly evolve toward more secure and practical defenses for IoT. There needs to be a shift in future research away from unrealistic, feature-space attacks and towards purely developing and testing against problem-space attacks that conform to domain constraints [13], [14], [37]. This includes broader investigations related to malware binary diversification [34] and packet-level manipulation that preserves valid protocol states [13], [37], [42].

4.3.5 Benchmark, New Datasets and Standardized Benchmarks

Finally, the community needs new and realistic large-scale benchmark datasets designed solely for AML in IoT [17], [32]). These datasets should contain labeled benign traffic as well as more recent botnet traffic and—most importantly—predefined, realistic adversarial samples with known domain constraints. It will allow for a standardized, reproducible and meaningful comparison between defense strategies [45].

5. CONCLUSION

This systematic reviewing of literature generated an extensive taxonomy and in-depth analysis that summarizes the state-of-the art adversarial attacks and defenses on AI-driven IoT botnet detectors. Our survey of the recent literature (2020–2025) validates that the threat model is dynamic. RQ1: Motivated By the fact that threat models in evolving paradigms present new challenges, we discovered that the attack landscape no longer a↵ects only classic ML models (e.g., SVM and RF), but is increasingly targeting more sophisticated Deep Learning (DL) architectures (DNNs and LSTMs) or distributed Federated Learning (FL) frameworks as adopted by recent modern IoT security. For RQ2, we constructed a high-dimensional attack taxonomy that underscores the pivotal transition from theoretical gradient-based evasion attacks (e.g., PGD) towards practical and powerful problem-space attacks like binary diversification and XAI-based methods which are more menacing in real applications. Regarding RQ3, we differentiated between proactive (model hardening) and reactive (detection) defenses. Adversarial Training (AT) was recognized as the most widely deployed proactive defense and focus on robust aggregation algorithms were deemed critical to secure FL environments. Finally, the results of our analysis of RQ4 showed that there is a significant methodological gap in the field, with many studies continuing to rely on old non-IoT-specific datasets and commonly not evaluating attacks given realistic operational constraints. This review confirms once again the crucial role played by the adversarial threat in IoT security. The fact that AI-based detectors can be fragile is no small academic point; it represents a key vulnerability. With the cyber-physical characteristic of the IoT systems, model failures are not just virtual incidents (e.g., crash of an AI-driven personal assistant service), but cascade into real-world consequences as well (e.g., plane crash due to malicious manipulation). therefore, providing robust and reliable detectors against adversarial attacks is critical to secure and trust our more connected critical infrastructure.

According to the critical challenges identified in our study (RQ5), we suggest several guidelines to researchers and developers:

- Prefer Realism: In the future, we should not entertain unrealistic feature-space attacks -- We must develop and test only problem-space attacks that adhere both to semantics and operations of IoT networks.
- Solution: Tackle Resource Limitations Novel, lightweight defenses (e.g., efficient input transformations, optimized detectors) that can be efficiently implemented for on-device deployment in resource-limited IoT contexts are crucial.
- Normalize Evaluation: The field needs to normalize over recent, IoT-specific datasets and the application of adaptive attack methodologies during evaluation; we move away from antiquated benchmarks.
- Investigate Hybrid and Proactive Approaches: Most promising R&D areas include hybrid defense solutions, Secure Federated Learning (FL), defensive use of Explainable AI (XAI), as well as application of Generative AI (GenAI) for proactive defenses against ‘unknown’ future threats.

In summarythe "arms race" of AI-based detection and adversarial evasion in IoT is still evolving. And that the future of secure IoT won't be in some single bullet-proof defense, but in building defenses on top of defenses that are more grounded and bounded to reality while bridging the rather wide gap from theoretical AML work to the practical realities of a typical IoT environment.

REFERENCES

[1] M. Lefoane, I. Ghafir, S. Kabir, and I.-U. Awan, “Internet of Things botnets: A survey on Artificial Intelligence based detection techniques,” J. Netw. Comput. Appl., vol. 236, p. 104110, 2025.
[2] S. Szymoniak, J. Piątkowski, and M. Kurkowski, “Defense and Security Mechanisms in the Internet of Things: A Review,” Appl. Sci., vol. 15, no. 2, p. 499, Jan. 2025.

