A Systematic Literature Review of Adversarial Domain Generation and Defense

Tomás Pelayo-Benedet[a], Ricardo J. Rodríguez[a,b,*]

[a]*Dpto. de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, Spain*
[b]*Aragón Institute of Engineering Research, Universidad de Zaragoza, Spain*

Abstract

Domain Generation Algorithms (DGAs) have long allowed malware to maintain persistent command and control channels by evading static blocklists. However, this dynamic has evolved into a sophisticated arms race: DGAs are no longer simply random but are now optimized to actively deceive detection systems. This paper presents a systematic literature review analyzing 32 primary studies (2016–2025) at the intersection of algorithmically generated domain detection and adversarial machine learning. We construct a comprehensive taxonomy of the evasion landscape, mapping the progression from simple character perturbations to advanced generative adversarial networks and semantic mimicry. Our analysis reveals two systemic flaws in the state of the art. First, we identify a significant deployment gap, where proposed defenses ignore operational realities, such as strict latency limits and the need for false positive rates below 0.1%. Second, we highlight a serious reproducibility crisis driven by a lack of public code and standardized datasets. We conclude by proposing a roadmap to standardize assessment frameworks and bridge the gap between theoretical soundness and operational feasibility.

Keywords: Domain Generation Algorithm, Adversarial Learning, Generative Adversarial Networks, Systematic Literature Review

1. Introduction

In today's digital landscape, cybercrime has become a highly lucrative and well-structured industry, reportedly generating more revenue than the global drug trade (Cybersecurity Ventures, 2025). This profitability fuels a continuous arms race: attackers develop increasingly sophisticated malware to maximize their return on investment, while security analysts strive to innovate in prevention, detection, and response mechanisms.

A critical component of modern malware operations is the ability to maintain a persistent communication channel with the attacker, a phase termed *Command and Control* (C&C) in the MITRE ATT&CK framework (MITRE) and the Cyber Kill Chain (Lockheed Martin). Through this channel, adversaries issue commands, exfiltrate sensitive data, or coordinate lateral movements within compromised networks. Traditionally, defenders disrupted these channels by blocklisting static IP addresses or domains embedded in malware. However, modern adversaries easily circumvent this static approach (Yong Wong et al., 2021).

To bypass simple blocklists, attackers have adopted Domain Generation Algorithms (DGAs), which were initially popularized by the `Conficker` worm (Porras et al., 2009). DGAs allow malware to dynamically generate a massive volume of pseudo-random domain names, known as *Algorithmically Generated Domains* (AGDs). By registering only a small fraction of these domains, attackers maintain communication while forcing defenders to inspect thousands of potential candidates, thus rendering static blocklisting ineffective.

In response to DGA proliferation, academic and industrial communities have spent the last decade developing advanced detection systems. While initial approaches relied on statistical feature engineering, the state of the art has shifted decisively toward Machine Learning (ML), particularly Deep Learning, techniques (Woodbridge et al., 2016; Yu et al., 2017; Drichel et al., 2024b). These models have demonstrated remarkable accuracy in distinguishing benign domains from malicious AGDs by learning complex patterns within character sequences.

However, the widespread adoption of ML-based detection systems has introduced a new vulnerability. Alongside the evolution of defensive models, attackers have begun to employ adversarial techniques to evade them. This creates a new paradigm where a DGA is no longer designed solely for randomness but is specifically optimized to deceive the detector. By leveraging techniques such as Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) or gradient-based perturbations (Wang et al., 2021), adversaries can generate domains that mimic the statistical and structural properties of benign domains, driving a complex "arms race" between offensive generation and defensive identification (Anderson et al., 2016; Sidi et al., 2020).

While researchers have made significant progress in AGD detection and adversarial attacks individually, there is no synthesis at the intersection of these two fields. While broader reviews exist on anomaly detection (Sabuhi et al., 2021) or general GAN applications (Nayak et al., 2024), the literature specifically addressing adversary techniques in the DGA domain remains scattered. Currently, there is no comprehen-

*Corresponding author
Email addresses: `tpelayo@unizar.es` (Tomás Pelayo-Benedet), `ricardo@unizar.es` (Ricardo J. Rodríguez)

sive overview summarizing the evasion techniques used against AGD detectors, nor a clear assessment of the operational feasibility of proposed defenses in production environments.

To address this gap, we present a Systematic Literature Review (SLR) on adversarial techniques applied to AGD detection. We follow a rigorous methodology Kitchenham et al. (2009); Siddaway et al. (2019) to analyze the current state of the art, aiming to understand how attackers use generative techniques to evade ML-based security systems and, conversely, how researchers are strengthening their models to resist these attacks.

In summary, this paper makes the following contributions:

- We conduct a systematic review of the literature on the intersection of AGD detection and adversarial techniques, analyzing 32 primary studies.
- We propose a comprehensive taxonomy of adversarial techniques in AGD detection, mapping their evolution from early proof-of-concept models to sophisticated semantic and gradient-based evasion methods.
- We identify a critical deployment gap between academic research and operational reality. Our analysis reveals that most proposed defenses fail to address real-world constraints, such as stringent latency requirements and false positive rates below 0.1%.
- We expose a serious reproducibility crisis in this field, analyzing the availability of public code and datasets. We analyze the impact of this lack of transparency on scientific progress and provide a roadmap for standardizing future assessment frameworks.

The rest of this paper is organized as follows. Section 2 provides background on DGAs and GANs. Section 3 discusses related work, differentiating our contribution from previous studies. Section 4 details the methodology used for this SLR, including the research questions and inclusion criteria. Section 5 presents a synthesis of the analyzed studies, detailing the taxonomy of attacks and defenses. Section 6 analyzes the implications of the findings, addressing specific research questions and identifying remaining challenges. Section 7 addresses the limitations of this work. Section 8 describes emerging trends and directions for future research. Finally, Section 9 concludes this work.

2. Background

This section provides the necessary theoretical foundations for understanding the intersection of AGD and adversarial machine learning techniques. We first introduce DGAs and their classification, followed by an overview of GANs and their specific application to discrete data generation.

2.1. Domain Generation Algorithms

A *Domain Generation Algorithm* (DGA) is a mechanism employed by malware to periodically generate a large volume of domain names, known as *Algorithmically Generated Domains* (AGDs). These algorithms serve as a critical resilience mechanisms for botnets, allowing infected hosts to dynamically locate their Command and Control (C&C) servers instead of relying on static, hard-coded IP addresses or domains (MITRE).

To ensure successful communication, the malware and the C&C server synchronize the generation of domains using a shared *seed* value (e.g., the current date, exchange rates, or a constant). This seed initializes the pseudo-random logic of the algorithm. Since the first operational use of DGAs by the `Conficker` family in 2008 (Porras et al., 2009), these techniques have evolved significantly in complexity to evade detection.

Researchers generally categorize DGAs based on their generation schemes. According to the comprehensive taxonomy proposed by Plohmann et al. (2016), the most common types include: (i) *arithmetic-based algorithms*, which employ mathematical operations to calculate numeric sequences, which are then mapped to ASCII characters to form the domain string; (ii) *hash-based algorithms*, which apply cryptographic hash functions (e.g., MD5, SHA-1, or SHA-256) to the seed, generating a hexadecimal digest that acts directly as a domain name; (iii) *dictionary-based algorithms*, which concatenate words extracted from embedded dictionaries to reduce entropy and evade statistical detection techniques; and (iv) *permutation-based algorithms*, which modify existing legitimate domain names by swapping characters or rearranging substrings.

While traditional detection systems effectively identify arithmetic or hash-based DGAs by analyzing entropy and lexical features, dictionary-based and hybrid variants pose a greater challenge. This resilience has motivated the decisive shift toward machine learning techniques.

2.2. Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. (2014), provide a framework for estimating generative models using an adversarial process. The architecture consists of two neural networks trained simultaneously in a zero-sum game: a *Generator* (G) and a *Discriminator* (D).

The goal of G is to capture the underlying data distribution p_{data} and produce synthetic samples $G(z)$ from a noise vector z that are indistinguishable from the real data. Conversely, D is a binary classifier optimized to differentiate between real samples from the training dataset and spurious samples generated by G.

In the context of AGD detection, the application of GANs represents a paradigm shift: from static generation logic to adaptive learning. Unlike traditional DGAs described in Section 2.1, which rely on fixed algorithms, GANs have the capacity to approximate the underlying statistical distribution of legitimate domain names. This capability allows malware to generate *adversarial examples*: domains specifically optimized to deceive detection classifiers while preserving the structural characteristics of benign domain names.

3. Related Work

Detecting AGDs remains a persistent and constantly evolving challenge in cybersecurity. As defenders innovate, attackers continually refine their DGAs, requiring defense strategies to evolve in parallel, from static heuristics to advanced Deep Learning (DL) architectures (Plohmann et al., 2016).

Early research laid the groundwork for automated detection through statistical feature engineering. The `EXPOSURE` system (Bilge et al., 2011, 2014) pioneered passive Domain Name System (DNS) analysis by correlating domain-specific features with temporal behavioral patterns. Building on this foundation, `Pleiades` (Antonakakis et al., 2012) introduced nonexistent domain response analysis (`NXDOMAIN`) to identify AGDs, while `Phoenix` advanced classification capabilities by leveraging linguistic and Internet Protocol features to categorize AGDs into distinct malware families (Schiavoni et al., 2014). These seminal works demonstrated the feasibility of feature-based machine learning but relied heavily on manual feature extraction.

A paradigm shift occurred with the application of deep learning, which enabled models to learn features directly from raw domain strings. Woodbridge pioneered this approach by applying Long-Term Memory Networks (LSTMs) to AGD detection, demonstrating superior performance in both binary and multiclass classification without manual feature engineering. Subsequently, Yu et al. (2017) provided a critical comparative analysis of LSTMs versus Convolutional Neural Networks (CNNs), demonstrating that convolutional layers could effectively capture local character patterns characteristic of machine-generated domains. This work laid the foundation for the current state of the art in detection.

However, the effectiveness of these DL models precipitated a counter-response: the adoption of adversarial machine learning by attackers. By leveraging GANs, adversaries can now synthesize domains specifically optimized to mimic the statistical distributions of benign domain names, thereby evading classifiers previously considered robust. This development has transformed AGD detection into a dynamic arms race, where static datasets and fixed models are rapidly becoming obsolete due to adaptive generative attacks.

Despite the critical nature of this threat, the literature lacks a consolidated analysis of this intersection. While there are extensive systematic reviews on anomaly detection (Sabuhi et al., 2021), logistic regression applications (Akram et al., 2021), and general GAN implementations (Nayak et al., 2024), there is a notable lack of reviews specifically addressing adversarial techniques in the DGA domain. This work addresses this deficiency through a SLR focused exclusively on generative evasion techniques. We provide a comprehensive taxonomy of the current state of the art, map the evolution of adversarial domain generation, and identify operational limitations that hinder the deployment of robust defenses.

4. Methodology

We conduct an SLR following the guidelines proposed by Kitchenham et al. (2009). SLRs offer significant advantages over traditional reviews by adhering to a rigorous and predefined protocol. This approach ensures reproducibility and minimizes potential bias, providing a comprehensive compilation of research findings through an objective and transparent process that allows for the identification of patterns, gaps, and trends, while maintaining methodological rigor (Siddaway et al., 2019).

This section details the methodology employed in this work. First, we present the research questions that guided our investigation. Next, we describe the search process, including the selection of scientific databases and the specific search terms used. We then describe the inclusion and exclusion criteria applied to filter the results. Finally, we present a quantitative overview of the article selection process and the final corpus of included studies.

4.1. Research Questions

The main objective of this research is to analyze the literature on the intersection of adversarial generative techniques and AGD detection. Specifically, we seek to understand the underlying principles of AGD evasion, the role of deep learning in this "arms race," and the reproducibility of the current state of the art. Formally, we address the following Research Questions (RQ), discussed in detail in Section 6:

RQ1.- What are the main adversarial techniques used to evade AGD detection, and how have they evolved to circumvent modern machine learning-based systems?

RQ2.- What specific features of AGDs or detection models are most susceptible to adversarial attacks, and how have countermeasures evolved to address these vulnerabilities?

RQ3.- What specific roles do DL architectures play in both adversary domain generation and detection?

RQ4.- What metrics, methodologies, and comparative frameworks are currently used to evaluate the robustness of adversary techniques against AGD detectors?

RQ5.- To what extent does the community adhere to open science practices, such as open source, dataset sharing, and documentation of experimental parameters?

RQ6.- Which defense strategies have proven most effective against adversary DGA attacks, and what is their operational feasibility in real-world scenarios?

RQ7.- What are the main computational challenges, limitations, and trade-offs when implementing adversary techniques against AGD detection systems?

4.2. Search Process

To ensure comprehensive coverage of the relevant literature, we selected four major scientific databases that index leading

conferences and journals in computer science and cybersecurity, namely: *IEEE Xplore*, *ACM Digital Library*, *Science Direct*, and *Scopus*. These platforms were chosen for their high impact factor and their compatibility with metadata export for systematic analysis.

We based the search strategy on selected keywords to capture the intersection of generative adversarial techniques and algorithmic domain generation. The final search string was defined as follows:

> *(GAN OR "Adversarial Networks" OR "Generative Adversarial" OR "Adversarial Learning") AND (DGA OR "Domain Generation Algorithm" OR AGD OR "Algorithmically Generated Domain" OR Botnet)*

We restricted the search to publications from 2008 to 2025. The starting year was selected to coincide with the first operational use of DGAs by the `Conficker` Porras et al. (2009) worm family.

Since this SLR focuses on high-quality scholarly contributions, we excluded gray literature (e.g., technical blogs, white papers, or non peer-reviewed publications). Finally, to mitigate potential indexing shortcomings, we performed a backward snowballing process (reference inspection) on the initially selected articles. This complementary step identified three additional relevant studies that were not included in the initial search results.

