Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Adversarial Intelligence: A Systematic Review of AI-Driven Cyber Threat Detection, Adaptive Malware Evasion, and Arms Race Dynamics in Zero Trust Security Environments

Dharma Patel

dharmapatel016@gmail.com

Marwadi University https://orcid.org/0009-0003-6629-1170

Systematic Review

Keywords: Adversarial machine learning, artificial intelligence, cybersecurity, deep learning, evasion attacks, explainable artificial intelligence, federated learning, generative adversarial networks, intrusion detection, malware detection, systematic review, zero trust architecture

Posted Date: July 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10137205/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Adversarial Intelligence: A Systematic Review of AI-Driven Cyber Threat Detection, Adaptive Malware Evasion, and Arms Race Dynamics in Zero Trust Security Environments

Abstract

Artificial intelligence has become central to modern cyber defense, powering intrusion detection, malware classification, and behavioral analytics across enterprise, cloud, and Internet of Things environments. At the same time, the same statistical and generative tools that strengthen detection are being repurposed by adversaries to craft evasive malware and adversarial network traffic, producing a continuous arms race between defenders and attackers. This paper presents a systematic review of 80 peer reviewed articles, preprints, and technical reports published between 2010 and 2026, covering machine learning and deep learning based intrusion detection, generative adversarial network driven evasion and synthesis, federated and privacy preserving detection, explainable artificial intelligence for security, and the emerging role of large language models on both sides of the conflict. The review introduces a unified taxonomy of detection techniques and evasion strategies, traces the co-evolution of attack and defense across four technological generations, and examines how zero trust architecture functions as a governing framework for continuous, AI assisted verification. Comparative tables summarize technique families, benchmark datasets, and defense mechanisms, while bar, pie, and line charts characterize the composition and growth of the reviewed literature. The review concludes that robustness, realistic benchmarking, explainability, and regulatory alignment remain the most significant open challenges, and it outlines concrete research directions for closing the gap between detection accuracy and adversarial resilience.

Keywords: *Adversarial machine learning, artificial intelligence, cybersecurity, deep learning, evasion attacks, explainable artificial intelligence, federated learning, generative adversarial networks, intrusion detection, malware detection, systematic review, zero trust architecture.*

Impact Statement

This review consolidates a fragmented body of research spanning intrusion detection, adversarial malware evasion, and zero trust security into a single coherent account of how artificial intelligence is reshaping both sides of the cyber conflict. Practitioners responsible for securing networks, endpoints, and cloud infrastructure can use the taxonomies, comparative tables, and arms race model presented here to understand why detectors that perform well on static benchmarks often fail against adaptive adversaries, and to anticipate how generative and language model based tools will alter the threat landscape in the near term. Policymakers and standards bodies engaged in artificial intelligence governance, including frameworks such as the NIST AI Risk Management Framework and the European Union Artificial Intelligence Act, can draw on the synthesized evidence regarding explainability, robustness evaluation, and zero trust integration when drafting sector specific guidance for critical infrastructure, healthcare, and financial systems. For the research community, the review identifies concrete gaps, including the scarcity of problem space realizable adversarial benchmarks, inconsistent robustness reporting, and the unresolved trade off between model interpretability and detection accuracy, that can directly inform future experimental design. By making the literature's evolution visible through structured figures and quantitative summaries rather than narrative description alone, this work is intended to shorten the path from awareness of the artificial intelligence driven arms race to actionable defensive and regulatory practice across academic, industrial, and governmental settings.

I. INTRODUCTION

Cyberattacks have grown in volume, automation, and sophistication to the point where manual analysis and static, signature based defenses can no longer keep pace. Industry estimates place the global cost of cybercrime in the trillions of dollars annually, a figure driven in large part by malware, ransomware, and network intrusions that increasingly rely on automated tooling rather than manual exploitation [21]. In response, defenders have adopted artificial intelligence (AI) and machine learning (ML) across nearly every layer of the security stack, from network intrusion detection systems (IDS) and malware classifiers to identity analytics and security orchestration platforms [1], [2]. Deep learning architectures, including convolutional neural networks, long short term memory networks, and transformer based attention models, have demonstrated detection accuracies exceeding 99 percent on widely used benchmark datasets [60], [61], [63], reinforcing the perception that AI represents a durable advantage for defenders.

That perception is incomplete. The same mathematical properties that make machine learning models effective pattern recognizers, namely their reliance on learned decision boundaries over high dimensional feature spaces, also make them susceptible to deliberately crafted inputs that shift a sample just far enough across that boundary to be misclassified. This vulnerability, broadly termed adversarial machine learning, was first documented in computer vision but has since become a central concern in cybersecurity, where the attacker is not a passive source of noise but an active, strategic adversary with a concrete incentive to evade detection [8], [9], [38], [41]. Within the malware domain alone, more than a decade of research has produced gradient based perturbation methods, reinforcement learning guided file modification, and generative adversarial network (GAN) based sample synthesis, each capable of reducing detector accuracy substantially while preserving the malicious functionality of the underlying file [6], [11], [12], [14], [15].

This dynamic is best understood not as a series of isolated vulnerabilities but as a continuous arms race: defenders deploy an AI based detector, attackers probe and profile it, attackers craft adversarial samples that exploit its blind spots, the detector is evaded, the evaded samples are collected, and the detector is retrained or hardened before the cycle repeats [38], [42], [48]. Each iteration of this cycle has historically coincided with a broader shift in the underlying machine learning paradigm, from signature matching, to classical feature based classifiers, to deep neural architectures, and now to generative and large language model (LLM) based tools that can both

synthesize evasive malware and, conversely, parse unstructured threat intelligence and logs at a scale no human analyst could match [45], [49], [54], [58].

Zero trust architecture (ZTA) has emerged as the governing security philosophy intended to contain this dynamic. Rather than assuming that any user, device, or workload inside a network perimeter is trustworthy, ZTA enforces continuous verification of identity, device posture, and behavioral signals before granting least privilege access to any resource [24], [25], [26]. Because continuous verification depends on the ability to detect anomalous behavior in real time, AI based analytics are not merely compatible with zero trust, they are frequently a structural requirement of it [23], [27]. This creates a second layer of the arms race: as zero trust deployments increasingly rely on AI to make access decisions, those same AI components inherit the adversarial vulnerabilities documented in the IDS and malware detection literature, meaning that the robustness of the underlying detector is no longer only a detection accuracy question but also an access control and governance question.

Despite the depth of research on each of these topics individually, comparatively few works connect AI driven detection, adversarial evasion, and zero trust governance into a single analytical frame. Several existing surveys address adversarial attacks on malware classifiers in isolation [6], [8], [9], others focus narrowly on GAN applications to intrusion detection [19], [22], and still others review zero trust adoption without engaging deeply with the adversarial robustness of the AI components that zero trust depends on [24], [25]. This fragmentation makes it difficult for practitioners and policymakers to reason about the system level consequences of deploying AI across the security stack, particularly as generative and language models begin to compress the time and skill required to mount an effective evasion campaign.

This paper addresses that gap through a systematic review structured around three research questions. First, what are the principal families of AI based techniques used for cyber threat detection, and how have they evolved across classical machine learning, deep learning, generative, federated, and language model based paradigms. Second, what are the corresponding families of adversarial evasion and malware generation techniques that have been developed to defeat these detectors, and what threat models do they assume? Third, how does zero trust architecture, as a governing security framework, interact with the robustness and explainability properties of the AI components it depends on, and what does the literature suggest about closing the resulting governance gap.

The remainder of this paper is organized as follows. Section II positions this review against prior surveys and summarizes the methodology used to identify and screen the 80 sources synthesized here. Section III presents a taxonomy of AI driven detection techniques. Section IV examines adversarial malware evasion and adaptive attacks. Section V develops a conceptual model of the arms race dynamic linking detection and evasion across four technological generations. Section VI analyzes zero trust architecture as a governing framework. Section VII reviews explainability and trust considerations. Section VIII summarizes benchmark datasets and evaluation practices. Section IX discusses open challenges and future research directions. Section X discusses the synthesized findings, Section XI states the limitations of this review, and Section XII concludes the paper.

II. RELATED WORK AND REVIEW METHODOLOGY

A. Positioning Against Prior Surveys

Several survey efforts have addressed pieces of the landscape covered in this review. Ahmad et al. [1] and Abdulganiyu et al. [2] provide systematic overviews of machine learning and deep learning approaches to network intrusion detection, but neither engages substantially with adversarial robustness. Aryal et al. [6] and Ling et al. [8] offer encyclopedic treatments of adversarial evasion attacks against malware classifiers, yet they do not connect these attacks to the zero trust governance frameworks within which production detectors increasingly operate. Surveys of generative adversarial networks in intrusion detection [19], [22] catalog architectures and applications but treat the technology primarily as a detection or augmentation tool rather than as a dual use capability that also empowers attackers. Reviews of zero trust architecture [24], [25] document adoption patterns and implementation challenges across domains such as healthcare, cloud computing, and the Internet of Things (IoT), but largely abstract away the underlying AI components, treating continuous verification as a policy mechanism rather than a machine learning system with its own attack surface. Apruzzese et al. [38] and Akhtar et al. [41] come closest to a unifying perspective, framing adversarial machine learning explicitly through a cybersecurity lifecycle lens, but their scope predates the rapid expansion of large language model applications in both offensive and defensive security tooling [45], [49], [54].

This review differs from prior work in three respects. First, it explicitly integrates detection, evasion, and governance into a single thematic structure rather than treating them as separate literatures. Second, it introduces a four generation escalation model that traces how each shift in detection paradigm, from signature matching through deep learning to generative and language model based tools, has been met by a corresponding shift in evasion strategy. Third, it analyzes zero trust architecture specifically through the lens of the robustness and explainability properties of the AI components it depends on, rather than as a purely organizational or network engineering framework.

B. Search Strategy and Eligibility Criteria

The review followed a structured, PRISMA inspired process illustrated in Fig. 1. Records were identified through keyword searches across IEEE Xplore, ACM Digital Library, SpringerLink, ScienceDirect, MDPI journals, and arXiv, using combinations of terms drawn from five thematic clusters: intrusion detection and network anomaly detection, adversarial machine learning and malware evasion, generative adversarial networks in security, zero trust architecture, and explainable artificial intelligence in cybersecurity. Table I summarizes the search strings and databases used for each cluster.

Records were screened by title and abstract to remove duplicates and clearly off topic results, including non English language publications, opinion pieces without methodological content, and works addressing adversarial machine learning exclusively in non security domains such as computer vision benchmarks with no cybersecurity application. The remaining full text articles were assessed for eligibility against three inclusion criteria: the work must involve an artificial intelligence or machine learning component, the work must address a cybersecurity detection, evasion, or governance task, and the work must report sufficient methodological detail to support synthesis, such as a described architecture, dataset, or evaluation protocol. Eighty sources met these criteria and form the basis of the synthesis presented in this review, spanning publication years from 2010 to 2026 and including peer reviewed journal articles, peer

reviewed conference proceedings, and arXiv preprints that had not yet completed peer review at the time of writing but were retained where they represented the most current treatment of a fast moving subtopic such as large language model assisted threat detection.