[3] R. Anne W, G. Kirubavathi, and U. K. Sridevi, "Detection of IoT Botnet using Machine learning and Deep learning Techniques," Research Square, Mar. 2023.
[4] M. S. Mohammed and H. A. T. Alothman, "Using Machine Learning Algorithms in Intrusion Detection Systems: A Review," Tikrit Journal of Pure Science, vol. 29, no. 3, pp. 63–74, Jun. 2024.
[5] M. Ahmed and Q. Abdullah, "Network Intrusion Detection Systems: Machine Learning-Based Attack and Remedy Strategies - A Review," Al-Salam Journal for Engineering and Technology, vol. 4, no. 2, pp. 11–29, 2025.
[6] Y. L. Khaleel, M. A. Habeeb, and H. Alnabulsi, "Adversarial Attacks in Machine Learning: Key Insights and Defense Approaches," Applied Data Science and Analysis, vol. 2024, pp. 121–147, Aug. 2024.
[7] P. M. Sánchez Sánchez, A. Huertas Celdrán, G. Bovet, and G. Martínez Pérez, "Adversarial attacks and defenses on ML- and hardware-based IoT device fingerprinting and identification," Future Gener. Comput. Syst., vol. 152, pp. 30–42, 2024.
[8] F. Aloraini, A. Javed, O. Rana, and P. Burnap, "Adversarial machine learning in IoT from an insider point of view," J. Inf. Secur. Appl., vol. 70, p. 103341, 2022.
[9] O. Ibitoye, O. Shafiq, and A. Matrawy, "Analyzing Adversarial Attacks Against Deep Learning for Intrusion Detection in IoT Networks," arXiv preprint arXiv:1905.05137v1, May 2019.
[10] A. T. Olutimehin et al., "Adversarial Threats to AI-Driven Systems: Exploring the Attack Surface of Machine Learning Models and Countermeasures," J. Eng. Res. Rep., vol. 27, no. 2, pp. 341–362, 2025.
[11] S. Sharma and Z. Chen, "A Systematic Study of Adversarial Attacks Against Network Intrusion Detection Systems," Electronics, vol. 13, no. 24, p. 5030, 2024.
[12] A. Namvar, "Adversarial Machine Learning in IoT: Vulnerability Analysis and Robustness," Ph.D. Thesis, School of Computer Science and Engineering, The University of New South Wales, Oct. 2023.
[13] H. Mohammadian, A. H. Lashkari, and A. A. Ghorbani, "Evaluating Deep Learning-based NIDS in Adversarial Settings," in Proc. 9th International Conference on Computer and Knowledge Engineering (ICCKE), 2022.
[14] Y. Wang et al., "Adversarial Attacks and Defenses in Machine Learning-Powered Networks: A Contemporary Survey," arXiv preprint arXiv:2303.06302v1, Mar. 2023.
[15] G.-Y. Lin, P.-Y. Wang, S.-M. Cheng, and H.-M. Lee, "Improving Robustness in IoT Malware Detection through Execution Order Analysis," ACM Trans. Embed. Comput. Syst., Aug. 2024. doi: 10.1145/3684278.
[16] F. R. Mughal et al., "Adaptive federated learning for resource-constrained IoT devices through edge intelligence and multi-edge clustering," Sci. Rep., vol. 13, no. 1, p. 20038, 2023.
[17] Z. Feng, "Federated Learning Security Threats and Defense Approaches," in Proc. 2023 2nd International Conference on Computer Science and Innovative Computations (CSIC), Highlights in Science, Engineering and Technology, vol. 85, pp. 120–127, 2024.
[18] A. Goel, A. Sharma, and D. Kejriwal, "IoT Device Authentication Using Adversarial Machine Learning," Journal of Advances in Developmental Research (IJAIDR), vol. 15, no. 12, Sep. 2024.
[19] A. Vassilev, A. Oprea, A. Fordyce, and H. Anderson, "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations," NIST AI 100-2e2023, U.S. Department of Commerce, Jan. 2024.
[20] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defences," CAAI Trans. Intell. Technol., vol. 6, no. 1, pp. 25–45, 2021.
[21] J. C. Costa, T. Roxo, H. Proença, and P. R. M. Inácio, "How Deep Learning Sees the World: A Survey on Adversarial Attacks & Defenses," arXiv preprint arXiv:2305.10862v1, May 2023.
[22] J. Sándor, R. Nagy, and L. Buttyán, "Increasing the Robustness of a Machine Learning-based IoT Malware Detection Method with Adversarial Training," in Proc. 2023 ACM Workshop on Wireless Security and Machine Learning (WiseML '23), Guildford, United Kingdom, Jun. 2023.
[23] M. M. Alani, A. Mashatan, and A. Miri, "Adversarial Explainability: Utilizing Explainable Machine Learning in Bypassing IoT Botnet Detection Systems," arXiv preprint arXiv:2310.00070v1, Sep. 2023.
[24] A. Abomakhelb, K. A. Jalil, A. G. Buja, A. Alhammadi, and A. M. Alenezi, "A Comprehensive Review of Adversarial Attacks and Defense Strategies in Deep Neural Networks," Technologies, vol. 13, no. 5, p. 202, May 2025.
[25] R. H. Randhawa, N. Aslam, M. Alauthman, H. Rafiq, and F. Comeau, "Security Hardening of Botnet Detectors Using Generative Adversarial Networks," IEEE Access, vol. 9, pp. 78278–78294, 2021.
[26] N. Capuano, G. Fenza, V. Loia, and C. Stanzione, "Explainable Artificial Intelligence in CyberSecurity: A Survey," IEEE Access, vol. 10, pp. 93575–93600, 2022.
[27] M. M. Hasan, R. Islam, Q. Mamun, M. Z. Islam, and J. Gaob, "Adversarial Attacks on Deep Learning-based Network Intrusion Detection Systems: A Taxonomy and Review," SSRN Electronic Journal, 2024. [Online]. Available: https://ssrn.com/abstract=4863302
[28] S. Ankalaki et al., "Cyber Attack Prediction: From Traditional Machine Learning to Generative Artificial Intelligence," IEEE Access, vol. 13, pp. 28863–28877, 2025.
[29] K. Barik and S. Misra, "A comprehensive defense approach of deep learning-based NIDS against adversarial attacks," Multimed. Tools Appl., vol. 84, pp. 37745–37791, 2025.
[30] M. B. Mwangi and S.-M. Cheng, "An Adversarial Attack on ML-Based IoT Malware Detection Using Binary Diversification Techniques," IEEE Access, vol. 12, pp. 170940–170953, 2024.
[31] G. Apruzzese, M. Colajanni, and M. Marchetti, "Evaluating the effectiveness of Adversarial Attacks against Botnet Detectors," in Proc. 2018 IEEE 17th International Symposium on Network Computing and Applications (NCA), Cambridge, MA, USA, 2019.
[32] J. Malik, R. Muthalagu, and P. M. Pawar, "A Systematic Review of Adversarial Machine Learning Attacks, Defensive Controls, and Technologies," IEEE Access, vol. 12, pp. 99402–99422, 2024.
[33] I. Debicha, J.-M. Dricot, and B. Piètu, "Adversarial Training for Deep Learning-based Intrusion Detection Systems," in Proc. The Sixteenth International Conference on Systems (ICONS 2021), 2021, pp. 40–45.
[34] E. Sanchez, L. Clark, S. Ramirez, A. Lewis, A. Robinson, and D. Esther, "Adversarial Attacks and Defenses in IoT Networks," Article, Nov. 2024. [Online]. Available: https://www.researchgate.net/publication/391270645
[35] V. A. Memos and K. E. Psannis, "AI-powered Honeypots for Enhanced IoT Botnet Detection," Presentation at 3rd World Symposium on Communication Engineering (WSCE), Oct. 2020.