4.3. Scope and Adversarial Definition

To ensure analytical precision, we explicitly establish the operational definition of "*adversarial*" adopted in this review, addressing the ambiguity between strict mathematical optimization and practical evasion. While definitions in classical Adversarial Machine Learning typically require gradient-based optimization against a differentiable loss function (Goodfellow et al., 2014; Madry et al., 2018), this SLR employs a broader evasion-oriented interpretation.

We apply the following decision rule for inclusion: a technique is considered adversarial if it exhibits (i) explicit evasion intent, specifically targeting the decision boundary of a detector rather than merely generating random noise; and (ii) adaptive mimicry, which attempts to approximate the statistical distribution of benign traffic. Consequently, we include heuristic and black-box techniques, recognizing them as adversarial baselines that effectively challenge detection logic without requiring access to internal gradients.

4.4. Inclusion and Exclusion Criteria

After the initial identification of candidate articles, we used the `StArt` tool (Zamboni et al., 2010; Hernandes et al., 2012) to streamline the selection workflow. `StArt` facilitates the systematic review lifecycle by automating duplicate detection and providing a keyword-based scoring mechanism to prioritize relevant studies.

Table 1: Inclusion (IC) and Exclusion (EC) Criteria.

Type	**Criterion**
IC1	The article explicitly investigates the intersection of adversarial machine learning and domain generation algorithms.
IC2	The main contribution is a novel technique for generating adversarial domains or a defense mechanism specifically designed to resist them.
EC1	The article is a short introductory paper, an early access draft, or a conference abstract/poster.
EC2	The article is less than 6 pages (excluding references).
EC2	The article is not written in English.
EC3	The article is a duplicate entry.
EC4	The article focuses on general AGD detection without specific adversarial generation or evasion component.
EC5	The full text is not accessible through standard academic repositories.

We used `StArt`'s scoring feature as a preliminary filter to rank the retrieved articles by relevance. This prioritization strategy assigns a weighted score based on the frequency and location of specific keywords within each article's metadata. The scoring metric was defined as follows: 5 points if the keyword appears in the article's title; 3 points if the keyword appears in the article's abstract; 2 points if the keyword appears in the article's keyword list.

The set of keywords used for this scoring process included terms encompassing both adversarial learning and domain generation aspects, such as "adversarial", "gan", "dga", "dga-based", "domain generation algorithm", "agd", "algorithmically generated domain", "malicious domain", "command and control", "novel dga", and "botnet". Variations and plural forms of these terms were also considered.

Articles that scored above the empirical threshold (set at 10 points) were subsequently reviewed manually according to the formal Inclusion (IC) and Exclusion Criteria (IC) described in Table 1. This rigorous filtering ensured that only studies directly addressing our RQs were retained for final analysis.

4.5. Articles Collected and Reviewed

Executing our search protocol yielded 1360 candidate articles across the four selected scientific databases. These records were processed using the preliminary scoring filter described above. This automated prioritization phase reduced the corpus to 182 potentially relevant studies that met or exceeded the established relevance threshold.

We then applied the formal inclusion and exclusion criteria to this subset. This manual selection resulted in the elimination of 141 articles. This process left 34 articles for full-text inspection.

After a detailed reading, we excluded five additional articles for not meeting the technical depth required by IC2.

Finally, to ensure comprehensiveness, we performed a backward snowball sampling procedure (reference list inspection) on the eligible articles. This step identified three additional relevant studies that were not included in the initial search results. Consequently, the final corpus selected for quantitative synthesis and in-depth analysis consisted of 32 primary studies.

Figure 1 illustrates the complete selection process, detailing the number of articles selected and rejected at each stage of the SLR protocol, in accordance with the PRISMA guidelines (Page et al., 2021).

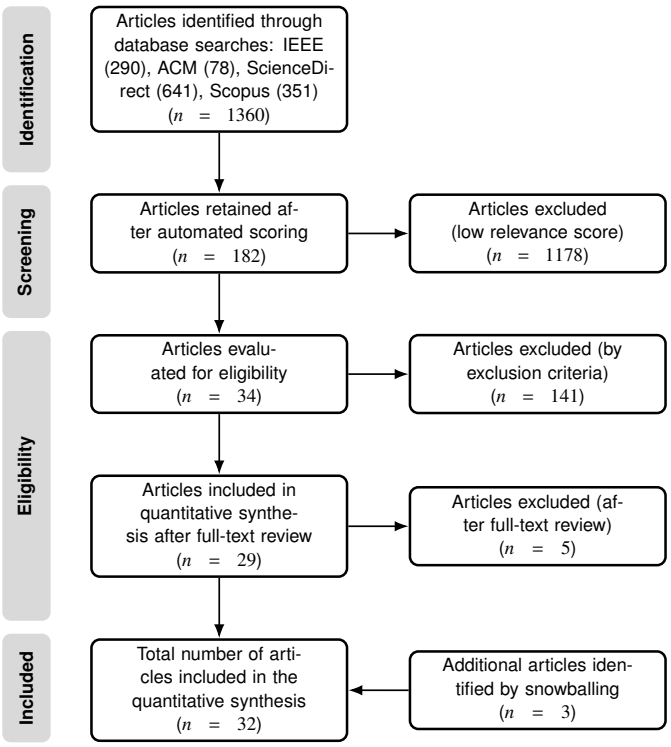

Figure 1: PRISMA flowchart illustrating the study selection process, from initial identification to final inclusion.

4.6. Threat Model and Operational Constraints

To systematically analyze adversarial scenarios found in the literature, we explicitly define a formal threat model based on the attacker's knowledge and the defender's operational constraints (Xiong and Lagerstrom, 2019).

Attacker Capabilities and Knowledge. We classify the adversary's perspective into two distinct categories. In the White-Box (WB) scenario, the adversary has complete knowledge of the target detection model, including its architecture, parameters, and gradients. This allows for the use of optimization attacks to mathematically calculate precise perturbations. Conversely, in the Black-Box (BB) scenario, the adversary does not have access to the model's internal weights or gradients. The attacker relies solely on query feedback (e.g., accepted/rejected status) or surrogate models to guide the generation of evasive domains, employing techniques such as reinforcement learning or heuristic mimicry.

Defender Observability and Constraints. We assume that the defender observes incoming DNS queries but operates under strict production constraints. To assess the feasibility of the defenses described in Section 5.3, we adopt specific thresholds as a basis. Regarding query latency, the defense mechanism must operate within a limit of < 100 ms per query to avoid network bottlenecks. Furthermore, to mitigate alert fatigue and ensure operational feasibility, the False Positive Rate (FPR) must be kept strictly below 0.1%.

4.7. Reproducibility Scoring Rubric

To systematically assess the transparency and replicability of the reviewed studies, we developed a standardized scoring rubric, which evaluates the extent to which authors provide sufficient artifacts to facilitate independent verification. Based on best practices in reproducible machine learning research (Pineau et al., 2021), the rubric defines five disclosure criteria: (i) *implementation availability*: public access to source code (e.g., GitHub, GitLab) or supplementary material containing accompanying scripts used for generation and detection; (ii) *preprocessing transparency*: documentation of data cleaning, feature extraction, and the split between training and test; (iii) *detailed architectures*: explicit description of layers, neuron counts, and structural components required for model reconstruction; (iv) *hyperparameter disclosure*: reports of training configurations, including learning rates, batch sizes, optimizers, and loss functions; and (v) *stochastic control*: explicit mention to random seeds to ensure deterministic initialization and training stability.

Based on compliance with these criteria, we categorized the studies on a three-level reproducibility scale. We assigned *high reproducibility* to studies that provide public source code, accessible datasets, and complete parameter documentation, allowing for exact replication. Studies are classified as *medium reproducibility* when source code is not included, but the work provides sufficient detail about the architecture and experimental setup to allow for approximate independent reimplementation. Finally, we categorized studies as *low reproducibility* if they effectively function as a "black box," lacking critical details about the model structure, hyperparameters, or data processing, thus preventing scientific replication.

5. Analysis of Results

In this section, we summarize the main findings derived from our systematic analysis of the 32 primary studies. To provide a holistic view of the adversarial DGA landscape, we structure our assessment along four analytical dimensions. First, we establish a taxonomy of generation techniques, mapping the technological evolution from simple perturbation heuristics to sophisticated deep learning architectures. Second, we deconstruct the mechanisms of evasion, identifying the specific structural

and statistical vulnerabilities exploited in current detection systems. Third, we assess the effectiveness of defensive countermeasures, specifically analyzing the trade-offs between adversarial robustness and operational performance. Finally, we examine experimental methodology, highlighting systemic shortcomings in dataset heterogeneity, metric selection, and reproducibility that currently hinder scientific progress in this field.

Unless otherwise stated, we calculate all percentages in this analysis relative to the total corpus of primary studies ($n = 32$). Regarding categorization, the taxonomy of techniques (see Section 5.1) imposes a mutually exclusive classification based on the study's main contribution. In contrast, the subsequent analyses of evasion mechanisms and defenses (Sections 5.2–5.5) allow for overlapping categories, recognizing that a single study may employ hybrid strategies or report multiple metrics simultaneously.

5.1. Taxonomy of Adversarial Techniques

Adversarial techniques applied to AGD detection has undergone a distinct evolutionary trajectory since 2016, characterized by increasing sophistication in both generation strategies and evasion logic. This section presents a comprehensive taxonomy of these techniques, analyzing their temporal distribution, architectural characteristics, and methodological approaches. Table 2 details the full corpus of reviewed works, classifying them by technique type, target architecture, code availability, datasets, and proposed defense mechanisms.

As shown in Figure 2, research activity in this field has accelerated significantly during the reviewed period. The field experienced a notable surge in 2020, accounting for approximately 25.00% of the main studies. This peak reflects widespread adoption and development of fundamental GAN-based frameworks (Anderson et al., 2016; Fu et al., 2017) as the standard for synthetic malicious traffic generation.

The following years (2023–2025) marked a qualitative shift toward defensive resilience and the integration of attention mechanisms (Drichel et al., 2024b; Nie et al., 2024; Luo et al., 2026). This transition suggests that while offensive generation is maturing, the focus is shifting toward overcoming the limitations of long-range dependency modeling. Furthermore, recent contributions demonstrate ongoing innovation, moving beyond simple recurrences to complex game-theory models (Nie et al., 2024), coordinated multi-instance attacks (Nazzal et al., 2024), and the adoption of transformers in GAN-based architectures (Pregardier et al., 2025; Luo et al., 2026) to master semantic mimicry and global structural coherence.

The distribution pattern illustrated in Figure 2 reveals three distinct evolutionary phases: (i) an emergence phase (2016–2019), focused primarily on proof-of-concept demonstrations and initial feasibility studies (Anderson et al., 2016; Fu et al., 2017; Peck et al., 2019); (ii) an expansion phase (2020), representing the peak of research activity. This year accounts for approximately 25.00% of the total corpus and is characterized by rapid methodological diversification; and (iii) a maturation phase (2021–2025), distinguished by the dual development of offensive sophistication (e.g., game theory, semantic attacks, transformers) and defensive resilience (Hu et al., 2023; Nazzal et al., 2024; Nie et al., 2024). After the peak in 2020, the volume of publications stabilized at approximately 3 to 5 annual contributions, suggesting a paradigm shift from exploratory proliferation toward highly specialized and robust research vectors.

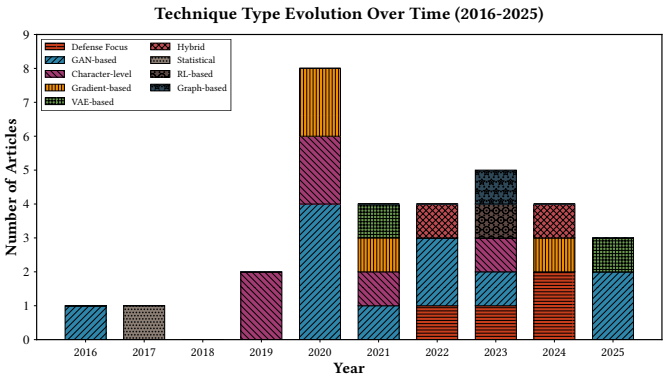

Figure 2: Temporal distribution of adversarial DGA publications (2016–2024). The histogram illustrates the annual frequency of primary studies, highlighting a significant peak in research activity during 2020.

To systematically analyze this diverse landscape, we organize the identified works according to a hierarchical taxonomy based on two dimensions: *Adversary Knowledge* and the primary *Generation Mechanism*. As shown in Figure 3, the field is first stratified according to attacker visibility: *White-Box* methods, which leverage internal gradients, and *Black-Box* approaches, restricted to query feedback. Within the latter, we identify three distinct generation paradigms. *Distribution Learning* approximates the probability distribution of benign traffic to synthesize plausible samples. *Loss Maximization* frames generation as an optimization task, employing gradient descent or reinforcement learning to maximize detector error. Finally, *Rule-Based and Heuristic* methods rely on deterministic strategies, such as character substitution or structural mutation, to exploit specific vulnerabilities without representational learning.

5.1.1. Distribution Learning Approaches

This category encompasses techniques whose primary objective is to approximate the probability distribution of benign domain names. Whether employing deep neural architectures or classical stochastic processes, these methods seek to sample new domains from the learned high-density regions of legitimate traffic, aiming to make them statistically indistinguishable from benign samples.

Adversarial Networks. GAN-based methods have evolved from unstable vanilla architectures to sophisticated, controllable systems. Table 3 summarizes the evolutionary progression of GAN architectures identified in the literature.