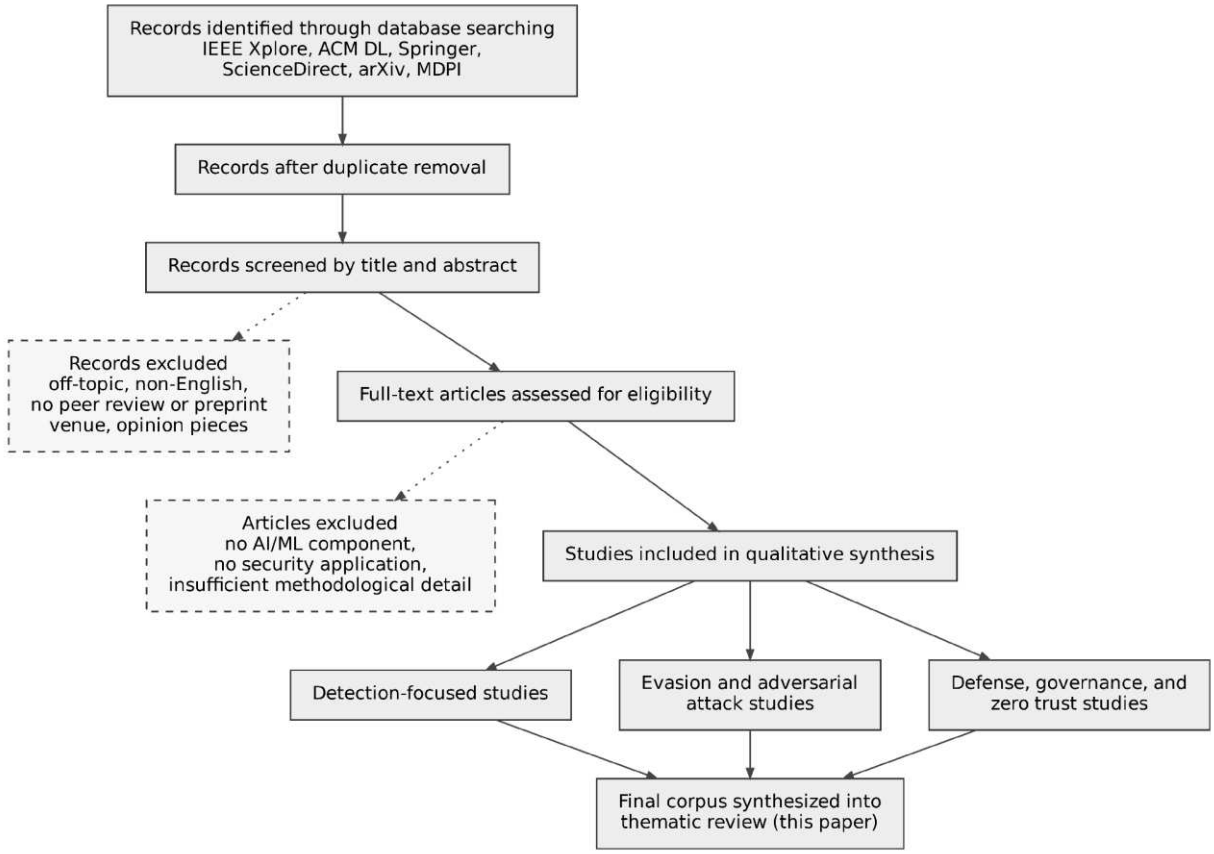


Fig. 1. PRISMA inspired a flow of record identification, screening, and inclusion for this systematic review.

C. Synthesis Approach

Because the included literature spans heterogeneous methodologies, ranging from purely conceptual taxonomies to empirical evaluations on benchmark datasets, this review adopts a thematic synthesis approach rather than a quantitative meta-analysis. Sources were grouped into four overlapping thematic clusters, detection focused studies, evasion and adversarial attack studies, defense and robustness studies, and governance and zero trust studies, with many studies contributing to more than one cluster. Within each cluster, recurring technique families, datasets, and evaluation metrics were extracted and tabulated to support the comparative tables presented in Sections III through VIII. The bar, pie, and line charts presented later in this paper characterize the composition and temporal growth of the reviewed corpus itself, and they should be read as a description of this review's literature sample rather than as an external bibliometric indicator of the field as a whole.

TABLE I. SEARCH STRATEGY AND DATABASES BY THEMATIC CLUSTER

Thematic Cluster	Example Search Terms	Databases Searched
Intrusion detection and anomaly detection	network intrusion detection; machine learning IDS; deep learning anomaly detection	IEEE Xplore, ACM DL, ScienceDirect, MDPI
Adversarial ML and malware evasion	adversarial malware; evasion attack; adversarial examples cybersecurity	arXiv, IEEE Xplore, ACM DL, ScienceDirect
Generative adversarial networks in security	GAN intrusion detection; adversarial traffic generation; MalGAN	SpringerLink, arXiv, PMC
Zero trust architecture	zero trust architecture; continuous verification; AI zero trust	PMC, MDPI, SpringerLink
Explainable AI in cybersecurity	explainable AI IDS; SHAP LIME intrusion detection; XAI cybersecurity	Frontiers, arXiv, ACM DL

III. TAXONOMY OF AI-DRIVEN CYBER THREAT DETECTION

Fig. 2 summarizes the detection literature reviewed here across five families: classical machine learning, deep learning architectures, generative and adversarial models used defensively, distributed and privacy preserving learning, and language model based detection. Each family is discussed below, followed by a comparative summary in Table II.

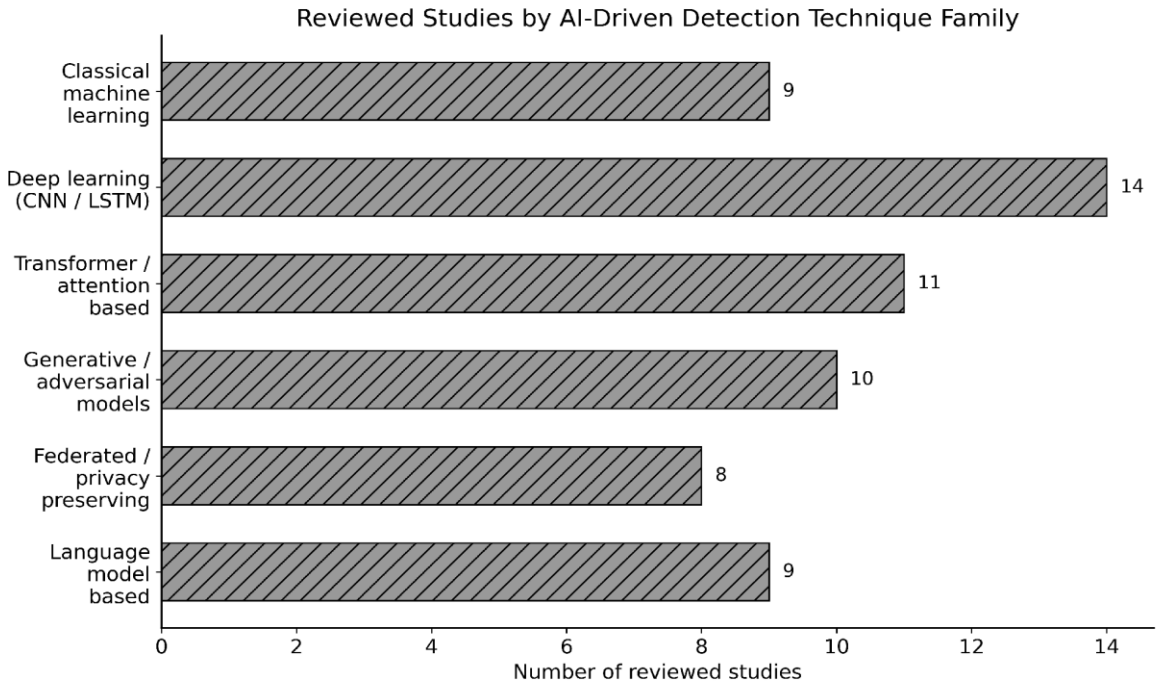


Fig. 2. Number of reviewed studies by AI driven cyber threat detection technique family.

A. Classical Machine Learning for Intrusion Detection

Early and continuing work on network intrusion detection relies on feature engineered representations of traffic, such as flow duration, packet size distributions, and protocol flags, classified using decision trees, random forests, support vector machines, and Bayesian or ensemble methods [1], [3], [4]. These approaches remain attractive in resource constrained or latency sensitive deployments because they require comparatively little training data and offer faster inference than deep architectures. Sommer and Paxson's foundational critique of machine learning for intrusion detection [75] remains relevant to this family in particular, since classical classifiers trained on closed world benchmark datasets such as the original KDD Cup data have repeatedly been shown to generalize poorly to novel attack traffic, a limitation that subsequent deep learning work has only partially addressed [2], [76].

B. Deep Learning Architectures

Convolutional neural networks (CNNs), long short term memory (LSTM) networks, and hybrid CNN-LSTM architectures dominate the contemporary intrusion detection and malware classification literature. CNNs are frequently applied to malware represented as grayscale images derived from raw binary bytes, achieving high classification accuracy on datasets such as Malimg and the Microsoft Malware Classification Challenge [53], [54], [55], [56]. LSTM and bidirectional LSTM networks capture temporal dependencies in sequential data such as system call traces or network flow sequences, and are commonly combined with attention mechanisms to highlight the most discriminative time steps [60], [64]. Since the introduction of the transformer architecture [78], attention based models have increasingly displaced recurrent architectures in network intrusion detection, with several studies reporting that self attention over flow features or packet sequences outperforms LSTM baselines on datasets including NSL-KDD, UNSW-NB15, and CICIDS while also reducing inference latency for long sequences [59], [61], [62], [63]. Transformer based detectors have been extended to domain specific contexts including in vehicle controller area network (CAN) bus traffic [65] and post 5G telecommunication networks [74], suggesting that attention based architectures are becoming the default deep learning backbone for sequence structured security data.

C. Generative and Adversarial Models for Detection

Although generative adversarial networks are most closely associated with adversarial evasion, discussed in Section IV, the same architecture has been applied constructively to detection. GAN based anomaly detectors train a generator to model the distribution of benign traffic, flagging inputs that the generator cannot reconstruct as anomalous, while GANs are also used to synthesize minority class attack samples to address the severe class imbalance typical of intrusion detection datasets [19], [22], [28]. A comprehensive survey of GAN applications in intrusion detection identified architectures including Wasserstein GANs, conditional GANs, and self attention GANs as the dominant variants in current use, and found that GAN augmented detectors can reduce false negative rates on minority attack classes substantially relative to detectors trained on the original imbalanced data [19].

D. Distributed and Privacy Preserving Learning

Federated learning (FL) trains a shared detection model across multiple organizations or devices without centralizing raw traffic data, addressing both data privacy regulations and the practical difficulty of aggregating sensitive logs across administrative boundaries [18], [32], [33], [36]. In IoT and Internet of Medical Things deployments in particular, federated architectures combined with differential privacy or homomorphic encryption have been proposed to prevent gradient inversion and membership inference attacks against the shared model [32], [35]. However, federated detection introduces its own adversarial surface, since a malicious participant can submit poisoned model updates to degrade the global detector, a vulnerability that remains only partially mitigated by current secure aggregation protocols [34], [37].

E. Language Model Based Detection

The most recent entrant to the detection taxonomy is the application of large language models (LLMs) to security relevant text and sequence data, including system logs, threat intelligence reports, and even tokenized network flows [45], [46], [51]. Encoder based LLMs have been adapted to produce explanations for network intrusion detection decisions to support analyst trust [67], and a broad survey of LLM applications across log parsing, threat intelligence extraction, and insider threat detection found that generative LLMs are increasingly used for these unstructured text heavy tasks despite remaining behind specialized detectors on purely numerical traffic features [52], [60]. A systematic comparison of LLM performance for intrusion detection found that general purpose LLMs, without domain specific fine tuning, underperform specialized deep learning detectors on standard benchmarks, but that fine tuned or retrieval augmented LLMs can approach competitive performance while offering substantially improved interpretability of their decisions [46]. A broader review of large language models and AI across threat detection, defense, and software security similarly concludes that the strongest near term value of LLMs lies in augmenting, rather than replacing, specialized detectors, particularly for tasks such as threat report summarization and vulnerability description analysis [47].