[36] V. P. Singh, R. Kumari, and M. Kaur, "Machine Learning for Intrusion Detection System in IoT Environment with Permutation Importance," in Proc. 1st International Conference on AI, IoT, and Next Generation Technologies (ICAINGT 2024), 2024, CEUR-WS.org, Vol-3774, Paper 2.
[37] K. S. Prasad et al., "A two-tier optimization strategy for feature selection in robust adversarial attack mitigation on internet of things network security," Sci. Rep., vol. 15, no. 1, p. 2235, 2025.
[38] Malik, Jasmita et al. "A Systematic Review of Adversarial Machine Learning Attacks, Defensive Controls, and Technologies." IEEE Access 12 (2024): 99382-99421.
[39] A Comprehensive Review of Learning-Based Anomaly Detection Techniques in IoT Security Systems", East Journal of Computer Science, vol. 1, no. 4, pp. 18–27, Sep. 2025.
[40] N. Hasan, Z. Chen, C. Zhao, Y. Zhu, and S. M. R. Islam, "IoT Botnet Detection framework from Network Behavior based on Extreme Learning Machine," in Proc. IEEE INFOCOM 2022 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2022.
[41] J. Vitorino, I. Praça, and E. Maia, "Towards adversarial realism and robust learning for IoT intrusion detection and classification," Ann. Telecommun., vol. 78, pp. 401–412, 2023.
[42] Z. Iqbal, A. Imran, A. U. Yasin, and A. Alvi, "Denial of Service (DoS) Defences against Adversarial Attacks in IoT Smart Home Networks using Machine Learning Methods," NUST Journal of Engineering Sciences, vol. 15, no. 1, pp. 19–25, 2022.
[43] Aparcana-Tasayco, A.J., Deng, X. & Park, J.H. A systematic review of anomaly detection in IoT security: towards quantum machine learning approach. EPJ Quantum Technol. **12**, 112 (2025).
[44] M. M. Hasan, R. Islam, Q. Mamun, M. Z. Islam, and J. Gao, "Adversarial Attacks on Deep Learning-based Network Intrusion Detection Systems: A Taxonomy and Review," SSRN Electronic Journal, 2024. [Online]. Available.
[45] T. Al-Shurbaji et al., "Deep Learning-Based Intrusion Detection System for Detecting IoT Botnet Attacks: A Review," in IEEE Access, vol. 13, pp. 11792-11822, 2025.
[46] S. Kumar and S. Soni, "Botnet Attack Prevention in Internet of Things (IOT) devices Using AI: A Systematic Review," International Journal of Computer Science Trends and Technology (IJCST), vol. 13, no. 2, pp. 67–78, 2025.