Anderson et al. (2016) introduced `DeepDGA`, the first architecture explicitly optimized to bypass AGD classifiers through character-by-character generation. Employing an autoencoder pre-training strategy, it reduced detection rates from 98.00% to 50.00% compared to traditional DGAs. However, these early

Table 2: Comprehensive overview of adversarial techniques on DGAs. The table highlights the scarcity of open source implementations and the evolution from character-level perturbations to GAN and Gradient-based attacks.

Reference	Year	Technique Type	Threat Model	Architecture	Target Detector	Open Source	Dataset	Defense Proposed
Anderson et al. (2016)	2016	GAN-based	BB	Char-AE-GAN	LSTM/RF	✗	Alexa + 10 DGA families	✓
Fu et al. (2017)	2017	Statistical	BB	HMM/PCFG	Statistical KL/ED/JI	✗	IPv4 rDNS	✓
Peck et al. (2019)	2019	Character-level	BB	2-Char Random Subst.	LSTM.MI/B-RF	✗	Alexa + Bambenek	✓
Spooren et al. (2019)	2019	Character-level	BB	Statistical Mimicry	RF (FANCI)/LSTM	✗	Alexa + DGArchive	✓
Sivaguru et al. (2020)	2020	Character-level	BB	CharBot (2-char mod)	LSTM.MI/B-RF	✗	Real DNS Traffic	✓
Sidi et al. (2020)	2020	Gradient-based	WB	Jacobian Saliency Map	LSTM/biLSTM/CNN	✓	DMD-2018	✓
Yilmaz et al. (2020)	2020	Gradient-based	WB	Gradient Descent+Cosine	LSTM (2-layer)	✗	Majestic + DGArchive	✓
Alaeiyan et al. (2020)	2020	Character-level	BB	Genetic Programming	RF (100 trees)	✗	Alexa + MasterDGA	✓
Charan et al. (2020)	2020	GAN-based	BB	CTGAN	C5.0 Ensemble	✗	Alexa + DGArchive	✓
Gould et al. (2020)	2020	GAN-based	BB	WGAN-GP+Softmax	CNN-LSTM Hybrid	✗	Tranco + 15 DGA families	✓
Yun et al. (2020)	2020	GAN-based	BB	WGAN-GP+ResNet	LSTM/RF/Graph	✗	Alexa + IPv4 rDNS	✓
Cao et al. (2020)	2020	GAN-based	BB	Vanilla GAN+LSTM	Bayes/J48/DT/RF	✗	Alexa + 360NetLab	✓
Zheng et al. (2021)	2021	GAN-based	BB	WGAN-GP+1D-CNN	FANCI/Endgame/NYU/MIT	✗	Tranco + DGArchive	✓
Wang and Guo (2021)	2021	VAE-based	BB	VAE+GCNN	FANCI/LSTM/BiLSTM	✗	Alexa + DGArchive	✗
Liu et al. (2021a)	2021	Gradient-based	BB	Geometric Vector Pert.	Endgame/CMU/MIT/Invincea	✗	Alexa + 3 DGA families	✗
Liu et al. (2021b)	2021	Character-level	BB	Influence Score Ranking	LSTM/CNN (char-level)	✗	Alexa + 360NetLab	✓
Zhai et al. (2022)	2022	GAN-based	BB	WGAN-GP+AE+NLP	LSTM/RF (FANCI)	✗	Tranco + DGArchive	✗
Shu et al. (2022)	2022	Hybrid	BB	VAE+GCNN+n-gram	Endgame/Invincea/MIT	✗	Tranco + HYDRA	✓
Zhang et al. (2022)	2022	GAN-based	BB	WGAN+ASCII Encoding	Bayes/J48/KNN/Bagging	✗	Alexa + 11 DGA families	✓
Suryotrisongko et al. (2022)	2022	Defense Focus	WB	FGSM/PGD/BIM	Quantum-Classical DL	✗	Alexa + 10 DGA families	✓
Hu et al. (2023)	2023	Character-level	BB	BiLSTM Semantic Model	FANCI/B-RF/LSTM.MI	✗	Alexa + OSINT + 360NetLab	✓
Ravi et al. (2023)	2023	Defense Focus	BB	DeepDGA/CharBot/MaskDGA	B-LSTM/B-GRU	✗	Homograph + AmritaDGA	✓
Ren et al. (2023)	2023	GAN-based	BB	WGAN+T5+BERT	HMM/FANCI/LSTM/BERT	✗	Alexa + 360DGA	✓
Gao et al. (2023)	2023	Graph-based	BB	Graph Adjacency Attack	hetGNN (HGNN)	✗	Alexa + Real DNS	✓
Nie et al. (2023)	2023	RL-based	BB	LSTM+Policy Gradient	LSTM/BiLSTM/CNN/RF	✓	Alexa + DGArchive	✓
Drichel et al. (2024a)	2024	Defense Focus	WB	PGD/HotFlip/C&W/BAT	ResNet (CNN)	✓	Real NXDomains + DGArchive	✓
Nie et al. (2024)	2024	Defense Focus	BB	Game Theory+MIRF	Stats/WordGraph/FANCI	✗	Alexa + DGArchive	✓
Nazzal et al. (2024)	2024	Gradient-based	BB	2-hop GCN Surrogate	hetGNN (meta-path)	✓	Real Enterprise DNS	✓
Selvaraj and Panjanathan (2024)	2024	Hybrid	BB	cWGAN+RCNN-BiLSTM	RCNN-BiLSTM/CNN-BiLSTM	✗	Tranco + UMUDGA + UTL_DGA	✓
Nangong and Wu (2025)	2025	VAE-based	BB	VAE+Transformer+LSTM	FANCI/Endgame	✗	Alexa + 360NetLab	✗
Pregardier et al. (2025)	2025	GAN-based	BB	Transformer-AE-GAN	FANCI/LSTM/BILBO	✗	Tranco + DGArchive	✓
Luo et al. (2026)	2025	GAN-based	BB	SAGAN (Self-Attention)	Endgame/MIT/BiLSTM	✗	Alexa + DGArchive	✗

WB=White-Box approach (gradients/parameters are known); **BB**=Black-Box approach (query-only/surrogate).

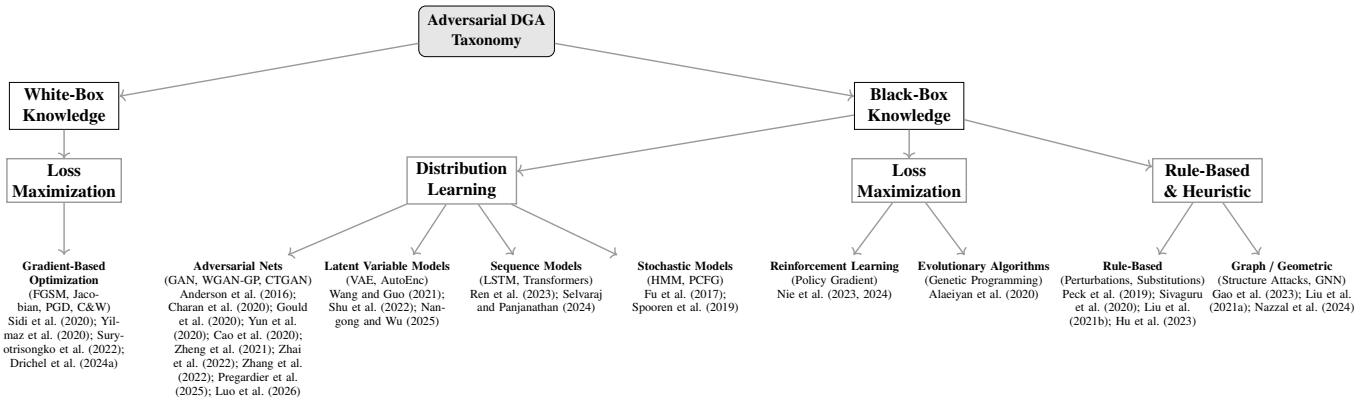

Figure 3: Taxonomy of adversarial DGA techniques organized by Adversary Knowledge (Level 1) and Generation Mechanism (Level 2). The hierarchy contrasts White Box strategies (which exploit internal gradients) with Black Box approaches, which are stratified into Distribution Learning (which mimics benign statistics), Loss Maximization (which optimizes evasion through feedback), and Rule-Based heuristics (deterministic modifications).

Table 3: Summary of GAN-based adversarial DGA architectures and their performance characteristics.

Architecture	#	Best Evasion	Main Limitation
Vanilla	3	50.00%	Training instability
Wasserstein	5	79.00%	Computational cost
Conditional	1	89.90%	Implementation complexity
Specialized	1	75.00%	Limited generalization
Self-Attention	2	95.35%	Inference latency

models suffered from severe training instability and mode collapse issues (Yun et al., 2020; Cao et al., 2020; Ravi et al., 2023). Charan et al. (2020) addressed domain type limitations by applying CTGAN to generate synthetic word-list DGAs, demonstrating that GANs could target specific semantic structures beyond random character strings.

To mitigate instability, researchers adopted Wasserstein GANs with Gradient Penalty (WGAN-GP). Yun et al. (2020) developed `Khaos`, the first n-gram-based WGAN-GP, which achieved superior stability. Subsequent systems demonstrated significant efficiency gains: Zheng et al. (2021) reported a reduction in training time from 14 hours (`DeepDGA`) to 37 minutes while generating domains with richer pattern diversity (872 vs. 545 bigrams). Similarly, Zhang et al. (2022) utilized WGANs for data augmentation, learning directly from 11 real DGA families to bolster training sets rather than solely generating evasive samples.

Advanced architectures have incorporated controllability mechanisms to enforce constraints. Zhai et al. (2022) proposed `CDGA`, combining WGAN-GP with controllable text generation to achieve zero repetition rates. Recently, the focus shifted towards capturing long-range dependencies. Luo et al. (2026) introduced `SADGA`, utilizing Self-Attention GANs to extract global features often missed by convolutional layers. Similarly, Pregardier et al. (2025) proposed `TITAN DGA`, unifying a Transformer-based autoencoder with a GAN and employing adversarial self-augmentation to iteratively refine the evasiveness of generated domains.

Latent Variable Models. Although less prevalent than GANs, Variational Autoencoders (VAEs) offer a probabilistic alternative for generation. Wang and Guo (2021) developed `NDG`, a VAE-based architecture that constructs a smooth latent space of domain characteristics. Unlike the adversarial game of GANs, `NDG` optimizes the evidence lower bound to generate domains. This approach proved effective against legacy detectors, with `NDG` domains evading detection 10.00%–16.00% of the time, significantly outperforming older HMM-based methods and vanilla GANs, though yielding lower evasion rates than state-of-the-art WGANs due to the characteristic "blurriness" of samples generated from the latent bottleneck.

Recently, Nangong and Wu (2025) enhanced this approach by embedding semantic information and integrating Transformer encoders within the VAE framework, alongside an S-type KL annealing strategy to mitigate the posterior collapse problem common in these architectures (Bowman et al., 2016).

Sequence-Based and Attention-Based Models. Recent research focuses on architectures that explicitly model sequential dependencies using Recurrent Neural Networks (RNNs) or Transformers, which function as sophisticated generators in adversarial environments.

Selvaraj and Panjanathan (2024) integrated Natural Language Processing (NLP) techniques by combining RCNN-BiLSTM networks with Gumbel Softmax relaxation. This architecture allows for discrete text generation while maintaining differentiability, resulting in word-level adversarial domains with an exceptionally low repetition rate (2.04%). Recent innovations also integrate NLP adaptations; Ren et al. (2023) developed `CL-GAN`, which uses T5-based generators (*Text-to-Text Transfer Transformers*) and teacher-student knowledge distillation. This continual learning approach prevents catastrophic forgetting, maintaining an F1 score of 75.00% on sequential learning tasks, where standard models dropped to 67.00%.

Statistical and Stochastic Modeling. Early adversarial generation techniques relied on mimicking the statistical properties of benign traffic rather than learning them using neural networks. Fu et al. (2017) pioneered the use of Hidden Markov Models

(HMMs) and context-free probabilistic grammars trained on legitimate IPv4 records. These stochastic models significantly outperformed contemporary randomness-based botnets by producing domains that statistically resembled human-generated strings. Spooren et al. (2019) further validated this approach with statistical mimicry, demonstrating that non-neural probabilistic methods could effectively evade feature-based detectors such as `FANCI` without the training overhead of deep learning.

5.1.2. Loss Maximization and Search Approaches

Unlike distribution learning, which aims to approximate benign statistics, optimization approaches frame evasion as a search problem that seeks to maximize the detector's loss function. We classify these works based on the adversary knowledge: *White-Box* methods leverage differentiable gradients to compute optimal perturbations, while *Black-Box* methods employ Reinforcement Learning (RL) or evolutionary strategies to navigate the decision boundary solely through query feedback.

Gradient-Based Optimization (White-Box). When the attacker has access to the detector's gradients, they can derive perturbations analytically using backpropagation instead of relying on stochastic search. Sidi et al. (2020) introduced `MaskDGA`, which uses Jacobian-based salience maps to identify the characters most influential in the classifier's decision. By perturbing only these high-impact features, the method achieved a true positive rate (TPR) as low as 5.00% with a strict false positive rate (FPR) of 1.00%, significantly outperforming random noise injection. Similarly, Yilmaz et al. (2020) applied gradient descent combined with cosine similarity to guide character-level replacements, reducing the detector's accuracy from 98.00% to 76.00%. More recently, Drichel et al. (2024a) extended this domain by evaluating advanced optimization attacks, such as projected gradient descent (Madry et al., 2018), HotFlip (Ebrahimi et al., 2018), and Carlini & Wagner (Carlini and Wagner, 2017), demonstrating that gradient-guided perturbations remain the most efficient evasion vector against differentiable classifiers.