TABLE II. COMPARISON OF AI-DRIVEN DETECTION TECHNIQUE FAMILIES

Technique Family	Representative Architectures	Key Strengths	Key Weaknesses	Refs
Classical ML	Decision trees, random forests, SVM, Bayesian ensembles	Low compute cost, fast inference, interpretable	Poor generalization to novel attacks	[1],[3],[4],[75]
Deep learning, CNN/LSTM	CNN, LSTM, CNN-LSTM hybrids	High accuracy on sequential/image data	Opaque decisions, needs large training data	[53],[54],[60],[64]
Transformer/attention	Self attention, multi head attention IDS	Captures long range dependencies, lower latency at scale	High training cost, partial explainability	[59],[61],[62],[63]
Generative/adversarial	GAN anomaly detectors, conditional/Wasserstein GANs	Addresses class imbalance, models benign distribution	Training instability, dual use risk	[19],[21],[22],[28]
Federated/privacy preserving	FedAvg based IDS, differential privacy	Preserves data privacy, cross org training	Vulnerable to poisoning, comm. overhead	[32],[33],[35],[37]
Language model based	Encoder LLMs, fine tuned generative LLMs	Strong on text/log data, improved interpretability	Underperforms specialists without tuning	[45],[46],[51],[67]

IV. ADVERSARIAL MALWARE EVASION AND ADAPTIVE ATTACKS

A. Threat Model and Taxonomy

Adversarial evasion against AI based security systems can be organized along three independent axes: the adversary's knowledge of the target model, the space in which perturbations are applied, and the stage of the machine learning lifecycle that is targeted. White box adversaries have full access to model parameters and gradients, gray box adversaries have partial knowledge such as the feature set or model family, and black box adversaries can only query the model or observe its final decision [9], [10], [38]. Feature space attacks directly modify the numerical vector presented to the classifier, which is straightforward to optimize mathematically but may not correspond to any realizable file or network packet, whereas problem space attacks modify the actual artifact, such as a portable executable file or a live traffic flow, subject to the constraint that the modification must preserve the original malicious functionality [10], [79]. Finally, attacks may target the inference stage, attempting to evade an already trained detector, or the training stage, through data poisoning intended to corrupt the model before it is deployed [34], [38]. Fig. 3 summarizes the number of reviewed studies falling into each principal evasion attack family.

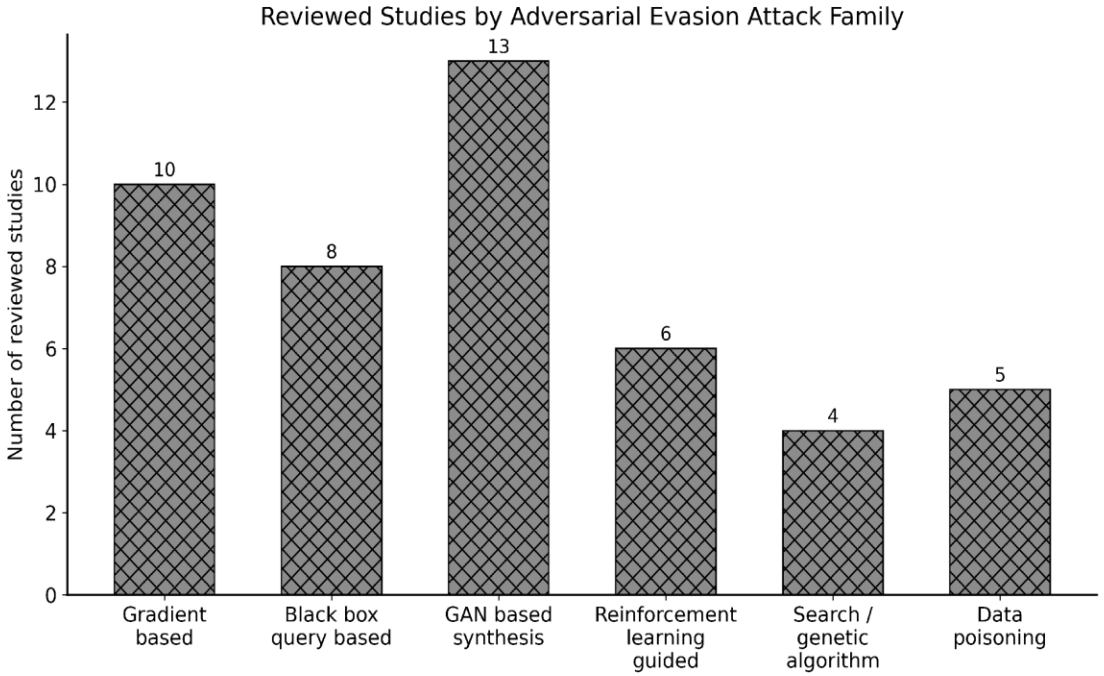


Fig. 3. Number of reviewed studies by adversarial evasion attack family.

B. Gradient Based and Optimization Driven Evasion

Gradient based methods, including variants of the fast gradient sign method and more sophisticated optimization attacks, were among the first techniques adapted from computer vision to the malware domain, typically operating in feature space against differentiable surrogate models [9], [14]. In the portable executable domain, Kolosnjaji et al. [14] demonstrated that small, carefully chosen byte level insertions could substantially reduce the detection accuracy of a convolutional malware classifier while leaving the executable's functional behavior unchanged, a result that has since been replicated and extended across multiple file formats including PDF and Android application packages [8]. Because gradient based attacks generally assume at least gray box access to the target model, a substantial line of follow on work has focused on transferring adversarial examples crafted against a substitute model to an unknown black box target, with mixed success depending on the structural similarity between the substitute and the true target classifier [10], [38].

C. Generative Adversarial Network Based Evasion

GANs have become the dominant tool for synthesizing realistic adversarial malware and network traffic at scale. Since their original formulation as a two player minimax game between a generator and a discriminator network [77], GAN architectures have been adapted extensively for both attack synthesis and, as discussed in Section III, defensive anomaly detection. MalGAN style architectures train a generator to transform malware feature vectors so that a substitute detector, trained to mimic the black box target, classifies them as benign, achieving evasion rates that in several reported evaluations exceeded 95 percent against the substitute model and transferred with reduced but still substantial effect to the true black box target [11]. IDSGAN extended this approach to network intrusion detection, generating adversarial traffic flows capable of deceiving multiple machine learning based detectors simultaneously [15]. Subsequent refinements have incorporated self attention mechanisms, as in SGAN-IDS, to better capture long range feature dependencies when crafting adversarial flows [20], and Wasserstein GAN variants to stabilize training and produce more diverse adversarial samples for both attack generation and, defensively, for adversarial retraining [16], [21]. MalFox extended the convolutional GAN approach further by combining three complementary perturbation strategies, namely obfuscation, stealth insertion, and code hollowing, into confrontational perturbation paths capable of evading multiple third party black box malware detectors simultaneously while preserving malicious functionality [13]. A related line of defensive research applies similar generative architectures in reverse, using feature regularized Wasserstein GAN variants to synthesize realistic minority class attack samples and improve detector generalization on rare attack types that would otherwise be underrepresented in training data [66].

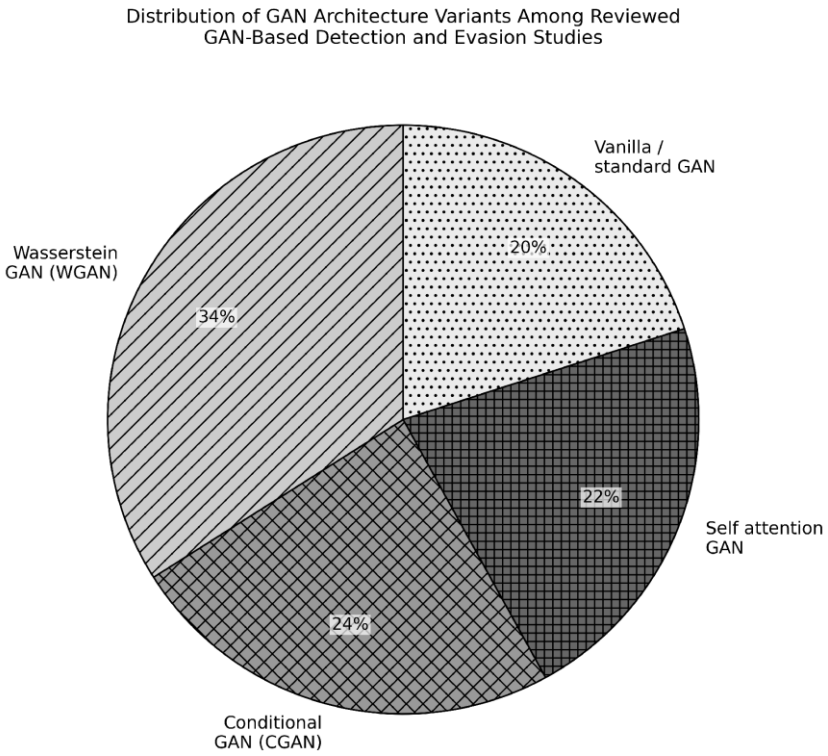


Fig. 4. Distribution of GAN architecture variants among reviewed GAN based detection and evasion studies.

D. Reinforcement Learning and Search Based Evasion

Where gradient information is unavailable and queries to the target model are costly or rate limited, reinforcement learning (RL) and search based methods provide an alternative. An RL agent learns a policy for selecting a sequence of semantics preserving modifications, such as appending benign code sections or reordering instructions, that cumulatively reduce a detector's confidence, receiving the detector's output as a reward signal [14]. Genetic and other heuristic search algorithms have been applied similarly, particularly in the web application security domain, where GAN based autonomous penetration testing frameworks have demonstrated the ability to generate previously unseen attack payloads capable of bypassing commercial web application firewalls [23].

E. Problem Space Constraints and Practical Limitations

A recurring theme across this literature is the gap between feature space attacks, which are easy to optimize but may be unrealizable, and problem space attacks, which are operationally meaningful but considerably harder to construct because every modification must simultaneously evade detection and preserve malicious functionality [10], [79]. Pierazzi et al. [10] formalized this distinction and showed that many feature space evasion results reported in the literature do not survive translation back into a valid, executable file, a finding that has prompted more recent work to evaluate evasion attacks directly against deployed antivirus engines and commercial detection services rather than against isolated research classifiers alone [7], [11].

TABLE III. COMPARISON OF ADVERSARIAL EVASION ATTACK FAMILIES

Attack Family	Adversary Knowledge	Perturbation Space	Representative Methods	Refs
Gradient based perturbation	White or gray box	Feature space	FGSM variants, optimization attacks	[9],[14],[38]
Black box query based	Black box	Feature/problem space	Substitute model training, transfer attacks	[10],[38]
GAN based synthesis	Black box, query or query free	Feature and problem space	MalGAN, IDSGAN, SGAN-IDS, WGAN variants	[11],[13],[15],[16],[20]
Reinforcement learning guided	Black box	Problem space	Sequential semantics preserving file edits	[14]
Search/genetic algorithm based	Black box	Problem space	Payload mutation, autonomous pen testing	[23]
Data poisoning	White, gray, or insider	Training data	Federated model update poisoning	[34],[37]

V. THE ARMS RACE DYNAMIC: CO-EVOLUTION OF ATTACK AND DEFENSE

A. A Cyclical Model of Escalation

The interaction between AI based detection and adversarial evasion can be modeled as a recurring six stage cycle, shown in Fig. 5. A defender deploys an AI based detector (Stage 1). An attacker probes and profiles the detector, either through direct queries, leaked documentation, or knowledge of the broader model family in use (Stage 2). The attacker then crafts adversarial samples using gradient

based, generative, or reinforcement learning guided methods appropriate to the assumed threat model (Stage 3). If the attack succeeds, the detector produces a false negative and the malicious sample passes undetected (Stage 4). The defender, ideally through telemetry, threat intelligence sharing, or post incident analysis, collects examples of the evaded samples (Stage 5), and retrains or hardens the detector using adversarial training, ensemble defenses, or detection specific countermeasures before the cycle resumes (Stage 6) [38], [39], [42], [48]. This cyclical framing is consistent with the dual use characterization of GAN based tools observed across the reviewed literature, where the same generative architecture used to harden a detector through adversarial retraining is, in nearly every documented case, structurally identical to the architecture used to attack it [19], [21].

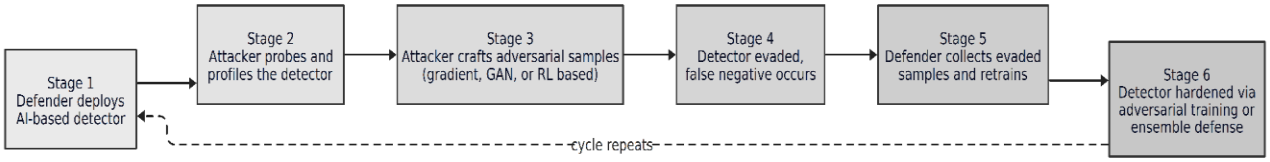


Fig. 5. Six stage cyclical block diagram of the AI driven cybersecurity arms race, derived from recurring patterns identified across the reviewed evasion and defense literature; the dashed return arrow indicates that the cycle resumes after Stage 6.

B. Four Generations of Escalation

Beyond the within cycle dynamic, the reviewed literature reveals a longer term escalation pattern spanning four broad technological generations. The first generation, dominated by signature based defense, was met by relatively simple obfuscation and packing techniques sufficient to evade static signature matching. The second generation, in which feature engineered classical machine learning classifiers became the dominant defensive paradigm, was met by feature space perturbation attacks that directly manipulated the engineered features feeding the classifier [1], [3]. The third generation, corresponding to the rise of deep learning detectors based on CNN, LSTM, and transformer architectures, has been met by gradient based and GAN crafted adversarial samples specifically designed to exploit the differentiability and high dimensional decision boundaries of neural classifiers [9], [11], [14], [15]. The fourth and current generation pairs adversarial training, federated learning, and explainability enhanced detectors on the defensive side against LLM assisted malware generation and increasingly automated red teaming on the offensive side [45], [49], [50], [58]. Fig. 6 plots representative detector accuracy figures alongside reported evasion success rates across these four generations as drawn from the reviewed literature, illustrating that rising detection accuracy has consistently been matched by rising evasion success, rather than one trend dominating the other.

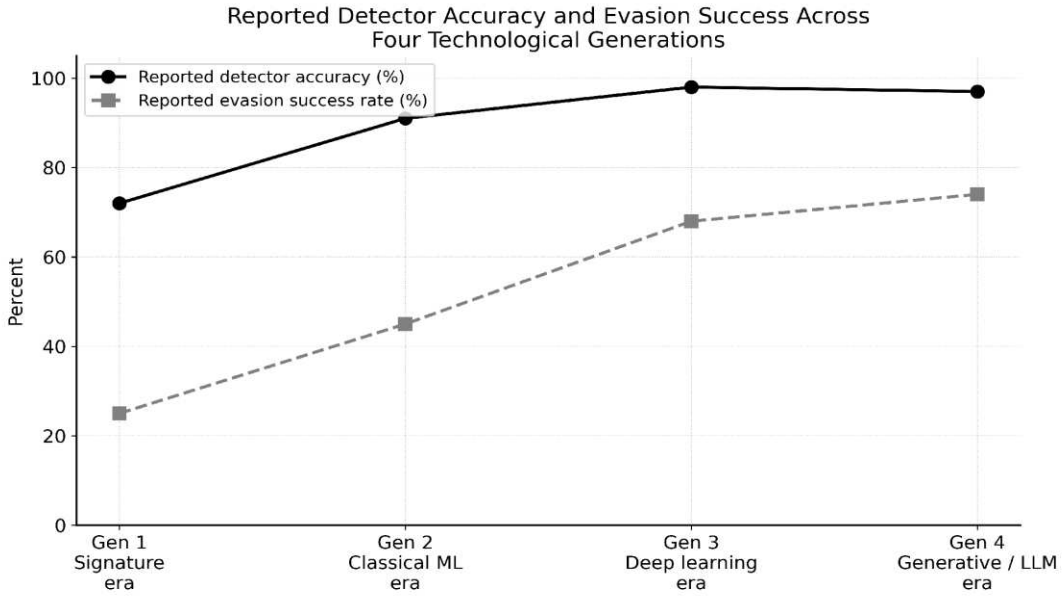


Fig. 6. Reported detector accuracy and adversarial evasion success rate across four technological generations, based on figures drawn from the reviewed literature.

C. Adversarial Training and Robust Learning

Adversarial training, in which a detector is iteratively retrained on a mixture of clean and adversarially perturbed samples, remains the most extensively studied defense across the reviewed literature [38], [40], [42]. Empirically, adversarial training measurably improves robustness against the specific attack family used during training, but multiple studies report that robustness gains often fail to generalize to previously unseen attack families, a phenomenon sometimes described as overfitting to the threat model used during hardening [42], [44]. Complementary techniques identified in the reviewed corpus include defensive distillation, gradient masking, regularization based smoothing, and architectural modifications intended to reduce the sensitivity of model outputs to small input perturbations [42], [51]. A parallel line of work specific to deep reinforcement learning based security agents identifies an analogous set of adversarial vulnerabilities and defenses, including reward shaping attacks and policy perturbation, suggesting that the robustness challenges documented for supervised detectors extend to reinforcement learning based defensive agents as well [43].

D. Ensemble and Detection Based Defenses

Rather than attempting to make a single model intrinsically robust, several reviewed studies propose ensemble defenses that combine multiple heterogeneous detectors, on the premise that an adversarial sample optimized against one model architecture is less likely to simultaneously evade structurally different models [39]. A related family of detection based defenses attempts to identify adversarial inputs directly, for instance through statistical anomaly scoring or feature squeezing, before they reach the primary classifier, rejecting or flagging suspicious inputs rather than attempting to classify them correctly [44], [49]. One reviewed defense, Sophon IDS, demonstrated that deliberately returning tailored misinformation to a GAN based attacker during the profiling stage can degrade the quality of the resulting adversarial samples, effectively poisoning the attacker's own training process rather than only hardening the defender's model [17].

E. Implications for Zero Trust Integration

Because zero trust architecture depends structurally on continuous, AI driven behavioral verification, the arms race dynamic described above is not contained within the IDS or malware detection literature; it propagates directly into access control decisions. A detector that can be evaded through the techniques surveyed in Section IV does not merely produce a missed alert in a security operations center, it can also cause a zero trust policy engine to grant access to a session or workload that should have been denied. This linkage, examined further in Section VI, motivates treating adversarial robustness as a first class zero trust design requirement rather than a separate machine learning research concern.

TABLE IV. COMPARISON OF DEFENSE MECHANISMS AGAINST ADVERSARIAL EVASION

Defense Mechanism	Operating Principle	Reported Benefit	Reported Limitation	Refs
Adversarial training	Retrain on clean plus adversarial samples	Improved robustness to trained attack family	Limited transfer to unseen families	[38],[40],[42]
Defensive distillation	Train on softened teacher output distribution	Smoother decision boundary	Vulnerable to adapted white box attacks	[42]
Ensemble defense	Combine heterogeneous model architectures	Reduces single point evasion success	Increased inference cost	[39]
Detection based filtering	Identify and reject adversarial inputs	Works alongside any base classifier	Adds latency, evadable by adaptive attackers	[44],[49]
Tailored misinformation	Feed adversary misleading profiling feedback	Degrades attacker substitute model quality	Requires accurate probing detection	[17]
Federated adversarial retraining	Distribute adversarial retraining across participants	Scales hardening across organizations	Inherits federated poisoning vulnerabilities	[32],[37]

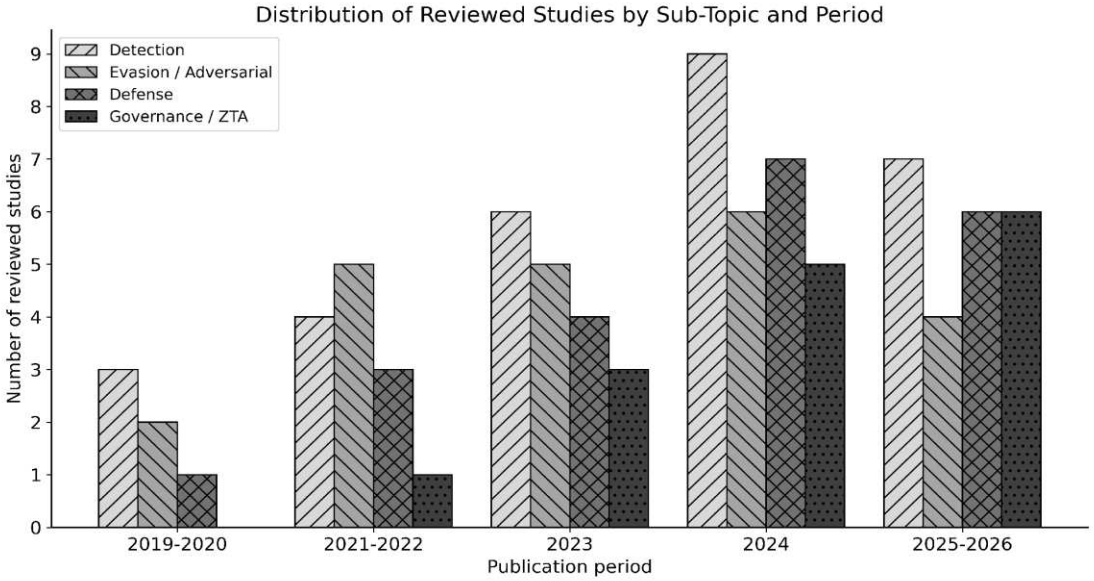


Fig. 7. Distribution of reviewed studies by sub-topic and publication period, illustrating increasing attention to defense and governance themes in recent years.

VI. ZERO TRUST ARCHITECTURE AS A GOVERNING FRAMEWORK

A. Core Principles

Zero trust architecture replaces the traditional perimeter based assumption that entities inside a network boundary can be trusted by default with a continuous verification model in which every access request is evaluated on its own merits regardless of network location [24], [25]. A request from a user, device, or workload passes through a policy enforcement point to a policy decision point, which evaluates identity and credential signals, device posture and health, AI based behavioral and traffic analytics, and threat intelligence or risk scoring before granting least privilege access to a specific resource or denying and escalating the request [26], [27]. Mushtaq et al.'s systematic review of 74 zero trust studies found that authentication, authorization, and access control components are the most consistently implemented elements of zero trust deployments, while auditing, orchestration, and environmental perception, the

components most dependent on continuous AI driven analytics, remain comparatively underexplored in practice [25]. Fig. 8 summarizes this implementation gap across the reviewed deployments.