Reinforcement Learning (Black-Box). When gradient information is unavailable, researchers employ agent-based learning to probe the detector. Nie et al. (2023) proposed `PKDGA`, a RL framework that leverages LSTM policy networks and Monte Carlo tree search. By interacting with the detector solely through query-response feedback, the agent learns to construct evasive sequences, achieving a 24.52% improvement over comparable baselines. This highlights the ability of RL to exploit the limitations of black-box decision boundaries. This concept was refined by Nie et al. (2024), who introduced a game-theory-based modeling approach that allows the attack strategy to dynamically adapt to changing defensive postures.

Evolutionary Algorithms (Black-Box). Unlike gradient-based methods, evolutionary approaches leverage the principles of biological selection to navigate the discrete, non-differentiable domain name space. Alaeiyan et al. (2020) applied this paradigm using `GADGA`, employing genetic algorithms with fitness functions derived from pronunciation scores and tests of statistical randomness. This approach demonstrated almost total evasion against traditional classifiers (reducing detection accuracy to 0.47%), showing that evolutionary search can efficiently generate adversarial samples without the computational overhead of gradient calculations or model training.

5.1.3. Rule-Based and Structural Heuristics

This category encompasses deterministic methods that rely on explicit rule sets, geometric properties, or domain-specific knowledge, rather than learning a statistical distribution or optimizing a loss function. These techniques typically exploit specific structural blind spots in the detection logic with minimal computational overhead.

Rule-Based and Character Perturbation. Heuristic perturbations typicallyoften produce high evasion rates with negligible computational cost. Peck et al. (2019) developed `CharBot`, a heuristic engine that randomly substitutes two characters in benign domains. Despite its simplicity, it achieved evasion rates comparable to those of complex GANs, requiring a fraction of the resources (e.g., 162 KiB vs. 6.5 MiB for `DeepDGA`). Similarly, Liu et al. (2021b) introduced `CLETer`, which calculates an influence score to identify and modify high-impact characters, and Hu et al. (2023) proposed `ReplaceDGA`, which uses semantic relationships to guide substitutions. Taken together, these works illustrate the fragility of current detectors, demonstrating that low-cost, rule-based noise is often sufficient to overcome decision boundaries.

Graph and Geometric Structure. Recent research has expanded the attack surface beyond alphanumeric sequences to encompass the structural properties of the detection logic itself. Liu et al. (2021a) introduced a geometric vector perturbation method, which provides mathematical proofs for adversary generation in feature space rather than string space. In the field of Graph Neural Networks (GNN), Gao et al. (2023) explored graph adjacency attacks, while Nazzal et al. (2024) proposed `MintA`, a multi-instance attack that coordinates perturbations between connected nodes. `MintA` demonstrated that manipulating relational edges in a DNS graph (rather than the domain content itself) could achieve an evasion rate of 80.00%, revealing critical vulnerabilities in modern graph-based defenses.

5.2. Mechanisms of Model Detection Evasion

Our analysis identifies three fundamental structural weaknesses that are systematically exploited in detection architectures.

Dependence on Surface Features (68.75%). The most frequent vulnerability stems from detectors over-reliance on surface statistical metrics, such as entropy, character frequency, and n-gram distributions. Adversaries exploit this by creating domains that adhere to the syntactic rules of benign domain names while maintaining malicious intent. Adversaries achieve this distribution mimicry through character substitution (Liu et al., 2021b; Hu et al., 2023), automated distribution matching (Yun et al., 2020; Zhai et al., 2022), and the generation of

semantic embeddings (Charan et al., 2020; Selvaraj and Panjanathan, 2024). Indeed, these attacks demonstrate that models that focus on the appearance of a domain (syntax) rather than its essence (semantics) are inherently fragile. Furthermore, recent research demonstrates that this fragility extends to long-range dependencies; attacks utilizing Self-Attention mechanisms (Luo et al., 2026) and Transformer-based autoencoders (Pregardier et al., 2025) successfully evade detection by mimicking global semantic structures that n-gram-based detectors fail to capture, effectively bypassing the entire class of frequency-based analysis.

Static Training Distributions (43.75%). A generalization gap exists due to the static nature of training datasets. Detectors trained exclusively on historical families of DGAs frequently fail to identify new adversarial variants (Anderson et al., 2016; Ren et al., 2023). This vulnerability highlights a fundamental tension between the rapid evolution of adversarial generative logic and the slower model implementation cycle. As noted in recent literature, models lacking continuous learning mechanisms suffer severe performance degradation when exposed to unseen attack families that deviate from the training distribution. This gap is further widened by novel adaptive strategies such as adversarial self-augmentation (Pregardier et al., 2025), where the generative model iteratively retrains itself on its own successful evasions, effectively automating the exploitation of static decision boundaries before the defender can even initiate an update cycle.

Differentiable Decision Boundaries (28.13%). The differentiability of deep neural networks, while necessary for training, creates a direct attack vector for white-box optimization. By accessing the model's gradient information, attackers can compute precise, non-random perturbations that maximize the probability of misclassification (Sidi et al., 2020; Yilmaz et al., 2020). Unlike statistical evasion, which guesses blind spots, gradient-based exploitation mathematically computes the optimal path through the decision boundary, allowing valid domains to be pushed into the benign classification region with minimal modification.

5.3. Countermeasures and Defensive Robustness

A *cat-and-mouse* dynamic characterizes the defensive landscape, where detectors must continually adapt to the constantly evolving generational logic. We classify the analyzed defensive mechanisms into three main categories: *adversarial training*, *contextual enrichment*, and *adaptive learning architectures*.

Adversarial Training and Augmentation (65.63%). This approach seeks to immunize detectors by explicitly incorporating adversarial samples into the training set (Anderson et al., 2016; Yilmaz et al., 2020; Liu et al., 2021b). The effectiveness of this strategy varies considerably. Alaeiyan et al. (2020) achieved near-perfect defense, improving accuracy from 0.47% to 98.15% after retraining with GADGA samples. In contrast, Peck et al. (2019) found that simple retraining was insufficient against heuristic attacks such as `CharBot`, where 79.80% evasion persisted. A main limitation is the lack of generalization across families. Zheng et al. (2021) showed that while retraining improved the detection of known adversary families (63.00% $\rightarrow$ 92.00%), it offered negligible protection against undetected variants (stagnating at $\approx$ 56.00%). This suggests that standard adversarial training leads to overfitting to specific attack signatures rather than learning robust, invariant features. This limitation has been further confirmed by Pregardier et al. (2025), who demonstrated that even rigorous adversarial retraining fails to contain self-augmenting generators, which iteratively learn from the detector's feedback to bypass the strengthened boundaries. This indicates that simple data augmentation is insufficient for long-term robustness against the semantic shifts of modern adaptive generators.

Contextual and Secondary Information Defense (25%). Recognizing the limitations of purely lexical analysis, researchers have integrated auxiliary data sources such as DNS metadata, WHOIS records, and network traffic characteristics (Sivaguru et al., 2020; Ravi et al., 2023). By expanding the feature space beyond the domain string, these systems aim to identify malicious behavior rather than just malicious syntax. However, this is not a panacea; Sivaguru et al. (2020) observed that even with enriched secondary information, 79.94% of domains generated by `CharBot` remained undetected, indicating that sophisticated mimics can evade multimodal detection if the attacker has a good understanding of the underlying feature dependencies.

Adaptive and Incremental Architectures (15.63%). Recent work focuses on dynamic systems that evolve alongside the adversaries. For instance, Ren et al. (2023) proposed frameworks that adapt to new DGA families without neglecting older ones, maintaining an F1 score of 0.75 where static models failed. Similarly, Nie et al. (2024) introduced game-based incremental learning, achieving an AUC of 0.98 using iterative detector-adversary cycles. While effective, this approach involves a high operational cost, as it requires continuous retraining cycles that can hinder scalability in real-time environments. Finally, Zheng et al. (2021) used an autoencoder-based approach to purify input samples, achieving an average improvement of 20.00% in accuracy without fully retraining the model.

Trade-offs between Performance and Robustness. Defensive strengthening often involves computational costs, although efficient alternatives exist. Charan et al. (2020) found that C5.0 decision trees offered a superior trade-off compared to Random Forests, maintaining 95.03% accuracy on adversarial samples with a degradation of only 3.05%, while halving training time. Similarly, Zhang et al. (2022) showed that WGAN-based data augmentation could increase robustness (Kappa 0.89–0.90) without sacrificing performance on clean data. Ultimately, the literature reveals a clear dichotomy: defenses are either highly effective but limited (stopping only known attacks (Sidi et al., 2020; Zheng et al., 2021)), or broad but computationally expensive (requiring continuous adaptation).

5.3.1. Operational Feasibility Framework

To bridge the gap between academic performance and industrial applicability, we apply the operational feasibility framework defined in Section 4.6. We evaluate the primary defense strategies against four critical thresholds that determine viability in production environments. Specifically: (i) *inference latency constraint (< 100 ms)*: While attackers achieve generation rates exceeding 6,000 domains per second (Spooren et al., 2019), defenses must operate within a strict margin (from sub-milliseconds to 100 ms per query) to avoid network bottlenecks (Nie et al., 2023); (ii) *alert fatigue threshold (FPR < 0.1%)*: High accuracy is not enough if the volume of false positives overwhelms security analysts. Operational viability requires a FPR strictly below 0.1% (Peck et al., 2019). Systems that exceed this threshold induce alert fatigue, rendering them virtually unusable, regardless of their TPR; (iii) *resource asymmetry*: The resource landscape significantly favors the attacker, who can operate with lightweight generators (160 KB – 1.86 MB) (Sidi et al., 2020; Yun et al., 2020). Defenses should be evaluated based on their scalability without requiring exponential memory growth or prohibitive hardware acceleration; (iv) *retraining frequency*: Given the rapid drift of DGA families, static models degrade quickly. Feasibility depends on the frequency of required updates (e.g., daily or monthly) and the risk of catastrophic forgetting during these cycles. Table 4 presents a structured evaluation of the main defense strategies according to these criteria.

5.4. Experimental Methodology and Validity

The validity of the adversarial results depends entirely on the experimental rigor applied. Our analysis reveals a fragmented methodological landscape, characterized by inconsistency in the selection of datasets and metrics, and a frequent disregard for operational constraints. This heterogeneity generates a reproducibility crisis, which hinders and often leads to misleading direct comparisons of performance between studies. In the sequel, we discuss more in depth our analysis results.

5.4.1. Data Heterogeneity and Comparative Analysis

The quality of underlying data fundamentally limits the reliability of evasion rates. As illustrated in Figure 4, the field lacks a standardized benchmark and instead relies on ad hoc combinations of public lists.

Baseline Shift. For benign traffic, the Alexa Top 1 Million list (Alexa) served as the standard in 65.63% of the reviewed papers. However, Figure 4 highlights a transition: Alexa's dominance has declined significantly since 2019, being replaced by the Tranco list (Tranco) in recent studies. This shift is methodologically sound (Tranco offers greater stability and resistance to manipulation), but it complicates longitudinal comparisons, as models trained on Alexa's older, noisier distribution cannot be fairly compared with those trained on Tranco.

Contamination and Privacy. Data quality remains a persistent problem. Alexa lists have been documented to contain malicious samples (Spooren et al., 2019; Peck et al., 2019), and even

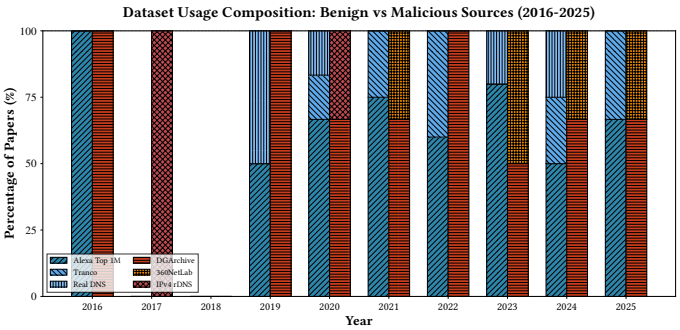

Figure 4: Relative composition of benign and malicious dataset use in adversarial DGA research. Each bar represents 100% of papers using that category per year.

the more reliable Tranco list is not entirely free of contamination (Pochat et al., 2019). While some researchers advocate for high-fidelity sources such as real DNS traffic (Peck et al., 2019; Sivaguru et al., 2020) or IPv4 reverse DNS records (Fu et al., 2017; Yun et al., 2020), privacy restrictions limit the public availability of these datasets, forcing the community to resort to synthetic or potentially contaminated public lists.

Fragmentation. Malicious sources also exhibit similar fragmentation. While DGArchive (Plohmann et al., 2016) remains the dominant source (43.75% of studies), recent work is increasingly fragmented among 360NetLab (netlab-360) , UMUDGA (Tuan et al., 2022), and Bambenek feeds (Bambenek). As a result, state-of-the-art claims are often based on completely unrelated datasets with varying levels of difficulty, severely limiting generalizability.

5.4.2. The Evaluation Gap: Metrics vs. Reality

There is a systematic disconnect between academic optimization goals and operational success criteria. The inconsistent adoption of operationally relevant metrics, compounded by the heterogeneity of datasets, fundamentally undermines the ability to make fair comparisons between adversarial technique studies. This variation reflects a broader divide between academic assessment protocols and real-world implementation realities. In operational environments, detection systems are governed by rigid Service Level Agreements that require the processing of millions of daily queries with sub-millisecond latency. These systems must maintain FPRs below 0.10% to avoid alert fatigue–a stringent constraint that is rarely enforced or reported in peer-reviewed academic literature. Consequently, while current research successfully demonstrates theoretical vulnerability, it often fails to validate the practical viability of threats under the stringent constraints of production networks.