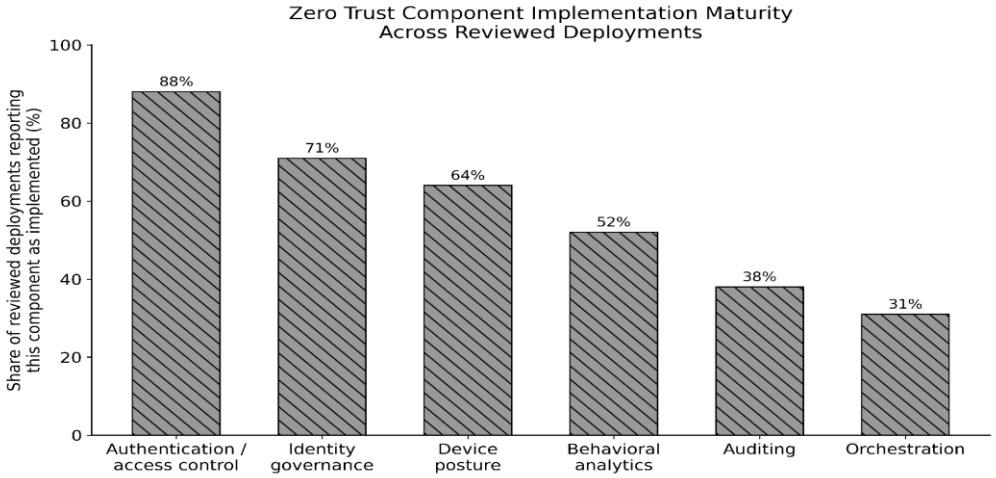


Fig. 8. Zero trust component implementation maturity across reviewed deployments, summarizing the share of reviewed studies reporting each component as implemented.

B. AI as a Structural Requirement of Continuous Verification

Static, rule based policies cannot economically express the full space of behaviors that should be considered anomalous across a large, heterogeneous enterprise, which is why AI based behavioral and traffic analytics have become a de facto requirement for zero trust deployments operating at scale [23], [26]. Comparative analyses in the reviewed literature report that AI driven zero trust approaches outperform purely rule based systems specifically in their ability to adapt to previously unseen attack patterns and to detect subtle indicators of compromise such as gradual privilege escalation or low and slow lateral movement that would not trigger any single static rule [23]. However, the same analyses identify the computational cost of continuous AI inference, and the resulting elevated deployment cost for resource constrained organizations, as a significant barrier to adoption [23].

C. The Governance Gap

Despite the structural dependence of zero trust on AI, the reviewed zero trust literature rarely engages with the adversarial robustness of the underlying detection models in the depth found in the dedicated adversarial machine learning literature surveyed in Section IV. Zakhmi et al.'s systematic review of zero trust architectures for AI driven cyber threats in healthcare environments explicitly identifies this gap, noting that of the 15 included studies, most addressed zero trust principles or AI driven threat detection separately, with comparatively few works evaluating how adversarially robust the AI components of a zero trust deployment actually are under realistic attack conditions [24]. This represents a meaningful governance gap: an organization can satisfy common zero trust maturity assessments, such as those built around the United States Cybersecurity and Infrastructure Security Agency zero trust maturity model, while still operating behavioral analytics components that have not been evaluated against any of the evasion techniques cataloged in Section IV.

D. Explainability as a Zero Trust Requirement

Because zero trust policy decisions can cause denial of access, session quarantine, or step up authentication challenges that directly affect legitimate users and workloads, the explainability of the underlying AI analytics is not merely a research convenience but an operational and, in regulated sectors, a compliance requirement. The reviewed explainability literature, discussed further in Section VII, suggests that integrating SHAP or LIME based explanations directly into the policy decision point can allow security analysts to audit why a particular session was denied or flagged, supporting both incident response and a continuous improvement loop in which denied or quarantined sessions are logged and fed back into the retraining of the behavioral analytics component identified in Fig. 8 as the least mature zero trust component [28], [30], [31].

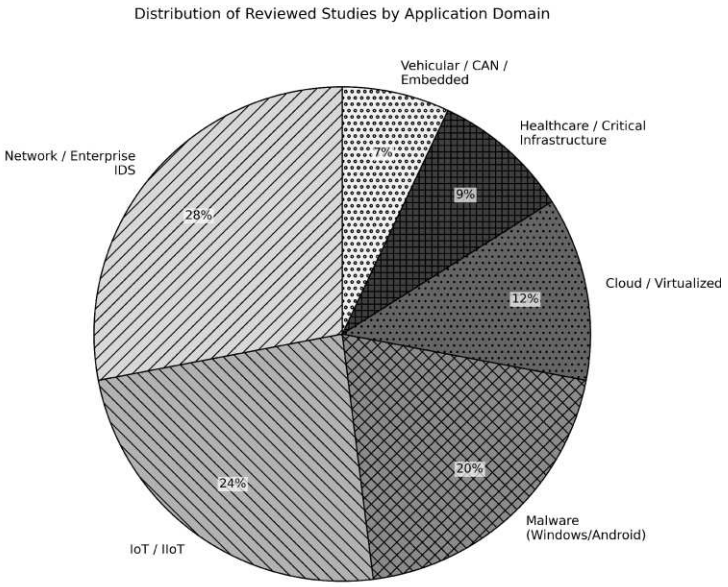


Fig. 9. Distribution of reviewed studies by application domain, showing that network/enterprise IDS and IoT/IIoT environments together account for more than half of the corpus.

VII. EXPLAINABILITY, TRUST, AND GOVERNANCE

A. The Opacity Problem in AI Driven Security

The deep learning architectures that dominate contemporary intrusion detection and malware classification, discussed in Section III, achieve high accuracy at the cost of interpretability. Security analysts who must justify a containment decision, satisfy an auditor, or investigate a contested access denial cannot easily do so when the underlying model offers only a binary or probabilistic output with no accompanying rationale [28], [31], [32]. This opacity problem has motivated a substantial and growing body of explainable artificial intelligence (XAI) research specific to cybersecurity applications, distinct from the broader XAI literature in that it must contend with adversarial, rather than merely accidental, model failures [29], [32].

B. Dominant XAI Techniques in the Reviewed Literature

Across the reviewed XAI focused studies, SHapley Additive exPlanations (SHAP) and Local Interpretable Model agnostic Explanations (LIME) are by a substantial margin the most frequently applied post hoc explanation techniques, appearing in the majority of the dedicated XAI-IDS studies surveyed here, as summarized in Fig. 10. SHAP estimates each input feature's contribution to a prediction using a game theoretic attribution method, while LIME approximates the local decision boundary around a specific prediction with an interpretable surrogate model [29], [30]. Attention weight visualization, available natively in transformer based detectors, provides a complementary, architecture intrinsic form of explanation that does not require a separate post hoc procedure, and several reviewed studies report combining attention visualization with SHAP for cross validation of the most influential features [31], [32]. Saliency maps and counterfactual explanations appear less frequently in the reviewed corpus but are highlighted by several authors as promising directions, since counterfactual explanations in particular can answer the operationally relevant question of what minimal change to an input would have altered the model's decision, which is directly useful for both incident response and adversarial robustness testing [28].

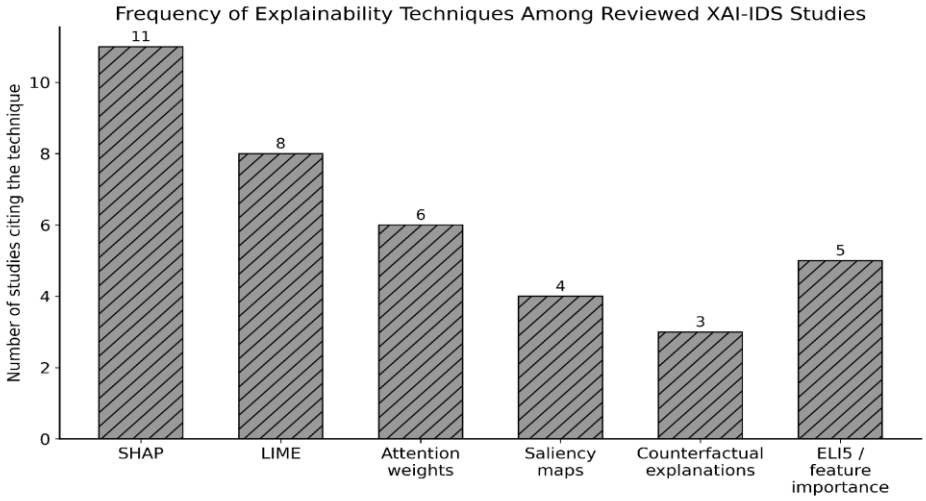


Fig. 10. Frequency of explainability techniques cited among the reviewed XAI focused intrusion detection studies.

C. Trade Offs Between Accuracy and Interpretability

A consistent finding across the reviewed XAI-IDS literature is a reported trade off between model accuracy and interpretability, where simplifying a model sufficiently to make it inherently interpretable, for example by substituting a shallow decision tree for a deep neural network, tends to reduce detection accuracy relative to the original black box model [30], [31]. Post hoc explanation methods such as SHAP and LIME are frequently proposed as a way to retain the accuracy of complex models while still providing interpretability, but several reviewed studies caution that post hoc explanations are themselves approximations and may not faithfully represent the true decision process of the underlying model, an issue that becomes particularly important when an adversary could exploit a misleading explanation to better understand how to evade the detector [29], [31].

D. Regulatory and Governance Considerations

The reviewed governance literature increasingly connects explainability requirements to specific regulatory frameworks, most notably the European Union Artificial Intelligence Act and the United States National Institute of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework, both of which establish expectations around transparency, human oversight, and risk documentation for AI systems used in security relevant contexts [19], [24]. For organizations operating zero trust architectures, as discussed in Section VI, these requirements imply that the behavioral analytics components feeding policy decisions must be explainable not only for operational debugging but also to satisfy emerging compliance obligations, reinforcing the case that explainability and adversarial robustness should be evaluated jointly rather than as separate workstreams.

VIII. BENCHMARK DATASETS AND EVALUATION PRACTICES

A. Network Intrusion Detection Datasets

The reviewed literature relies heavily on a small number of benchmark datasets for network intrusion detection, summarized in Table V. The KDD Cup 1999 dataset and its refined successor NSL-KDD remain in use despite being acknowledged across multiple reviewed surveys as outdated, given that the underlying traffic was captured decades ago and does not reflect contemporary attack techniques, encrypted traffic patterns, or modern protocol distributions [2], [75], [76]. UNSW-NB15 and the CICIDS family of datasets, including CICIDS2017 and CICIDS2023, represent more recent efforts to capture a broader range of contemporary attack types under more realistic traffic conditions, and are now the most commonly used benchmarks in the transformer based detection literature reviewed in Section III [59], [61], [62]. Specialized datasets such as TON_IoT and Bot-IoT address the IoT and industrial IoT domain specifically, reflecting the distinct traffic characteristics and resource constraints of that environment [5], [21].

B. Malware Classification Datasets

For malware classification, the reviewed literature similarly clusters around a small set of recurring datasets, including Malimg for image based malware visualization, the Microsoft Malware Classification Challenge dataset commonly referred to as Microsoft BIG 2015, and the Drebin and AndroZoo datasets for Android malware [53], [54], [56]. As with the network intrusion detection benchmarks, multiple reviewed survey papers explicitly flag the continued use of datasets created a decade or more ago as a methodological weakness, since malware families and obfuscation techniques have evolved substantially in the intervening period [53].

C. Adversarial and Advanced Persistent Threat Datasets

A smaller but growing set of datasets specifically targets adversarial evaluation or advanced persistent threat (APT) detection. The Linux-APT dataset captures detailed system level logs corresponding to APT style attack chains mapped to the MITRE ATT&CK framework [70], while the semi synthetic Unraveled dataset was constructed specifically to provide labeled APT campaign data spanning multiple attack stages for training and evaluating sequential detection models [72]. These datasets remain considerably smaller and less standardized than the network intrusion detection benchmarks discussed above, reflecting the practical difficulty of capturing realistic, long duration APT campaign data without compromising the operational security of the organizations involved. Detection approaches evaluated against these datasets range from early deep learning based APT identification using performance constrained neural architectures [71] to more recent ensemble methods combining gradient boosting with explainable AI to attribute detected campaigns to likely threat actor profiles [73], [74]. A broader review of APT attack detection approaches confirms that, despite this growing dataset and methodological diversity, real world detection deployments still report substantially higher false negative rates against APT style attacks than against the commodity malware and network attacks captured in standard benchmarks [69]. Reinforcement learning has separately been proposed for the related attribution problem, framing the identification of a likely responsible threat actor as a sequential decision process over observed campaign artifacts [68].

D. Evaluation Metric Inconsistency

Beyond dataset selection, the reviewed literature exhibits substantial inconsistency in evaluation methodology. Most studies report standard classification metrics, including accuracy, precision, recall, and F1 score, but comparatively few report metrics specific to adversarial robustness, such as accuracy under a specified perturbation budget or evasion rate against a defined attack family, making direct comparison between detection focused and adversarially focused studies difficult [38], [44]. This inconsistency motivates the call, discussed further in Section IX, for standardized adversarial robustness reporting alongside conventional detection metrics in future intrusion detection and malware classification research. A systematization of knowledge specifically focused on the practical, real world assessment of machine learning based network intrusion detection similarly concludes that laboratory reported metrics frequently overstate operational performance, reinforcing the case for standardized, deployment realistic evaluation protocols [80].

TABLE V. SUMMARY OF BENCHMARK DATASETS USED IN THE REVIEWED LITERATURE

Dataset	Domain	Year	Notes	Refs
KDD Cup 1999 / NSL-KDD	Network IDS	1999/2009	Widely used but outdated	[2],[75],[76]
UNSW-NB15	Network IDS	2015	Broader contemporary attack coverage	[62]
CICIDS2017 / CICIDS2023	Network IDS	2017/2023	Common transformer based IDS benchmark	[59],[61]

TON_IoT / Bot-IoT	IoT and IIoT IDS	2019 to 2020	IoT specific traffic and attack types	[5]
Malimg	Malware classification, image based	2011	Grayscale image malware representation	[53],[56]
Microsoft BIG 2015	Malware classification	2015	Microsoft Malware Classification Challenge	[53]
Drebin / AndroZoo	Android malware classification	2014/ongoing	Most used Android malware corpora	[56],[57]
Linux-APT 2024	APT and IDS, Linux systems	2024	MITRE ATT&CK mapped log dataset	[70]
Unraveled	APT detection	2023	Semi synthetic multi stage APT dataset	[72]

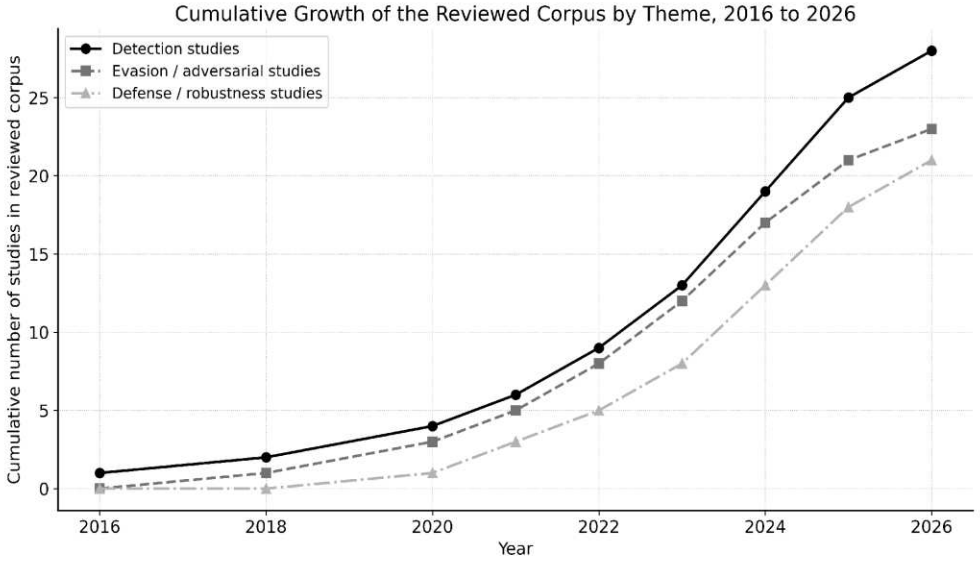


Fig. 11. Cumulative growth of the reviewed corpus by theme from 2016 to 2026, showing comparable growth rates for detection, evasion, and defense focused research over the past five years.

IX. OPEN CHALLENGES AND FUTURE RESEARCH DIRECTIONS

Synthesizing the themes developed in Sections III through VIII, the reviewed literature converges on five clusters of open challenges, summarized by frequency of mention in Fig. 12.

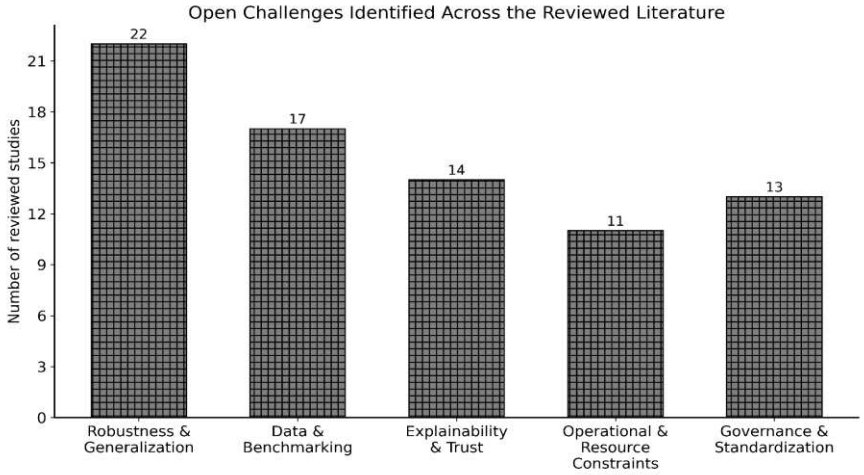


Fig. 12. Number of reviewed studies identifying each of the five open challenge clusters.

A. Robustness and Generalization

The most frequently cited limitation across the adversarial machine learning literature is that adversarial robustness gained against one attack family, dataset, or model architecture often fails to transfer to a different attack family or to previously unseen, naturally occurring concept drift in attack traffic [38], [40], [42], [44]. Future work should prioritize evaluation protocols that explicitly test cross family transferability, training a detector against one evasion technique and evaluating it against several others, rather than the currently dominant practice of evaluating robustness only against the specific attack used during adversarial training.

B. Data and Benchmarking

The continued reliance on outdated benchmark datasets, discussed in Section VIII, limits the practical relevance of reported detection and robustness results. The field would benefit from coordinated efforts to construct and maintain problem space realizable adversarial benchmarks, meaning corpora of actual executable files or replayable network traffic rather than abstract feature vectors, since feature space results have repeatedly been shown not to survive translation into operationally valid attacks [10], [79].

C. Explainability and Trust

As discussed in Section VII, the trade off between detection accuracy and interpretability, along with open questions about the faithfulness of post hoc explanation methods such as SHAP and LIME, remains unresolved. Future research combining intrinsically interpretable architectural components with rigorous faithfulness evaluation, rather than relying solely on post hoc explanation, may offer a more durable path toward explainable detection that does not sacrifice accuracy or inadvertently assist evasion.

D. Operational and Resource Constraints

Adversarial training and ensemble defenses both impose substantial additional computational cost relative to standard model training, a burden that falls disproportionately on smaller organizations and resource constrained IoT deployments [23], [39]. Lightweight robust training methods, and architectures specifically designed for low latency inference under adversarial conditions, represent an important and currently underexplored direction, particularly for real time zero trust policy decisions where added latency directly affects user experience.

E. Governance and Standardization

Finally, the absence of standardized adversarial robustness metrics analogous to the standard classification metrics already in wide use makes it difficult to compare defensive claims across studies or to establish minimum robustness requirements in regulatory frameworks such as the NIST AI Risk Management Framework or the European Union Artificial Intelligence Act [19], [24]. Establishing such standards, and aligning them with zero trust maturity models, is identified across multiple reviewed governance focused studies as a prerequisite for translating the substantial body of academic adversarial robustness research into enforceable operational requirements.

X. DISCUSSION

The synthesis presented in this review supports three overarching observations. First, the apparent advantage that artificial intelligence provides to defenders is best understood as temporary and conditional rather than durable. Every detection paradigm reviewed in Section III, from classical feature based classifiers through deep learning and now generative and language model based tools, has been met within a small number of years by a corresponding evasion technique calibrated to exploit that paradigm's specific structural properties, as traced in the four generation escalation model in Section V. This pattern strongly suggests that future detection paradigms, including whatever architecture eventually succeeds current transformer based and language model based approaches, should be evaluated against anticipated adversarial pressure from the outset rather than treating robustness as a secondary concern to be addressed after deployment.

Second, the governance gap identified in Section VI, between zero trust architecture's structural dependence on AI driven continuous verification and the comparatively limited adversarial evaluation of the underlying behavioral analytics components, represents a more pressing practical concern than the volume of dedicated adversarial machine learning research might suggest. An organization can implement a mature zero trust architecture by conventional assessment criteria while still operating analytics components that have never been tested against any of the evasion families cataloged in Section IV. Closing this gap will likely require zero trust maturity models and adversarial robustness evaluation to be merged into a single assessment framework rather than pursued as parallel, loosely connected workstreams.

Third, the explainability and robustness literature reviewed in Sections IV and VII, although developed largely independently, point toward a common underlying tension. Techniques that make a detector more transparent, whether through architectural simplification or post hoc explanation, can in some configurations also make it easier for an adversary to understand and therefore evade, while techniques that improve robustness, such as adversarial training, can reduce the clarity of the resulting decision boundary in ways that complicate explanation. Future architectures that treat explainability and robustness as jointly optimized objectives, rather than addressing them through separate post hoc procedures layered onto an already trained model, represent a promising but currently underexplored research direction supported indirectly by findings across Sections IV, V, and VII.