The Accuracy Trap. Although Precision, Recall, and F1-score provide more informative assessments and appear in 46.88%, 46.88%, and 43.75% of studies, respectively, Accuracy remains the prioritized metric in 53.13% of the literature. However, its inadequacy for unbalanced datasets is well-documented (He and Garcia, 2009). In real-world scenarios where benign traffic outnumbers malicious queries by orders of magnitude, a high

Table 4: Operational Feasibility Assessment. Comparison of defense strategies against strict production constraints. Values indicate the risk or operational cost associated with each dimension.

Defense Strategy	Latency	FPR Risk	Resources	Retrain. Freq.	Primary Bottleneck
Adversarial Training	Low	High	Med	High	Poor generalization (Zheng et al., 2021)
Contextual Enrichment	High	Low	High	Low	External query overhead (Sivaguru et al., 2020)
Adaptive Architectures	Med	Med	High	Low	System complexity (Ren et al., 2023)

accuracy score frequently masks a functionally useless detector that completely fails to identify the minority adversarial class.

Standard metrics, such as Accuracy and AUC, are inherently misleading in this field due to extreme class imbalance. A detector achieving 99% accuracy remains operationally unviable if a 1% error rate generates thousands of false alerts daily, leading to alert fatigue. Therefore, evaluation should focus on cost-weighted risk metrics, prioritizing operational FPR (false positives per unit of time) over overall performance scores. Rigorous comparison requires setting FPR at strict production thresholds (e.g., $< 0.1\%$) and measuring the resulting TPR, rather than integrating performance in operationally irrelevant regions of the ROC curve.

Attack-Specific Metrics and Transferability. To assess generative quality, researchers use specialized metrics that go beyond standard classification metrics. While the evasion rate remains the dominant metric (90.63%), recent literature increasingly validates the usefulness of generated samples using specific structural metrics such as collision rate (25.00%), which measures the probability of generating a domain that is already registered (and therefore unusable), and repetition rate (15.63%), which quantifies the pseudo-random generator's internal redundancy, indicating how frequently it produces duplicate domains within the same campaign. Recently, evaluation criteria have expanded towards semantic realism; Luo et al. (2026) introduced metrics for word-level element distribution to quantify how closely generated domains mimic natural language meaningfulness, moving beyond simple character-level statistics. Additionally, few studies assess transferability (25.00%), measuring the success of attacks on different detector architectures (e.g., GAN-generated domains tested against heterogeneous black-box models).

5.4.3. Disregard of Operational Constraints

Perhaps the most significant gap in the literature is widespread neglect of the operational framework established in Section 5.3.1. Only 40.63% of the reviewed works explicitly address or report metrics related to the deployment environment, often ignoring hardware limitations and latency constraints inherent to botnet infrastructure.

Performance and Latency. Despite the critical inference latency constraint identified previously, few studies document their inference times. This omission is critical given the structural asymmetry of this field: while attackers can generate thousands of domains per second (Spooren et al., 2019), defenders often fail to report whether their advanced deep learning models can process wire-speed traffic without causing bottlenecks. This oversight could introduce a vulnerability to volume-based denial-of-service attacks, induced by the defense mechanism itself.

Malware Footprint. Similarly, the attacker's limitations are rarely considered in defensive modeling. While we identified that viable adversarial generators are lightweight (e.g., 160 KB – 1.86 MB) (Sidi et al., 2020; Yun et al., 2020), many proposed defenses operate under the assumption of unlimited computational resources. This overlooks a critical deployment asymmetry: while sophisticated generators are optimized for portability within malware binaries, many proposed defenses rely on high-performance hardware that is not typically available in standard network edge devices.

Irrelevance of Training Cost. Finally, a common error in reviewed works is the emphasis on training time as a feasibility metric. While training requires considerable resources (e.g., 14 hours for older architectures (Anderson et al., 2016)), this is a one-time offline development cost. Once trained, the generator is efficient. Therefore, criticisms of adversary attacks based on "high training cost" lack operational validity; the primary feasibility metric should be inference efficiency, which continues to be systematically underestimated.

5.5. Reproducibility Assessment

Reproducibility is the cornerstone of scientific validity; however, it remains a systemic challenge in adversarial DGA research. To rigorously assess the state of the field, we evaluated all 32 primary studies using the scoring rubric defined in Section 4.7. The detailed results of this assessment are presented in Table 5.

Our analysis reveals a critical transparency gap. Only 12.50% of the reviewed studies (4 of 32) achieve a *high* level of reproducibility (Sidi et al. (2020); Nie et al. (2023); Drichel et al. (2024a), and Nazzal et al. (2024)). These works are distinguished by providing publicly accessible source code along with detailed experimental documentation. A moderate segment of the literature (25.00%) falls into the *medium* category. While these studies do not publish code, they provide sufficient granularity regarding architecture, preprocessing steps, and hyperparameters (such as Fu et al. (2017); Selvaraj and Panjanathan (2024), and Pregardier et al. (2025)) to allow for an approximate independent reimplementation. However, the majority of the works (62.50%) are classified as having *low* reproducibility. These studies function as "black boxes," often omitting hyperparameters or architectural details, thus preventing scientific replication.

Table 5: Detailed evaluation of the reproducibility of the 32 reviewed studies. Columns indicate the availability of source code (Code), preprocessing steps (Prep.), architecture details (Arch.), hyperparameters (Params.), and fixed random seeds (Seed). Our score classifies studies as High (public artifacts), Medium (sufficient documentation), or Low (insufficient details). Notably, there is a widespread lack of random seed reports.

Reference	Code	Prep.	Arch.	Params.	Seed	Score
Anderson et al. (2016)	✗	✓	✓	✓	✗	Medium
Fu et al. (2017)	✗	✓	✓	✓	✗	Medium
Peck et al. (2019)	✗	✓	✓	✗	✗	Low
Spooren et al. (2019)	✗	✓	✓	✓	✗	Medium
Sivaguru et al. (2020)	✗	✓	✗	✗	✗	Low
Sidi et al. (2020)	✓	✓	✓	✓	✗	High
Yilmaz et al. (2020)	✗	✓	✗	✗	✗	Low
Alaeiyan et al. (2020)	✗	✓	✓	✓	✗	Medium
Charan et al. (2020)	✗	✓	✓	✗	✗	Low
Gould et al. (2020)	✗	✓	✓	✗	✗	Low
Yun et al. (2020)	✗	✓	✓	✗	✗	Low
Cao et al. (2020)	✗	✓	✓	✗	✗	Low
Zheng et al. (2021)	✗	✗	✓	✓	✗	Low
Wang and Guo (2021)	✗	✓	✓	✓	✗	Medium
Liu et al. (2021a)	✗	✓	✗	✗	✗	Low
Liu et al. (2021b)	✗	✓	✓	✗	✗	Low
Zhai et al. (2022)	✗	✓	✗	✗	✗	Low
Shu et al. (2022)	✗	✓	✓	✗	✗	Low
Zhang et al. (2022)	✗	✓	✓	✗	✗	Low
Suryotrisongko et al. (2022)	✗	✓	✓	✗	✗	Low
Hu et al. (2023)	✗	✓	✓	✗	✗	Low
Ravi et al. (2023)	✗	✓	✓	✗	✗	Low
Ren et al. (2023)	✗	✓	✓	✗	✗	Low
Gao et al. (2023)	✗	✓	✗	✗	✗	Low
Nie et al. (2023)	✓	✓	✓	✓	✗	High
Drichel et al. (2024a)	✓	✓	✓	✓	✗	High
Nie et al. (2024)	✗	✓	✓	✓	✗	Medium
Nazzal et al. (2024)	✓	✓	✓	✓	✗	High
Selvaraj and Panjanathan (2024)	✗	✓	✓	✓	✗	Medium
Nangong and Wu (2025)	✗	✓	✗	✗	✗	Low
Pregardier et al. (2025)	✗	✓	✓	✓	✗	Medium
Luo et al. (2026)	✗	✓	✗	✗	✗	Low

A particularly troubling finding is the universal absence of deterministic controls. None of the analyzed studies reported the random seeds used for initialization or data splitting. Given this widespread deficiency in the industry, we assigned the label *High* to studies that meet all other criteria (Code, Preprocessing, Architecture, and Hyperparameters), considering them the current upper limit of reproducibility in the field despite the absence of seeds. This lack of information makes the exact replication of stochastic training procedures impossible.

This reproducibility deficiency stems from a confluence of ethical and systemic factors, ranging from dual-use security concerns and competitive academic pressures to genuine technical obstacles such as the prohibitive size of datasets (Selvaraj and Panjanathan, 2024). The impact of this opacity extends beyond individual studies, as the lack of standardized benchmarks deprives the community of a consensus on cutting-edge performance. Without the ability to directly test new countermeasures against proven adversary samples, defensive research is forced into a cycle of guesswork, where improvements cannot be reliably distinguished from experimental noise or specific setup biases, ultimately transforming reproducibility from an artifact-sharing problem into a fundamental barrier to cumulative scientific progress.

6. Discussion

In this section, we synthesize the findings of our systematic analysis to directly address the research questions formulated in Section 4.1. We structure the discussion around three thematic pillars. First, we interpret the technological trajectory of the adversary's arms race (RQ1–RQ3), analyzing how offensive innovation has outpaced defensive adaptation. Second, we critically examine the methodological foundations of the field (RQ4–RQ5), exposing systemic flaws in assessment standards and reproducibility that undermine scientific validity. Finally, we contrast academic defensive strategies with the operational realities of deployment (RQ6–RQ7), highlighting the profound deployment gap and structural asymmetries that currently favor the adversary.

6.1. RQ1: Evolution of Adversarial Evasion Techniques

A temporal analysis of 32 primary studies reveals that research on adversarial DGA has not only grown in volume but has also undergone a distinct methodological metamorphosis. As illustrated in Figure 2, this evolution follows three identifiable phases, from established initial proofs of concept to an intense diversification and a shift toward semantic stealth and defensive robustness.

This trajectory reflects a fundamental change in research objectives: initial efforts (Anderson et al., 2016; Fu et al., 2017) focused on feasibility, demonstrating that deep learning could automate domain generation. The peak in 2020 marked a shift toward effectiveness, introducing various paradigms such as Wasserstein stability (Yun et al., 2020) and gradient-based optimization(Sidi et al., 2020; Yilmaz et al., 2020) to address the modal collapse and instability issues of earlier models. The current stabilization phase indicates a transition toward reliability, where the focus has moved beyond simple evasion to ensure domain registerability, semantic validity, and resilience to countermeasures (Drichel et al., 2024a; Nie et al., 2024).

Technologically, this evolution is defined by four converging innovations that have progressively narrowed the gap between benign and malicious traffic profiles. First, gradient-based optimization (18.75% of studies) shifted the paradigm from guessing to calculating evasion, enabling precise, minimal perturbations that are computationally cheaper than training full generative models. Second, controllable generation (Zhai et al., 2022; Ren et al., 2023) introduced constraints that force adversary domains to adhere to valid syntax and logging rules, directly addressing the practical limitations of previous noisy GANs. Third, the focus on transferability revealed that blind attacks could succeed across diverse architectures (Spooren et al., 2019; Sidi et al., 2020; Liu et al., 2021b), with 46.88% of studies validating black-box scenarios, demonstrating that attackers do not require white-box access to threaten deployed systems. Finally, semantic integration (Shu et al., 2022; Selvaraj and Panjanathan, 2024) represents the current frontier, moving away from character-level randomness toward word-based imitation that creates natural-looking domains indistinguishable to human observers.

Contemporary research (2023–2025) exemplifies this accumulated sophistication. Cutting-edge techniques now combine these innovations in hybrid systems: `ReplaceDGA` (Hu et al., 2023) merges semantic modeling with character perturbation; Multi-instance attacks (Nazzal et al., 2024) extends evasion to the structural level of graph-based detectors; while novel transformer GAN-based architectures (Pregardier et al., 2025; Luo et al., 2026) leverage Self-Attention mechanisms to master global dependencies. These advances suggest that the field has not peaked, but is entering an era of high precision, where adversarial techniques are evolving to evade not only statistical detectors, but also semantic and structural defenses.

Takeaways from RQ1

Adversarial DGA research has evolved through three distinct phases (emergence, expansion, and maturation), progressing from unstable character-level generative models to high-precision, semantically integrated frameworks. Contemporary techniques have shifted their focus from simple evasion to operational feasibility, leveraging gradient-based optimization, controllable generation, and game-theory adaptation. These modern methods prioritize domain registerability and black-box transferability, effectively overcoming static defensive measures by mimicking the semantic and structural patterns of benign domain names.

6.2. RQ2: Vulnerability Landscape and Defensive Efficacy

Our analysis reveals a distinct hierarchical structure of vulnerabilities that adversaries exploit with varying degrees of success. As summarized in Table 6, these vulnerabilities stem primarily from the over-reliance of vulnerability detectors on shallow statistics and static training distributions, rather than on the depth of the architecture.

Syntactic overdependence is the most widespread weakness (appearing in 68.75% of the studies). This vulnerability arises because most deep learning detectors optimize form rather than intent, relying on proxy metrics such as entropy, character frequency, and n-gram transitions. Adversaries exploit this syntactic bias through distribution mimicry: Yun et al. (2020) showed that by matching the syllabic structure of benign domains, a WGAN could degrade the AUC of the LSTM detector to 0.57 (a near-random assumption). Similarly, word-based attacks (Selvaraj and Panjanathan, 2024) bypass these filters entirely by constructing semantically valid domains that are statistically indistinguishable from legitimate traffic.

The generalization gap represents the second critical failure mode. There is an intrinsic tension between the static nature of training datasets and the dynamic evolution of adversary logic. Detectors trained primarily on historical artifacts suffer severe performance degradation when faced with new artifact families. For example, Ren et al. (2023) quantified a drop in the F1 score to 0.67 for static models, highlighting that without continuous learning, detectors quickly become obsolete.