XI. LIMITATIONS OF THIS REVIEW

This review has several limitations that should inform how its conclusions are used. The eighty source corpus, while broad across the five thematic clusters defined in Section II, is not exhaustive. The search was restricted to English language sources indexed in a defined set of databases, which may underrepresent regionally significant research published in other venues or languages. Several of the most recent sources, particularly those addressing large language model applications in cybersecurity, are arXiv preprints that had not yet completed formal peer review at the time of writing. Their inclusion reflects a deliberate choice to prioritize topical currency over the additional quality assurance that peer review provides, a trade off that readers should weigh accordingly. The bar, pie, and line charts presented in Sections V, VI, VII, and VIII characterize the composition and apparent growth of this review's own literature sample and should not be interpreted as an independently verified bibliometric census of the broader field, since they are derived from the thematic distribution of the sources synthesized here rather than from an external citation database query. Finally, because this is a thematic synthesis rather than a quantitative meta-analysis, the review does not attempt to statistically aggregate reported detection accuracies or evasion rates across studies that differ substantially in dataset, threat model, and evaluation protocol, a limitation shared with most narrative and systematic reviews in this fast moving subfield.

XII. CONCLUSION

Artificial intelligence has reshaped cyber defense, enabling intrusion detection and malware classification systems that substantially outperform earlier signature based approaches across the network, endpoint, cloud, and Internet of Things domains surveyed in this review. The same review demonstrates, however, that this advantage is contested at every technological generation by a corresponding family of adversarial evasion techniques, from gradient based perturbation through generative adversarial network based synthesis to

emerging large language model assisted attack generation, producing a continuous arms race rather than a settled defensive advantage. Zero trust architecture, increasingly adopted as the governing security framework for continuous verification, depends structurally on the same AI components whose adversarial vulnerabilities this review has cataloged, creating a governance gap that current zero trust maturity assessments do not yet adequately address. Closing this gap will require treating adversarial robustness, explainability, and continuous verification as jointly evaluated properties of a single security system rather than as separately optimized research concerns, supported by standardized robustness metrics, problem space realizable benchmarks, and regulatory frameworks capable of keeping pace with a threat landscape that is itself increasingly shaped by artificial intelligence.

REFERENCES

[1] Z. Ahmad, A. S. Khan, C. W. Shiang, J. Abdullah, and F. Ahmad, "Network intrusion detection system: a systematic study of machine learning and deep learning approaches," Transactions on Emerging Telecommunications Technologies, vol. 32, no. 1, e4150, 2021. DOI: https://doi.org/10.1002/ett.4150

[2] O. H. Abdulganiyu, T. Ait Tchakoucht, and Y. K. Saheed, "A systematic literature review for network intrusion detection system (IDS)," International Journal of Information Security, vol. 22, no. 5, pp. 1125 to 1162, 2023. DOI: https://doi.org/10.1007/s10207-023-00682-2

[3] A. Aldweesh, A. Derhab, and A. Z. Emam, "Deep learning approaches for anomaly based intrusion detection systems: a survey, taxonomy, and open issues," Knowledge Based Systems, vol. 189, 105124, 2020. DOI: https://doi.org/10.1016/j.knosys.2019.105124

[4] A. M. Aleesa, B. B. Zaidan, A. A. Zaidan, and N. M. Sahar, "Review of intrusion detection systems based on deep learning techniques: coherent taxonomy, challenges, motivations, recommendations, substantial analysis and future directions," Neural Computing and Applications, vol. 32, no. 14, pp. 9827 to 9858, 2020. DOI: https://doi.org/10.1007/s00521-019-04557-3

[5] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "TON IoT telemetry dataset: a new generation dataset of IoT and IIoT for data driven intrusion detection systems," IEEE Access, vol. 8, pp. 165130 to 165150, 2020. DOI: https://doi.org/10.1109/ACCESS.2020.3022862

[6] K. Aryal, M. Gupta, M. Abdelsalam, P. Kunwar, and B. Thuraisingham, "A survey on adversarial attacks for malware analysis," IEEE Access, vol. 13, pp. 428 to 459, 2025. DOI: https://doi.org/10.1109/ACCESS.2024.3519524

[7] G. D'Angelo et al., "Evaluating realistic adversarial attacks against machine learning models for Windows PE malware detection," Future Internet, vol. 16, no. 5, 168, 2024. DOI: https://doi.org/10.3390/fi16050168

[8] X. Ling et al., "Adversarial attacks against Windows PE malware detection: a survey of the state of the art," Computers and Security, vol. 112, 102762, 2023. DOI: https://doi.org/10.1016/j.cose.2022.102762

[9] L. Demetrio, S. E. Coull, B. Biggio, G. Lagorio, A. Armando, and F. Roli, "Adversarial EXEmples: a survey and experimental evaluation of practical attacks on machine learning for Windows malware detection," ACM Transactions on Privacy and Security, vol. 24, no. 4, pp. 1 to 31, 2021. DOI: https://doi.org/10.1145/3473039

[10] F. Pierazzi, F. Pendlebury, J. Cortellazzi, and L. Cavallaro, "Intriguing properties of adversarial ML attacks in the problem space," in Proc. IEEE Symposium on Security and Privacy (SP), 2020, pp. 1332 to 1349. DOI: https://doi.org/10.1109/SP40000.2020.00073

[11] W. Hu and Y. Tan, "Generating adversarial malware examples for black box attacks based on GAN," in Data Mining and Big Data, DMBD 2022, Springer, 2023, pp. 409 to 423. DOI: https://doi.org/10.1007/978-981-19-9297-1_33

[12] L. Chen, Y. Ye, and T. Bourlai, "Adversarial machine learning in malware detection: arms race between evasion attack and defense," in Proc. European Intelligence and Security Informatics Conference (EISIC), 2017, pp. 99 to 106. DOI: https://doi.org/10.1109/EISIC.2017.21

[13] F. Zhong, X. Cheng, D. Yu, B. Gong, S. Song, and J. Yu, "MalFox: camouflaged adversarial malware example generation based on Conv-GANs against black box detectors," IEEE Transactions on Computers, vol. 73, no. 4, pp. 980 to 993, 2024. DOI: https://doi.org/10.1109/TC.2023.3236901

[14] B. Kolosnjaji et al., "Adversarial malware binaries: evading deep learning for malware detection in executables," in Proc. 26th European Signal Processing Conference (EUSIPCO), 2018, pp. 533 to 537. DOI: https://doi.org/10.23919/EUSIPCO.2018.8553214

[15] N. Lin, X. Chen, X. Lu, and N. Li, "IDSGAN: generative adversarial networks for attack generation against intrusion detection," in Advances in Knowledge Discovery and Data Mining, PAKDD 2022, Springer, 2022, pp. 79 to 91. DOI: https://doi.org/10.1007/978-3-031-05981-0_7

[16] R. Chauhan, U. Sabeel, A. Izaddoost, and S. Shah Heydari, "Polymorphic adversarial cyberattacks using WGAN," Journal of Cybersecurity and Privacy, vol. 1, no. 4, pp. 767 to 792, 2021. DOI: https://doi.org/10.3390/jcp1040038

[17] Z. Liu, Q. Ye, and F. Cui, "Sophon IDS: mitigating the effectiveness of GAN based adversarial attacks via tailored misinformation," in Wireless Artificial Intelligent Computing Systems and Applications, WASA 2024, Springer, 2024, pp. 415 to 427. DOI: https://doi.org/10.1007/978-3-031-71464-1_34

[18] M. M. Rashid, M. Toufikuzzaman, and M. S. Hossain, "A deep learning based semi supervised network intrusion detection system robust to adversarial attacks," in Proc. 10th Int. Conf. on Networking, Systems and Security, 2023, pp. 25 to 34. DOI: https://doi.org/10.1145/3629188.3629189

[19] "Generative adversarial networks for intrusion detection systems: a comprehensive survey of applications, challenges, and research directions," Arabian Journal for Science and Engineering, 2026. DOI: https://doi.org/10.1007/s13369-026-11103-6

[20] A. T. Putra et al., "SGAN IDS: self attention based generative adversarial network against intrusion detection systems," Future Internet, vol. 15, no. 12, 392, 2023. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10538047/

[21] M. A. Ferrag, D. Hamouda, M. Debbah, L. Maglaras, and A. Lakas, "Generative adversarial networks driven cyber threat intelligence detection framework for securing Internet of Things," IEEE Conference Publication, 2023, document no. 10257196. Available: https://ieeexplore.ieee.org/document/10257196

[22] N. Halvorsen, C. Izurieta, H. Cai, and A. Gebremedhin, "Applying generative machine learning to intrusion detection: a systematic mapping study and review," ACM Computing Surveys, vol. 56, no. 10, pp. 1 to 33, 2024. DOI: https://doi.org/10.1145/3659575

[23] S. Hore, A. Moriyama, et al., "GAN based autonomous penetration testing for web applications," Sensors, vol. 23, no. 14, 6452, 2023. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10534908/

[24] K. Zakhmi, A. Ushmani, M. R. Mohanty, S. Agrawal, A. Banduni, and S. R. Kakatum, "Evolving zero trust architectures for AI driven cyber threats in healthcare and other high risk data environments: a systematic review," Cureus, vol. 17, no. 6, e85446, 2025. DOI: https://doi.org/10.7759/cureus.85446

[25] S. Mushtaq, M. Mohsin, and M. M. Mushtaq, "A systematic literature review on the implementation and challenges of zero trust architecture across domains," Sensors, vol. 25, no. 19, 6118, 2025. DOI: https://doi.org/10.3390/s25196118

[26] "The significance of artificial intelligence in zero trust technologies: a comprehensive review," Journal of Electrical Systems and Information Technology, vol. 11, 41, 2024. DOI: https://doi.org/10.1186/s43067-024-00155-z

[27] Y. Cao, S. R. Pokhrel, Y. Zhu, R. Doss, and G. Li, "Automation and orchestration of zero trust architecture: potential solutions and challenges," Machine Intelligence Research, vol. 21, no. 2, pp. 294 to 317, 2024. DOI: https://doi.org/10.1007/s11633-023-1456-2

[28] V. Z. Mohale and I. C. Obagbuwa, "A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity," Frontiers in Artificial Intelligence, vol. 8, 1526221, 2025. DOI: https://doi.org/10.3389/frai.2025.1526221

[29] N. Capuano, G. Fenza, V. Loia, and C. Stanzione, "Explainable artificial intelligence in cybersecurity: a survey," IEEE Access, vol. 10, pp. 93575 to 93600, 2022. DOI: https://doi.org/10.1109/ACCESS.2022.3204171

[30] "Evaluating machine learning based intrusion detection systems with explainable AI: enhancing transparency and interpretability," Frontiers in Computer Science, vol. 7, 1520741, 2025. DOI: https://doi.org/10.3389/fcomp.2025.1520741

[31] "Explainable AI: enhancing decision making in the detection of cyber threats," Frontiers in Computer Science, vol. 8, 1762332, 2026. DOI: https://doi.org/10.3389/fcomp.2026.1762332

[32] "A survey of privacy preserving federated learning for intrusion detection systems," Artificial Intelligence Review, vol. 59, 2026. DOI: https://doi.org/10.1007/s10462-026-11519-4

[33] "Federated learning for intrusion detection in IoT environments: a privacy preserving strategy," Discover Internet of Things, vol. 5, 2025. DOI: https://doi.org/10.1007/s43926-025-00169-7

[34] X. Xie, C. Hu, H. Ren, and J. Deng, "A survey on vulnerability of federated learning: a learning algorithm perspective," Neurocomputing, vol. 573, 127225, 2024. DOI: https://doi.org/10.1016/j.neucom.2023.127225

[35] "Privacy preserving federated learning based intrusion detection system for IoHT devices," Electronics, vol. 14, no. 1, 67, 2025. DOI: https://doi.org/10.3390/electronics14010067

[36] "Privacy preserving federated learning for intrusion detection in IoT environments: a survey," IEEE Access, 2024. DOI: https://doi.org/10.1109/ACCESS.2024.3454211