Finally, gradient fragility exploits the architectural nature of neural networks. The differentiability required for training becomes a disadvantage during deployment, allowing white-box attackers to mathematically calculate the optimal perturbation needed to cross the decision boundary, reducing accuracy from 98.00% to 76.00% with minimal noise injection.

Regarding the defensive landscape, it has evolved in response but remains reactive. We categorize current strategies into three levels of effectiveness:

Tier 1: Adversarial Training (The Brute-Force Approach). Employed in 65.63% of studies, this method attempts to plug gaps in the decision boundary by including adversarial samples in training. While it can achieve near-perfect recovery against known attacks (e.g., 98.15% accuracy (Alaeiyan et al., 2020)), it suffers from poor generalization across families. As Zheng et al. (2021) noted, robustness against one generator rarely transfers to another, leading to a perpetual game of "hit the mole." This asymmetry is further exacerbated by novel self-augmenting architectures (Pregardier et al., 2025), which automate this learning loop on the offensive side, rendering static defensive snapshots obsolete almost immediately upon deployment.

Tier 2: Contextual Enrichment. Strategies that integrate secondary information (DNS metadata, traffic volume) consistently outperform purely lexical detectors. By expanding the feature space beyond the domain string, these systems force adversaries to mimic not only the name but also the behavior of benign actors, significantly increasing the cost of attack.

Tier 3: Adaptive Architectures. The frontier of defense lies in dynamic systems. Continuous learning frameworks (Ren et al., 2023) and game-theory models (Nie et al., 2024) abandon the notion of a static detector in favor of systems that evolve alongside the adversary. While Nie et al. (2024) demonstrated that such adaptation can maintain a high AUC, these approaches introduce significant operational complexity and require continuous retraining cycles that can hinder scalability in high-performance environments.

Takeaways from RQ2

The main weakness of current detectors is an over-reliance on superficial statistics that generative models can easily mimic. This is exacerbated by static training sets that do not generalize to new adversary families. Regarding countermeasures, standard adversary training offers only limited and fragile robustness. A sustainable defense requires moving beyond data augmentation to contextual enrichment (e.g., domain metadata) and adaptive architectures (e.g., continuous learning) that can dynamically adjust to changing adversary conditions.

6.3. RQ3: The Dual Role of Deep Learning Architectures

Deep learning architectures have transcended their initial role as simple classifiers to become the fundamental engine driving the adversarial arms race. Our analysis identifies a distinct bifurcation in architectural roles: on the offensive side, architectures function as distribution approximators seeking to mimic

Table 6: Ranking of Exploited Vulnerabilities in DGA Detectors

Vulnerability Rank	#	Primary Exploitation Vectors
Character-level Dependency	23	N-gram mimicry, Homograph attacks, Word-based substitution
Training Data Limitations	15	Novel family generation, Catastrophic forgetting
Gradient-based Weakness	9	Perturbation optimization, Boundary identification

benign statistics; on the defensive side, they act as feature extractors attempting to discern subtle structural deviations.

GANs constitute the dominant architectural paradigm, appearing in 40.63% of approaches. Their primacy stems from their ability to map random noise to realistic domain distributions. The field has largely abandoned *vanilla* GANs in favor of WGAN-GP. This architectural shift was necessitated by the discrete nature of text data, where vanilla GANs frequently suffered from mode collapse. On the contrary, WGAN-GP provides the necessary gradient stability to generate diverse, high-quality character sequences. Recent innovations have moved beyond random generation to conditional GANs, which introduce "control knobs", allowing adversaries to dictate specific evasion properties (e.g., "*generate a domain that bypasses Filter X*") while strictly maintaining validity constraints, effectively turning random generation into targeted weaponization. Most recently, the architectural focus has shifted toward mastering global dependencies through Self-Attention mechanisms (Luo et al., 2026; Pregardier et al., 2025). By integrating Transformers into the generative framework, these models transcend the limitations of local n-gram approximation, enabling the synthesis of domains with long-range semantic coherence that closely mimics human language structure. Conversely, VAEs offer a more stable, probabilistic alternative. However, a clear trade-off exists: while VAEs minimize training instability, they typically yield lower evasion rates than the aggressive optimization of GANs, resulting in blurrier samples that fail to penetrate high-confidence detectors.

However, our synthesis reveals a critical divergence: While GAN-based architectures dominate the academic landscape (representing over 40% of studies), simple heuristic perturbations often rival them in effectiveness, while offering superior operational feasibility. Techniques such as `CharBot` (Peck et al., 2019) or stochastic character injection (Sivaguru et al., 2020) achieve comparable evasion rates at negligible computational cost, avoiding the instability inherent in adversarial training. This discrepancy highlights potential academic bias toward architectural complexity, obscuring the practical reality that an attacker's optimal strategy (i.e, maximizing utility and minimizing cost) typically favors deterministic rules over deep generative models.

Defensive architectures have consolidated around capturing dual-modality features: local character patterns and long-range sequential dependencies. The combination of CNN and LSTM networks has emerged as the industry standard. In this symbiotic arrangement, CNN layers function as feature extractors for local n-grams (mimicking lexical analysis), while LSTM layers capture the temporal dependencies of the sequence. Empirically, these LSTM-based approaches consistently outperform traditional machine learning, with average accuracies of 98.70% compared to 93.70% for Random Forests (Spooren et al., 2019). A significant architectural pivot is the adoption of GNNs, which redefine detection from content analysis to context analysis. By modeling domain-IP and client relationships, GNNs detect malicious patterns that are invisible to sequence-based detectors.

Despite the increasing complexity of defensive models, our review highlights a critical paradox: architectural depth does not equate to adversarial robustness. Hybrid systems, for all their accuracy, remain vulnerable to the same fundamental weaknesses as simpler models. Gradient-based perturbations exploit the differentiability of the loss function regardless of whether the architecture is a CNN, LSTM, or Transformer (Yilmaz et al., 2020). Furthermore, topological defenses have introduced new attack surfaces, where surrogate models can successfully transfer structural perturbations to complex heterogeneous graphs (Nazzal et al., 2024). This suggests that current architectural innovations are optimizing for performance on static distributions, rather than solving the fundamental generalization problem against dynamic adversaries.

Takeways from RQ3

In their offensive role, DL serves as a distribution approximator. The field has standardized on WGAN-GP to solve stability issues in discrete data generation, with a shift toward Conditional GANs for targeted, constraint-aware evasion.
Regarding their defensive role, the model architecture serves as a feature extractor. While Hybrid CNN-LSTM models capture local and sequential dependencies effectively, and GNNs capture structural context, increased complexity has *not* yielded intrinsic robustness. Differentiability remains a universal vulnerability, allowing gradient-based attacks to succeed regardless of architectural depth.

6.4. RQ4: Evaluation Methodologies and Standardization

The synthesis of experimental settings (see Section 5.4) reveals that the evaluation of adversarial techniques is currently compromised by three systemic flaws: heterogeneity of datasets, metric inconsistency, and a lack of standardized reference frames.

The "Apples-to-Oranges" Problem (Datasets). The heterogeneity of datasets constitutes the main impediment to scientific validity. While researchers rely on common resources such

as DGArchive (Plohmann et al., 2016) and 360NetLab (netlab-360), the composition of training and test sets varies considerably across studies. Similarly, benign baselines have fractured between traditional studies using the Alexa Top 1 Million (Alexa) and modern work adopting the Tranco list for stability (Tranco). Since model performance is inextricably linked to the difficulty of the specific segment of the dataset used, claims that "Technique X (98% evasion)" outperforms "Technique Y (95% evasion)" lack scientific validity unless tested on identical data partitions. Currently, the results are only valid in their isolated experimental contexts, preventing a true cutting-edge evaluation. Metric selection also reveals a fundamental gap between academic optimization and operational reality. Accuracy remains the dominant metric (53.13% of studies) despite its mathematical inadequacy for highly unbalanced datasets (He and Garcia, 2009). Similarly, while the evasion rate is standard for measuring attack success (90.63%), critical operational metrics such as TPR with low FPR thresholds (e.g., 0.1%) are frequently omitted. This suggests that many "successful" attacks could, in fact, generate prohibitively high false alarm rates in production, making them practically viable only in paper evaluations.

The Framework Gap. Unlike the AGD detection research field, which has begun to establish reference implementations (Pelayo-Benedet et al., 2025), adversarial generation lacks a unified framework. There is no standard protocol enforcing mandatory transferability tests or latency constraints. This absence forces each new study to reimplement baselines from scratch, often introducing implementation biases, and prevents the community from maintaining a reliable ranking of adversarial techniques.

Takeways from RQ4

The evaluation of adversarial dynamic data analysis techniques is fundamentally compromised by three systemic flaws: the heterogeneity of datasets, the inconsistency of metrics, and the absence of standardized benchmarks. Scientific validity is undermined by the fragmented use of benign baselines, which invalidates performance comparisons between studies. This comparability problem is compounded by the reliance on misleading aggregate accuracy metrics that ignore strict false-positive constraints, which are necessary to the operation. Furthermore, the lack of a unified evaluation framework forces the community to resort to unverified reimplementations, hindering the reliable identification of truly cutting-edge techniques.

6.5. RQ5: Reproducibility and Open Science Practices

As quantified in Section 5.5, the field currently faces a systemic transparency crisis that threatens its scientific validity. Only 12.50% of studies provide public code. This pervasive opacity creates a "black box" research environment where the absence of shared model weights and contrasting datasets makes it impossible to distinguish genuine algorithmic advances from experimental noise or configuration artifacts. Even when implementation details are partially reported, the widespread omission of deterministic random seeds (often replaced by non-reproducible, time-dependent initializations) makes exact replication of stochastic training procedures impossible, reducing peer review to a mere check of theoretical plausibility rather than empirical verification.

To restore credibility, the community must move from partial reporting to a comprehensive open science. Research groups should prioritize publishing public code through persistent repositories (e.g., Zenodo, GitHub) and, where security concerns exist regarding powerful generators, adopt tiered access systems instead of full retention. Furthermore, documentation standards must evolve beyond simple architecture diagrams to include experimentation details: precise training hyperparameters, convergence criteria, and hardware-specific constraints. Ultimately, progress depends on redefining reproducibility not as an optional add-on, but as a non-negotiable requirement for legitimate scientific contribution in adversarial DGA research.

The proliferation of high-performance adversarial generators, though, also raises significant ethical concerns regarding dual-use technology. While our review advocates for open science to address the reproducibility crisis, we acknowledge the tension between scientific transparency and operational security. Releasing code for potent generators (e.g., those achieving $> 90\%$ evasion) lowers the barrier to entry for threat actors.

To mitigate this risk without hindering research, we recommend adopting a *responsible benchmarking* framework. This implies: (i) distinguishing between *verification artifacts* (datasets, pre-calculated adversarial samples), which should remain open, and *weaponizable code* (training scripts for active generators), which may require restricted access; and (ii) refocusing the evaluation toward *defensive hardening* rather than purely offensive metrics. Ultimately, the scientific community must ensure that the democratization of adversarial capabilities serves exclusively to stress-test and strengthen defensive infrastructure. By adhering to these standards of responsible disclosure, researchers can validate detection robustness without inadvertently providing a model for next-generation botnet operators.

Takeways from RQ5

Current practices are scientifically inadequate and characterized by a severe lack of public artifacts (only 12.50% code availability) and undocumented stochastic controls (0% random seed reports). Restoring validity requires a mandatory shift towards Open Science standards (specifically, publishing source code and trained weights) to ensure that reported performance gains are attributable to algorithmic superiority rather than hidden configuration variations.
However, greater transparency should not compromise operational security. We recommend adopting a responsible benchmarking framework that separates verification artifacts (openly shared datasets used to demonstrate validity) from weaponizable code (restricted training scripts). This ensures scientific robustness without lowering the barrier to entry for malicious actors.

6.6. RQ6: Defense Strategies and Operational Feasibility

Based on our systematic analysis of defensive mechanisms (see Section 5.3), we have categorized the literature into four distinct strategic approaches. As summarized in Table 7, while adversary training remains the dominant paradigm (present in approximately 66.67% of defensive studies), our review identifies a fundamental discrepancy between laboratory effectiveness and operational feasibility.

The effectiveness of these strategies is highly context-dependent and often overstated in isolation. Adversarial training, despite being the industry standard, exhibits significant fragility when faced with new families; studies report persistent avoidance rates exceeding 77.74%, even after retraining (Peck et al., 2019). This indicates that simply increasing the amount of data does not provide long-term robustness against the semantic changes of modern generators. In contrast, emerging adaptive mechanisms, such as continuous learning (Ren et al., 2023) and game-theory approaches (Nie et al., 2024), show significant promise in mitigating catastrophic forgetting (maintaining F1 scores of 0.75 against evolving threats), but they introduce architectural complexities and stability risks that have not yet been validated in large-scale production environments.

Fundamentally, there is a profound disconnect between academic evaluation metrics and real-world implementation requirements. Only 40.63% of reviewed articles explicitly acknowledge operational limitations, and quantitative reports of feasibility metrics are extremely scarce. While detection accuracy is universally reported, the metrics of interest for production systems (latency, throughput, and strict false-positive limits) are systematically ignored. For example, accurate inference times are rarely documented, with outliers such as Nie et al. (2024) reporting 1.91 ms/query, despite stringent industry requirements for sub-100 ms processing times. This defensive latency stands in stark contrast to the scalability of offensive tools, which can generate more than 6,000 domains per second (Spooren et al., 2019). The most concerning aspect is that operational environments demand FPRs below 0.10% to avoid service disruption (Peck et al., 2019). However, our analysis reveals that applying this threshold often cripples detection capabilities. Sivaguru et al. (2020) noted that 80.00% of adversary domains successfully evaded detection when the system was restricted to an FPR of 0.10%, highlighting that many state-of-the-art defenses are only viable if the cost of false alarms is ignored.

Takeaways from RQ6

The defensive landscape remains dominated by adversary training, which provides only fragile and specific robustness against known attacks. While emerging adaptive methods offer improved generalization, they introduce prohibitive complexity. Critically, the application of the operational feasibility framework (see Table 4) reveals a significant deployment gap. When evaluated under strict production constraints (specifically, an inference latency of less than 100 ms and an FPR strictly below 0.10%) detection performance frequently degrades. This demonstrates a critical misalignment: Current defenses are typically optimized for static classification metrics rather than the dynamic requirements of real-time traffic filtering.

6.7. RQ7: Systemic Challenges and Computational Asymmetry

Beyond the operational failures detailed above, our analysis identifies three systemic challenges that create a clear advantage for the adversary: structural resource asymmetry, the scalability of adaptation, and the inevitable trade-off between robustness and efficiency. Regarding resource asymmetry, a fundamental imbalance exists between the resource requirements for generation and detection. Adversarial generators have a lightweight architecture; models such as `CharBot` (Peck et al., 2019) and `MaskDGA` (Sidi et al., 2020) require only 160 KB to 684 KB, making it easy to integrate them into malware payloads or execute them on compromised IoT devices. In contrast, effective defense requires a disproportionate investment in resources. Robust detectors often rely on deep, memory-intensive architectures (e.g., heterogeneous graph networks) or require real-time access to external threat intelligence (WHOIS, DNS history) to function. This asymmetry means that while the adversary can scale horizontally at minimal cost, the defender faces exponential resource growth to maintain coverage.

Likewise, the scalability challenge is not limited to inference speed but extends to adaptation speed. While attackers can generate thousands of new domains per second, defenders face a retraining bottleneck. As DGA families evolve (conceptual drift), static models degrade rapidly. However, continuous retraining is computationally expensive and introduces the risk of catastrophic forgetting, where the model loses the ability to detect older families as it adapts to newer ones. Maintaining an up-to-date and comprehensive defense requires complex repetition mechanisms or incremental learning frameworks that are significantly more difficult to design than the attacks they counter.

Consequently, this arms race confirms that AGD detection is fundamentally a *non-stationary learning problem*, rather than

Table 7: Comparative analysis of defense strategies against adversarial DGAs.

Defense Strategy	#	Primary Limitation	Operational Trade-Off
Adversarial Training	21	Poor cross-family generalization	Robustness vs. clean accuracy
Side Information	4	High latency / Data access	Detection rate vs. throughput
Specialized	4	Narrow threat scope	Complexity vs. evasion rate
Adaptive / Continual	3	Retraining overhead	Adaptation vs. system stability

a static classification task. From a game-theoretic perspective, there is no permanent Nash equilibrium; any effective defense acts as selective pressure, forcing the adversary to mutate. Therefore, the strategic objective must shift from the search for a "perfect" classifier–an illusory goal of static robustness–to the development of co-evolutionary frameworks capable of continuous adaptation. We must accept that the lifespan of any fixed decision boundary is effectively limited by the adversary's rate of adaptation.

Finally, our review highlights a zero-sum trade-off, effectively a *robustness tax*: there is a consistent inverse relationship between a model's adversarial resilience and its performance in benign traffic. For instance, Ren et al. (2023) reported that implementing continuous learning strategies to counter evasion resulted in a significant drop in the F1 score for clean data (from 0.93 to 0.75). Therefore, proponents are forced to choose between optimizing efficiency and accuracy (risk of evasion) or optimizing robustness (risk of higher false positives and computational overhead). Currently, no solution exists that simultaneously maximizes both without incurring significant penalties.

Takeaways from RQ7

A structural imbalance favors the attacker, who uses lightweight and portable (KB-scale) generators, while defenders require heavyweight, memory-intensive architectures to detect them. Furthermore, the retraining bottleneck prevents defenses from keeping pace with attack generation; frequent updates risk catastrophically overlooking older threats. Finally, there is a robustness tax when increased resistance to adversarial attacks consistently degrades performance on benign traffic or strictly increases computational overhead.

7. Limitations

While this systematic review adheres to rigorous methodological protocols to minimize bias and ensure comprehensiveness, it is subject to inherent threats to validity (Ampatzoglou et al., 2019), common in secondary research. We discuss these limitations below.

Search Strategy and Linguistic Bias. Our inclusion criteria introduced specific restrictions regarding language and scope. By restricting the review to English-language publications, we may have inadvertently excluded relevant contributions from non-English-speaking research communities, particularly those in regions with active cyberwarfare research programs. Furthermore, the retrieval strategy relied on a predefined Boolean query structure. While these keywords were iteratively refined to capture core terminology in the field, the rapid evolution of adversarial machine learning means that newer, non-standardized terminology may exist outside the scope of our search string.

Exclusion of Gray Literature. We deliberately chose to exclude gray literature, such as technical white papers, security advisories, and code repositories, in favor of peer-reviewed scientific contributions. While this ensures a basic level of methodological rigor, it represents a significant disadvantage in the context of cybersecurity. The operational security landscape evolves faster than the academic publication cycle; consequently, cutting-edge attack vectors and proof-of-concept exploits often appear in industry reports or GitHub repositories months or years before they are formally characterized in scientific journals. Therefore, our review reflects the state of the art in academic research, which may lag slightly behind the state of practice in operational threat environments.

Reproducibility and Comparative Analysis. Perhaps the most significant limitation governing this review is the systematic lack of research artifacts. As noted in our reproducibility assessment (see Section 5.5), the vast majority of reviewed studies do not publish source code, trained model weights, or adversary datasets. This opacity precludes any possibility of a direct quantitative meta-analysis or experimental benchmarking. Without access to the original implementations, we were unable to empirically validate the reported evasion rates or test the models under identical conditions. Consequently, the comparative analysis presented in this paper is qualitative rather than quantitative, relying on reported metrics that, as discussed in Section 5.4, are often derived from heterogeneous experimental settings. Readers are advised to interpret performance comparisons between studies as indicative trends, rather than absolute measures of superiority.

8. Trends and Future Directions in Research

The analysis of the current state of adversarial DGA research highlights a critical turning point. While offensive techniques have matured into semantically and structurally sophisticated engines, defensive strategies and assessment methodologies lag behind, trapped in static paradigms. To remain relevant, this field requires a fundamental shift in how threats are modeled, defenses are designed, and progress is measured.

The analysis of the current state of adversarial DGA research highlights a critical inflection point. While offensive techniques have evolved towards semantic and structural sophistication, defensive strategies and evaluation methodologies require a corresponding paradigm shift to maintain relevance.

The Evolving Threat Landscape. Adversaries are rapidly moving from simple randomization to contextual evasion. LLMs present a transformative challenge. As Ren et al. (2023) points out, the generative capabilities of LLMs allow the creation of unknown DGA types that mimic the fluency of natural language, completely evading traditional entropy-based filters. A significant evolution lies in hybrid generation approaches (Selvaraj and Panjanathan, 2024), which connect character- and word-based analysis. By combining linguistic meaning with specific statistical patterns, these attacks create blind spots that neither purely lexical nor purely statistical detectors can cover. The threat surface is expanding into the structural dimension. Nazzal et al. (2024) demonstrated that multi-instance attacks can now leverage DNS relation graphs to coordinate evasion against graph neural networks. Future research should anticipate multimodal adversaries capable of exploiting semantic, statistical, and topological vulnerabilities simultaneously.

Defense Imperatives. To counter these constantly evolving threats, defensive research must shift from maximizing accuracy on static datasets to ensuring operational resilience. Reliance on lexical detection of a single feature is becoming obsolete. Future architectures must prioritize integrating secondary information, such as WHOIS data (Charan et al., 2020), DNS metadata (Sivaguru et al., 2020), and traffic volume analysis (Hu et al., 2023), to create systems where successful evasion requires the simultaneous manipulation of consistent features across multiple network layers. Furthermore, given the rapid conceptual drift in DGA families, static models are unsustainable. Researchers should explore lightweight continuous learning frameworks (Ren et al., 2023) and game-theory approaches (Nie et al., 2024) that allow for frequent, low-cost retraining (e.g., daily or weekly) to mitigate catastrophic forgetting without prohibitive computational overhead. Academic effectiveness often does not translate into production. Future proposals should be rigorously evaluated considering strict operational constraints, in particular, keeping FPR below 0.10% and meeting latency requirements of less than 100 ms for online detection (Peck et al., 2019; Sivaguru et al., 2020). Finally, zero-day DGAs remains a priority through cross-family generalization. Future work should investigate transfer learning and zero-shot detection capabilities Ren et al. (2023); Selvaraj and Panjanathan (2024) to identify novel DGA families that share latent structural characteristics with known families, reducing the dependency on exhaustive training datasets.

Methodological Standardization. Progress is currently hampered by a reproducibility crisis, as discussed above. To distinguish genuine scientific advances from experimental noise, the field must adopt rigorous standards. The scarcity of public code is a major barrier. We advocate for community standards that require the publication of model weights, random seeds, and generation code as a prerequisite for publication. Similarly, the use of heterogeneous datasets prevents valid comparisons between studies. Adopting unified benchmarks that standardize benign and malicious sources is mandatory to ensure that performance improvements are due to architectural superiority and not differences in data quality.

9. Conclusion

We presented a systematic literature review on the adversarial DGA research, comprising 32 primary studies conducted between 2016 and 2025. Our analysis traced the trajectory of evasion techniques, from early heuristic modifications to advanced deep generative models, while simultaneously assessing the robustness of defensive countermeasures and the methodological rigor of the field. Our analysis also reveals a structural imbalance that favors the adversary. Attackers benefit from unlimited offline optimization for perfect evasion, whereas defenders face strict online latency and near-zero false positive constraints. Consequently, static defenses based on superficial syntactic features do not generalize against modern generative models, rendering reactive strategies unsustainable. We also found that a reproducibility crisis fundamentally hampers progress. With source code available for only 12.50% of studies and a lack of standardized benchmarks, the field suffers from an inability to verify results or fairly compare techniques. Furthermore, a disconnect exists between academic optimization (aggregate accuracy) and operational reality (performance and accuracy), leading to the proposal of theoretically sound but practically unworkable solutions.

Ultimately, the advancement of DGA defense depends less on architectural complexity and more on methodological rigor. To bridge the gap between theoretical evasion and operational robustness, the community must demand open science practices such as releasing reproducible artifacts and adopt evaluation protocols that rigorously reflect the constraints of production environments.

Acknowledgments

This research was supported in part by grant PID2023-151467OA-I00 (CRAPER), funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, by grant TED2021-131115A-I00 (MIMFA), funded by MICIU/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, by grant *Proyecto Estratégico Ciberseguridad EINA UNIZAR*, funded by the Spanish National Cybersecurity Institute (INCIBE) and the European Union NextGenerationEU/PRTR, by grant *Cátedra Internacional de Ciberseguridad UNIZAR*, funded by the Spanish National Cybersecurity Institute (INCIBE) and the European Union NextGenerationEU/PRTR, by grant *Programa de Proyectos Estratégicos de Grupos de Investigación* (DisCo research group, ref. T21-23R), funded by the University, Industry and Innovation Department of the Aragonese Government.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used Gemini 3 Pro AI model to improve readability and language. After using this tool/service, the authors reviewed and edited the content as needed and assume full responsibility for the content of the publication.

References

Akram, Z., Majid, M., Habib, S., 2021. A systematic literature review: Usage of logistic regression for malware detection, in: 2021 International Conference on Innovative Computing (ICIC), pp. 1–8.

Alaeiyan, M., Parsa, S., P., V., Conti, M., 2020. Detection of algorithmically-generated domains: An adversarial machine learning approach. Computer Communications 160, 661–673.

Alexa, . Alexa Top 1 Million Domains feed. [Online; http://s3.amazonaws.com/alexa-static/top-1m.csv.zip].

Ampatzoglou, A., Bibi, S., Avgeriou, P., Verbeek, M., Chatzigeorgiou, A., 2019. Identifying, categorizing and mitigating threats to validity in software engineering secondary studies. Information and Software Technology 106, 201–230.

Anderson, H.S., Woodbridge, J., Filar, B., 2016. DeepDGA: Adversarially-Tuned Domain Generation and Detection, in: Proceedings of the 2016 ACM Workshop on Artificial Intelligence and Security, Association for Computing Machinery, New York, NY, USA. p. 13–21.

Antonakakis, M., Perdisci, R., Nadji, Y., Vasiloglou, N., Abu-Nimeh, S., Lee, W., Dagon, D., 2012. From Throw-Away Traffic to Bots: Detecting the Rise of DGA-Based Malware, in: 21st USENIX Security Symposium (USENIX Security 12), USENIX Association. pp. 491–506.

Bambenek, . Bambenek Consulting - master feeds. [Online; https://osint.bambenekconsulting.com/feeds/].

Bilge, L., Kirda, E., Kruegel, C., Balduzzi, M., 2011. Exposure: Finding Malicious Domains Using Passive DNS Analysis, in: Network and Distributed System Security, pp. 1–17.

Bilge, L., Sen, S., Balzarotti, D., Kirda, E., Kruegel, C., 2014. Exposure: A Passive DNS Analysis Service to Detect and Report Malicious Domains. ACM Trans. Inf. Syst. Secur. 16.

Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A.M., Józefowicz, R., Bengio, S., 2016. Generating Sentences from a Continuous Space, in: Goldberg, Y., Riezler, S. (Eds.), Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016, ACL. pp. 10–21.