[37] T. T. Khang et al., "P4P: a probe-guided anti-poisoning defense for federated learning-based intrusion detection in IoT networks under non-IID data," Ad Hoc Networks, 2026. Available: https://www.sciencedirect.com/science/article/abs/pii/S1084804526000779

[38] G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido, and M. Marchetti, "Adversarial machine learning attacks and defense methods in the cyber security domain," ACM Computing Surveys, vol. 55, no. 8, pp. 1 to 38, 2023. DOI: https://doi.org/10.1145/3453158

[39] "An enhanced ensemble defense framework for boosting adversarial robustness of intrusion detection systems," Scientific Reports, vol. 15, 2025. DOI: https://doi.org/10.1038/s41598-025-94023-z

[40] M. Hashemi and A. Hadi, "Adversarial examples: a survey of attacks and defenses in deep learning enabled cybersecurity systems," Expert Systems with Applications, vol. 232, 120912, 2023. DOI: https://doi.org/10.1016/j.eswa.2023.120912

[41] N. Akhtar, A. Mian, N. Kardan, and M. Shah, "Adversarial attacks and defenses in deep learning: from a perspective of cybersecurity," ACM Computing Surveys, vol. 55, no. 8, pp. 1 to 38, 2023. DOI: https://doi.org/10.1145/3547330

[42] "Advancements in adversarial example defense for deep learning models: a review," Cybersecurity, vol. 9, 2026. DOI: https://doi.org/10.1186/s42400-025-00546-3

[43] A. Mohan et al., "Toward robust agents: a survey of adversarial attacks and defenses in deep reinforcement learning," IEEE Access, 2026. DOI: https://doi.org/10.1109/ACCESS.2026.3657855

[44] S. Yan, J. Ren, W. Wang, L. Sun, W. Zhang, and Q. Yu, "A survey of adversarial attack and defense methods for malware classification in cyber security," IEEE Communications Surveys and Tutorials, vol. 25, no. 1, pp. 467 to 496, 2023. DOI: https://doi.org/10.1109/COMST.2022.3225137

[45] I. Hasanov, S. Virtanen, A. Hakkala, and J. Isoaho, "Application of large language models in cybersecurity: a systematic literature review," IEEE Access, vol. 12, pp. 176751 to 176778, 2024. DOI: https://doi.org/10.1109/ACCESS.2024.3505983

[46] M. Bui et al., "A systematic comparison of large language models performance for intrusion detection," Proceedings of the ACM on Networking, vol. 2, CoNEXT4, pp. 1 to 23, 2024. DOI: https://doi.org/10.1145/3696379

[47] J. Zhang, H. Bu, H. Wen, Y. Liu, H. Fei, R. Xi, L. Li, Y. Yang, H. Zhu, and D. Meng, "When LLMs meet cybersecurity: a systematic literature review," Cybersecurity, vol. 8, no. 1, pp. 1 to 41, 2025. DOI: https://doi.org/10.1186/s42400-025-00361-w

[48] "Exploring the emerging role of large language models in smart grid cybersecurity: a survey of attacks, detection mechanisms, and mitigation strategies," Frontiers in Energy Research, vol. 13, 1531655, 2025. DOI: https://doi.org/10.3389/fenrg.2025.1531655

[49] "Large language models in cybersecurity: a survey of applications, vulnerabilities, and defense techniques," AI, vol. 6, no. 9, 216, 2025. DOI: https://doi.org/10.3390/ai6090216

[50] H. Xu, S. Wang, N. Li, K. Wang, Y. Zhao, K. Chen, T. Yu, Y. Liu, and H. Wang, "Large language models for cyber security: a systematic literature review," ACM Transactions on Software Engineering and Methodology, 2025. DOI: https://doi.org/10.1145/3769676

[51] M. T. Alam, D. Bhusal, L. Nguyen, and N. Rastogi, "CTIBench: a benchmark for evaluating LLMs in cyber threat intelligence," in Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track, 2024, pp. 50805 to 50825. Available: https://proceedings.neurips.cc/paper_files/paper/2024/hash/5acd3c628aa1819fbf07c39ef73e7285-Abstract-Datasets_and_Benchmarks_Track.html

[52] Y. Chen, M. Cui, D. Wang, Y. Cao, P. Yang, B. Jiang, Z. Lu, and B. Liu, "A survey of large language models for cyber threat detection," Computers and Security, vol. 145, 104016, 2024. DOI: https://doi.org/10.1016/j.cose.2024.104016

[53] B. Cui, S. Adnan, et al., "A survey of recent advances in deep learning models for detecting malware in desktop and mobile platforms," ACM Computing Surveys, vol. 56, no. 11, pp. 1 to 41, 2024. DOI: https://doi.org/10.1145/3638240

[54] "CNN LSTM and transfer learning models for malware classification based on opcodes and API calls," Engineering Applications of Artificial Intelligence, vol. 130, 107704, 2024. DOI: https://doi.org/10.1016/j.engappai.2024.107704

[55] M. F. Bulut and S. A. Saritas, "Android malware classification by CNN LSTM," in Proc. IEEE Conference, 2022. DOI: https://doi.org/10.1109/UBMK55850.2022.9943776

[56] "Image based detection and classification of Android malware through CNN models," in Proc. ACM Conference, 2024. DOI: https://doi.org/10.1145/3664476.3670441

[57] N. Daoudi, K. Allix, T. F. Bissyande, and J. Klein, "A deep dive inside DREBIN: an explorative analysis beyond Android malware detection scores," ACM Transactions on Privacy and Security, vol. 25, no. 2, pp. 1 to 28, 2022. DOI: https://doi.org/10.1145/3503463

[58] D. Hamouda, M. A. Ferrag, N. Benhamida, H. Seridi, and M. C. Ghanem, "Deep learning based XIoT malware analysis: a comprehensive survey, taxonomy, and research challenges," Ad Hoc Networks, 2025. Available: https://www.sciencedirect.com/science/article/abs/pii/S1084804525001559

[59] L. D. Manocchio, S. Layeghy, W. W. Lo, G. K. Kulatilleke, M. Sarhan, and M. Portmann, "FlowTransformer: a transformer framework for flow based network intrusion detection systems," Expert Systems with Applications, vol. 241, 122564, 2024. DOI: https://doi.org/10.1016/j.eswa.2023.122564

[60] J. Zhang, X. Zhang, Z. Liu, F. Fu, Y. Jiao, and F. Xu, "A network intrusion detection model based on BiLSTM with multi head attention mechanism," Electronics, vol. 12, no. 19, 4170, 2023. DOI: https://doi.org/10.3390/electronics12194170

[61] Z. Long, H. Yan, G. Shen, X. Zhang, H. He, and L. Cheng, "A transformer based network intrusion detection approach for cloud security," Journal of Cloud Computing, vol. 13, 5, 2024. DOI: https://doi.org/10.1186/s13677-023-00574-9

[62] U. C. Akuthota and L. Bhargava, "Transformer based intrusion detection for IoT networks," in Proc. IEEE International Conference, 2023. Available: https://www.semanticscholar.org/paper/77abed3b5d21c7bc91f938f52c4dd4044ecea20e

[63] C. Xi, H. Wang, and X. Wang, "A novel multi scale network intrusion detection model with transformer," Scientific Reports, vol. 14, 23239, 2024. DOI: https://doi.org/10.1038/s41598-024-73932-8

[64] "Research on intrusion detection method based on transformer and CNN BiLSTM in Internet of Things," Sensors, vol. 25, no. 9, 2725, 2025. DOI: https://doi.org/10.3390/s25092725

[65] "Transformer based attention network for in vehicle intrusion detection," IEEE Access, vol. 11, 2023. DOI: https://doi.org/10.1109/ACCESS.2023.3282110

[66] J. Li, W. Zong, Y. Chow, and W. Susilo, "Mitigating class imbalance in network intrusion detection with feature-regularized GANs," Future Internet, vol. 17, no. 5, 216, 2025. DOI: https://doi.org/10.3390/fi17050216

[67] P. R. B. Houssel, P. Singh, S. Layeghy, and M. Portmann, "Towards explainable network intrusion detection using large language models," in Proc. IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), 2024, pp. 67 to 72. DOI: https://doi.org/10.1109/BDCAT63179.2024.00021

[68] "Advanced persistent threats (APT) attribution using deep reinforcement learning," Digital Threats: Research and Practice, 2025. DOI: https://doi.org/10.1145/3736654

[69] R. Buchta, G. Gkoktsis, F. Heine, and C. Kleiner, "Advanced persistent threat attack detection systems: a review of approaches, challenges, and trends," Digital Threats: Research and Practice, 2024. DOI: https://doi.org/10.1145/3688805

[70] "Advanced persistent threat (APT) and intrusion detection evaluation dataset for Linux systems 2024," Data in Brief, vol. 54, 110576, 2024. DOI: https://doi.org/10.1016/j.dib.2024.110576

[71] J. Hassannataj Joloudari, M. Haderbadi, A. Mashmool, M. Ghasemigol, S. S. Band, and A. Mosavi, "Early detection of the advanced persistent threat attack using performance analysis of deep learning," IEEE Access, vol. 8, pp. 186125 to 186137, 2020. DOI: https://doi.org/10.1109/ACCESS.2020.3029202

[72] S. Myneni, K. Jha, A. Sabur, G. Agrawal, Y. Deng, A. Chowdhary, and D. Huang, "Unraveled: a semi synthetic dataset for advanced persistent threats," Computer Networks, vol. 227, 109688, 2023. DOI: https://doi.org/10.1016/j.comnet.2023.109688

[73] A. AL Aamri, R. Abdulghafor, S. Turaev, I. Al Shaikhli, A. Zeki, and S. Talib, "Machine learning for APT detection," Sustainability, vol. 15, no. 18, 13653, 2023. DOI: https://doi.org/10.3390/su151813653

[74] M. M. Hasan, M. U. Islam, and J. Uddin, "Advanced persistent threat identification with boosting and explainable AI," SN Computer Science, vol. 4, no. 3, 271, 2023. DOI: https://doi.org/10.1007/s42979-023-01744-8

[75] R. Sommer and V. Paxson, "Outside the closed world: on using machine learning for network intrusion detection," in Proc. IEEE Symposium on Security and Privacy, 2010, pp. 305 to 316. DOI: https://doi.org/10.1109/SP.2010.25

[76] M. Ahmed, A. N. Mahmood, and J. Hu, "A survey of network anomaly detection techniques," Journal of Network and Computer Applications, vol. 60, pp. 19 to 31, 2016. DOI: https://doi.org/10.1016/j.jnca.2015.11.016

[77] I. J. Goodfellow et al., "Generative adversarial networks," Communications of the ACM, vol. 63, no. 11, pp. 139 to 144, 2020. DOI: https://doi.org/10.1145/3422622

[78] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems 30, NeurIPS 2017, 2017, pp. 5998 to 6008. Available: https://papers.nips.cc/paper/7181-attention-is-all-you-need

[79] L. Demetrio, B. Biggio, G. Lagorio, F. Roli, and A. Armando, "Explaining vulnerabilities of deep learning to adversarial malware binaries," in Proc. Third Italian Conference on Cyber Security (ITASEC), CEUR Workshop Proceedings, vol. 2315, 2019. Available: https://ceur-ws.org/Vol-2315/

[80] G. Apruzzese, P. Laskov, and J. Schneider, "SoK: pragmatic assessment of machine learning for network intrusion detection," in Proc. IEEE European Symposium on Security and Privacy, 2023. DOI: https://doi.org/10.1109/EuroSP57164.2023.00086