Cao, H., Wang, C., Huang, L., Cheng, X., Fu, H., 2020. Adversarial DGA Domain Examples Generation and Detection, in: Proceedings of the 2020 1st International Conference on Control, Robotics and Intelligent System, Association for Computing Machinery, New York, NY, USA. p. 202–206.

Carlini, N., Wagner, D., 2017. Towards Evaluating the Robustness of Neural Networks, in: 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57.

Charan, P.V.S., Shukla, S.K., Anand, P.M., 2020. Detecting word based DGA domains using ensemble models, in: Krenn, S., Shulman, H., Vaudenay, S. (Eds.), Cryptology and Network Security, Springer International Publishing. pp. 127–143.

Cybersecurity Ventures, 2025. Cybercrime To Cost The World $12.2 Trillion Annually By 2031. [Online; https://cybersecurityventures.com/official-cybercrime-report-2025/]. Accessed on 1 December, 2025.

Drichel, A., Meyer, M., Meyer, U., 2024a. Towards Robust Domain Generation Algorithm Classification, in: Proceedings of the 19th ACM Asia Conference on Computer and Communications Security, Association for Computing Machinery, New York, NY, USA. pp. 2–18.

Drichel, A., von Querfurth, B., Meyer, U., 2024b. Extended Abstract: A Transfer Learning-Based Training Approach for DGA Classification, in: Maggi, F., Egele, M., Payer, M., Carminati, M. (Eds.), Detection of Intrusions and Malware, and Vulnerability Assessment, Springer Nature Switzerland, Cham. pp. 381–391.

Ebrahimi, J., Rao, A., Lowd, D., Dou, D., 2018. HotFlip: White-Box Adversarial Examples for Text Classification, in: Gurevych, I., Miyao, Y. (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia. pp. 31–36.

Fu, Y., Yu, L., Hambolu, O., Ozcelik, I., Husain, B., Sun, J., Sapra, K., Du, D., Beasley, C.T., Brooks, R.R., 2017. Stealthy Domain Generation Algorithms. IEEE Transactions on Information Forensics and Security 12, 1430–1443.

Gao, Y., Li, Z., Yuan, F., Zhang, X., Wang, D., Cao, C., Liu, Y., 2023. Robust Malicious Domain Detection Against Adversarial Attacks on Heterogeneous Graph, in: 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2028–2033.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial networks, in: Advances in Neural Information Processing Systems (NIPS), pp. 2672–2680.

Gould, N., Nishiyama, T., Kamiya, K., 2020. Domain Generation Algorithm Detection Utilizing Model Hardening Through GAN-Generated Adversarial Examples, in: Wang, G., Ciptadi, A., Ahmadzadeh, A. (Eds.), Deployable Machine Learning for Security Defense, Springer International Publishing, Cham. pp. 84–101.

He, H., Garcia, E.A., 2009. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering 21, 1263–1284.

Hernandes, E., Zamboni, A., Fabbri, S., Di Thommazo, A., 2012. Using GQM and TAM to evaluate StArt-A tool that supports systematic review. CLEI Electronic Journal 15.

Hu, X., Chen, H., Li, M., Cheng, G., Li, R., Wu, H., Yuan, Y., 2023. ReplaceDGA: BiLSTM-Based Adversarial DGA With High Anti-Detection Ability. IEEE Transactions on Information Forensics and Security 18, 4406–4421.

Kitchenham, B., Brereton, O.P., Budgen, D., Turner, M., Bailey, J., Linkman, S., 2009. Systematic literature reviews in software engineering–a systematic literature review, in: Information and Software Technology, pp. 7–15.

Liu, Q., Yu, G., Wang, Y., Yi, Z., 2021a. A Novel DGA Domain Adversarial Sample Generation Method By Geometric Perturbation, in: Proceedings of the 3rd International Conference on Advanced Information Science and System, Association for Computing Machinery, New York, NY, USA.

Liu, W., Zhang, Z., Huang, C., Fang, Y., 2021b. CLETer: A Character-level Evasion Technique Against Deep Learning DGA Classifiers. EAI Endorsed Trans. Security Safety 7, e5.

Lockheed Martin, . The Cyber Kill Chain. [Online; https://lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html]. Accessed on 1 December, 2025.

Luo, J., Qin, S.H., Wang, Z., 2026. SADGA: A Self Attention GAN-Based Adversarial DGA with High Anti-detection Ability, in: Han, J., Xiang, Y., Cheng, G., Susilo, W., Chen, L. (Eds.), Information and Communications Security, Springer Nature Singapore, Singapore. pp. 550–567.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A., 2018. Towards Deep Learning Models Resistant to Adversarial Attacks, in: International Conference on Learning Representations (ICLR).

MITRE, . Dynamic Resolution: Domain Generation Algorithms . [Online; https://attack.mitre.org/techniques/T1568/002/]. Accessed on 1 December, 2025.

Nangong, S., Wu, Z., 2025. Adversarial Sample of Domain Generation Algorithm Based on Variational Autoencoder, in: 2025 6th International Conference on Computer Engineering and Application (ICCEA), pp. 1–4.

Nayak, A.A., Venugopala, P.S., Ashwini, B., 2024. A Systematic Review on Generative Adversarial Network (GAN): Challenges and Future Directions. Archives of Computational Methods in Engineering 31, 4739–4772.

Nazzal, M., Khalil, I., Khreishah, A., Phan, N., Ma, Y., 2024. Multi-Instance Adversarial Attack on GNN-Based Malicious Domain Detection, in: 2024 IEEE Symposium on Security and Privacy (SP), pp. 1236–1254.

netlab-360, . Netlab-360 feed. [Online; https://data.netlab.360.com/dga/].

Nie, L., Shan, X., Zhao, L., Li, K., 2023. PKDGA: A Partial Knowledge-Based Domain Generation Algorithm for Botnets. IEEE Transactions on Information Forensics and Security 18, 4854–4869.

Nie, L., Zhao, L., Li, K., Shan, X., Qiu, T., 2024. A Game-Based Adversarial DGA Detection Scheme Using Multi-Level Incremental Random Forest. IEEE Transactions on Network Science and Engineering 11, 779–792.

Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., Chou, R., Glanville, J., Grimshaw, J.M., Hróbjartsson, A., Lalu, M.M., Li, T., Loder, E.W., Mayo-Wilson, E., McDonald, S., McGuinness, L.A., Stewart, L.A., Thomas, J., Tricco, A.C., Welch, V.A., Whiting, P., Moher, D., 2021. The

PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372.

Peck, J., Nie, C., Sivaguru, R., Grumer, C., Olumofin, F., Yu, B., Nascimento, A., De Cock, M., 2019. CharBot: A Simple and Effective Method for Evading DGA Classifiers. IEEE Access 7, 91759–91771.

Pelayo-Benedet, T., Rodríguez, R.J., Gañán, C.H., 2025. RAMPAGE: a software framework to ensure reproducibility in algorithmically generated domains detection. Expert Systems with Applications 293, 128629.

Pineau, J., Vincent-Lamarre, P., Sinha, K., Lariviere, V., Beygelzimer, A., d'Alche Buc, F., Fox, E., Larochelle, H., 2021. Improving Reproducibility in Machine Learning Research(A Report from the NeurIPS 2019 Reproducibility Program). Journal of Machine Learning Research 22, 1–20.

Plohmann, D., Yakdan, K., Klatt, M., Bader, J., Gerhards-Padilla, E., 2016. A comprehensive measurement study of domain generating malware. USENIX Security Symposium , 263–278.

Pochat, V.L., Goethem, T.V., Joosen, W., 2019. Evaluating the Long-term Effects of Parameters on the Characteristics of the Tranco Top Sites Ranking, in: 12th USENIX Workshop on Cyber Security Experimentation and Test (CSET 19), USENIX Association, Santa Clara, CA. p. 10.

Porras, P.A., Saïdi, H., Yegneswaran, V., 2009. A Foray into Conficker's Logic and Rendezvous Points. LEET 9, 7.

Pregardier, R.C., Bianchi, L.A.C., Garcia, V.F., Stiller, B., Silva, L.A.L., Santos, C.R.P.D., 2025. TITAN DGA: Enhancing DGA Evasiveness through a Transformer-based Autoencoder and Adversarial Self-Augmentation, in: 2025 21st International Conference on Network and Service Management (CNSM), pp. 1–9.

Ravi, V., Alazab, M., Srinivasan, S., Arunachalam, A., Soman, K.P., 2023. Adversarial Defense: DGA-Based Botnets and DNS Homographs Detection Through Integrated Deep Learning. IEEE Transactions on Engineering Management 70, 249–266.

Ren, Y., Li, H., Liu, P., Liu, J., Zhu, H., Sun, L., 2023. CL-GAN: A GAN-based continual learning model for generating and detecting AGDs. Computers & Security 131, 103317.

Sabuhi, M., Zhou, M., Bezemer, C.P., Musilek, P., 2021. Applications of Generative Adversarial Networks in Anomaly Detection: A Systematic Literature Review. IEEE Access 9, 161003–161029.

Schiavoni, S., Maggi, F., Cavallaro, L., Zanero, S., 2014. Phoenix: DGA-Based Botnet Tracking and Intelligence, in: Dietrich, S. (Ed.), Detection of Intrusions and Malware, and Vulnerability Assessment, Springer International Publishing. pp. 192–211.

Selvaraj, S., Panjanathan, R., 2024. WordDGA: Hybrid Knowledge-Based Word-Level Domain Names Against DGA Classifiers and Adversarial DGAs. Informatics 11. Cited by: 0; All Open Access, Gold Open Access.

Shu, X., Cao, C., Wang, L., Tao, F., 2022. GWDGA: An Effective Adversarial DGA, in: Cao, C., Zhang, Y., Hong, Y., Wang, D. (Eds.), Frontiers in Cyber Security, Springer Singapore, Singapore. pp. 30–48.

Siddaway, A.P., Wood, A.M., Hedges, L.V., 2019. How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. Annual review of psychology 70, 747–770.

Sidi, L., Nadler, A., Shabtai, A., 2020. MaskDGA: An Evasion Attack Against DGA Classifiers and Adversarial Defenses. IEEE Access 8, 161580–161592.

Sivaguru, R., Peck, J., Olumofin, F., Nascimento, A., De Cock, M., 2020. Inline Detection of DGA Domains Using Side Information. IEEE Access 8, 141910–141922.

Spooren, J., Preuveneers, D., Desmet, L., Janssen, P., Joosen, W., 2019. Detection of algorithmically generated domain names used by botnets: a dual arms race, in: Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, Association for Computing Machinery, New York, NY, USA. pp. 1916–1923.

Suryotrisongko, H., Musashi, Y., Tsuneda, A., Sugitani, K., 2022. Adversarial Robustness in Hybrid Quantum-Classical Deep Learning for Botnet DGA Detection. Journal of Information Processing 30, 636–644.

Tranco, . Tranco List. [Online; https://tranco-list.eu/].

Tuan, T.A., Long, H.V., Taniar, D., 2022. On Detecting and Classifying DGA Botnets and their Families. Computers & Security 113, 102549.

Wang, J., Chang, X., Wang, Y., Rodríguez, R.J., Zhang, J., 2021. LSGAN-AT: Enhancing Malware Detector Robustness against Adversarial Examples. Cybersecurity 4:38, 15.

Wang, Z., Guo, Y., 2021. Neural networks based domain name generation. Journal of Information Security and Applications 61, 102948.

Woodbridge, J., Anderson, H.S., Ahuja, A., Grant, D., 2016. Predicting Domain Generation Algorithms with Long Short-Term Memory Networks. CoRR abs/1611.00791. arXiv:1611.00791.

Xiong, W., Lagerstrom, R., 2019. Threat modeling - A systematic literature review. Computers & Security 84, 53–69.

Yilmaz, I., Siraj, A., Ulybyshev, D., 2020. Improving DGA-Based Malicious Domain Classifiers for Malware Defense with Adversarial Machine Learning, in: 2020 IEEE 4th Conference on Information & Communication Technology (CICT), pp. 1–6.

Yong Wong, M., Landen, M., Antonakakis, M., Blough, D.M., Redmiles, E.M., Ahamad, M., 2021. An Inside Look into the Practice of Malware Analysis, in: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, Association for Computing Machinery, New York, NY, USA. pp. 3053–3069.

Yu, B., Gray, D.L., Pan, J., Cock, M.D., Nascimento, A.C.A., 2017. Inline DGA Detection with Deep Networks, in: 2017 IEEE International Conference on Data Mining Workshops (ICDMW), IEEE. pp. 683–692.

Yun, X., Huang, J., Wang, Y., Zang, T., Zhou, Y., Zhang, Y., 2020. Khaos: An Adversarial Neural Network DGA With High Anti-Detection Ability. IEEE Transactions on Information Forensics and Security 15, 2225–2240.

Zamboni, A., Thommazo, A., Hernandes, E., Fabbri, S., 2010. StArt uma ferramenta computacional de apoio à revisão sistemática, in: Congresso Brasileiro de Software (CBSoft), Salvador, Brazil. pp. 91–96.

Zhai, Y., Yang, J., Wang, Z., He, L., Yang, L., Li, Z., 2022. Cdga: A GAN-based Controllable Domain Generation Algorithm, in: 2022 IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 352–360.

Zhang, K., Huang, B., Wu, Y., Chai, C., Zhang, J., Bao, Z., 2022. A WGAN-Based Method for Generating Malicious Domain Training Data, in: Sun, X., Zhang, X., Xia, Z., Bertino, E. (Eds.), Artificial Intelligence and Security, Springer International Publishing, Cham. pp. 257–270.

Zheng, Y., Yang, C., Yang, Y., Ren, Q., Li, Y., Ma, J., 2021. ShadowDGA: Toward Evading DGA Detectors with GANs, in: 2021 International Conference on Computer Communications and Networks (ICCCN), pp. 1–8.