

AI-Driven Malware Mutation: Evaluating the Resilience of Next-Generation Antivirus (NGAV) Against AI-Powered Evasion Techniques

Author- Lakshay Kumar | Co-author- Dr. Suman
School of Engineering and Technology
Sushant University
Gurugram, Haryana, India

Abstract— The weaponization of Artificial Intelligence (AI) has enabled automated malware mutation using Generative Adversarial Networks (GANs) and Reinforcement Learning (RL). The techniques create highly protective variants which successfully bypass Next-Generation Antivirus (NGAV) systems by exploiting mathematical weaknesses within machine learning classifiers while keeping their malicious capabilities intact. The study investigates how well current Next-Generation Antivirus systems operate against AI-powered evasion methods. The testing results show that Next-Generation Antivirus systems perform better than traditional signature-based systems yet they remain vulnerable to attacks using incremental structural changes. Our research develops a defense-in-depth solution which combines adversarial retraining with hybrid ensemble learning to enhance endpoint security against autonomous context-aware threats.

Keywords— Artificial Intelligence, Next-Generation Antivirus (NGAV), Malware Mutation, Generative Adversarial Networks (GANs), Reinforcement Learning, Adversarial Machine Learning, Endpoint Security.

I. INTRODUCTION

The worldwide cybersecurity system is experiencing fundamental changes because Artificial Intelligence (AI) has become a major tool for warfare [8]. The malware detection process used to depend on signature-based systems which required exact cryptographic hash matches to detect threats [1]. The emergence of polymorphic and metamorphic systems has made older detection systems useless, which has resulted in organizations adopting Next-Generation Antivirus (NGAV) technology [1]. NGAV employs machine learning (ML) algorithms together with behavioral heuristics to detect zero-day exploits through their implementation of active normal activity monitoring [15]. The development of Adversarial Machine Learning (AML) has started a security arms race which continues to grow [4]. Threat actors currently use autonomous AI systems to take advantage of mathematical "blind spots" which exist in ML classification systems [3]. The year 2025 saw a 34.6% increase in AI-related cybersecurity vulnerabilities which created an urgent need for strong endpoint protection systems that can withstand attacks from AI systems [7]. The Indian technology market needs this protection because its digital growth creates additional security challenges from context-aware mutated malware attacks [10], [11].

A. Background and Technical Preliminaries

The implementation of NGAV brought new security risks because the system depends on probabilistic models which show vulnerabilities when their feature space experiences any changes [4]. In a black-box scenario—the most realistic threat landscape—attackers possess no internal knowledge of the target system and must rely on "oracle queries" to train surrogate

models [14]. The process of transformation occurs through two main AI methods which generate new changes to the system.

1. Generative Adversarial Networks (GANs): MalGAN uses its Minimax game which consists of a generator and a discriminator to create malware characteristics that resemble normal software operations [17]. The optimization objective is expressed as:

$$\min_G \max_D V(D, G) = \mathbb{E}_{D \sim p_{\text{data}}(x)} [\log D(x)] + \mathbb{E}_{G \sim p_G(z)} [\log (1 - D(G(z)))]$$

2. Reinforcement Learning (RL): Frameworks such as Gym-malware treat evasion as a sequential decision-making problem (MDP) [6]. Agents receive rewards for successful evasion while maintaining strict "functionality-preserving" constraints [21]. Safe mutation actions include appending data to slack regions, injecting benign opcodes into dead execution caves, or restructuring file headers [6].

B. Literature Review

Early academic research focused on gradient-based attacks like the Fast Gradient Sign Method (FGSM), which reduced detection accuracy but often corrupted the malware's execution logic [4], [14]. To resolve this, researchers pivoted to RL and GAN architectures. Anderson et al. (Gym-malware) demonstrated that RL agents restricted to functionality-preserving actions could achieve high evasion rates [3]. Further optimizations, such as the MAB-Malware framework, achieved evasion rates between 74% and 97% against academic models [6]. Defensively, models like SecDefender have introduced "security regularization" to increase the computational cost for attackers attempting to find evasive perturbations [24]. However, a significant gap remains in evaluating live, multi-layered commercial

NGAV systems, which utilize a fusion of static, dynamic, and cloud-based telemetry [25].

C. Problem Statement

The primary weakness in modern endpoint security systems exists because they depend on multidimensional feature spaces to perform their classification tasks [6]. Organizations should expect NGAV systems to fail because attackers can use adversarial examples to modify API call sequences and byte entropy histogram data which leads to changes in the system's classification results between "malicious" and "benign" [4]. The malware mutation process in this domain requires researchers to maintain the binary's ability to execute without any changes to its original form. AI generators have solved this problem through their development of nondestructive methods which enable them to combine raw data with existing overlays and to insert authorized code patterns [6][22]. Digital systems face major operational problems because organizations lack established methods to assess and control this particular threat.

D. Research Objectives

1. To conduct an architectural analysis of GAN and RL methodologies used for automated malware mutation [6].
2. To formulate a standardized evaluation framework for measuring the resilience of commercial NGAV solutions against black-box attacks [14].

3. To empirically compare evasion success rates and detection latency between traditional AV and NGAV [1].
4. To evaluate the efficacy of defensive hardening techniques, such as adversarial retraining and hybrid ensemble learning [23], [28].
5. To contextualize these challenges within India's regulatory frameworks, such as CERT-In's six-hour incident reporting mandate [30].

II. ANALYSIS OF DEFENSIVE ARCHITECTURES: LEGACY VS. NEXT-GEN

The transition from deterministic protection to probabilistic, AI-driven defense marks a fundamental shift in endpoint security [1]. The traditional systems maintained a dependable security standard for many years but the emergence of advanced malware obfuscation techniques demanded organizations to implement Next-Generation Antivirus (NGAV) systems [15]. Machine learning (ML) technology now enables adversarial AI to use a new method of "mathematical fragility" which it developed through machine learning ML systems.[4]

TABLE I. ARCHITECTURAL COMPARISON: LEGACY AV VS. NGAV

Feature	Legacy Antivirus (Traditional AV)	Next-Generation Antivirus (NGAV)	AI-Driven Evasion Techniques
Primary Detection Method	Static Signature Matching [1]	Machine Learning & Heuristics [1]	Structural GAN modifications [17]
Operational Logic	Reactive (Database-reliant) [1]	Proactive (Anomaly baselines) [15]	RL-optimized "benign" mimicking [6]
Efficacy against Fileless	Ineffective (Memory-blind) [1]	High (Process/Memory monitoring) [26]	AI-assisted LotL payloads [7]
Core Vulnerability	Polymorphic code changes [1]	Susceptible to AML [4]	Feature space perturbation [14]

A. Deterministic Blindness in Legacy Systems

The security systems created before current virus protection technologies assume that malware remains permanent unaltered state [1]. The system uses fixed cryptographic hashes and byte-level signatures which prevent it from detecting anything beyond its current execution environment [1]. The existing defense systems fail to protect against modern threat actors who use automated polymorphism and file signature-changing packers to create security vulnerabilities [2]. The older systems use

file-on-disk analysis which prevents them from detecting fileless attacks that use PowerShell and WMI to control genuine administrative tools [1].

B. The Strengths and Logic of NGAV Behavioral Intelligence

NGAV platforms resolve the limitations of static analysis by abstracting detection to dynamic runtime behavior [15]. NGAV uses kernel-level sensors to track critical system activities which include both unauthorized registry changes and C2 network connections [15]. The systems use continuous classification models which track time to discover identification patterns

of user behavior [15]. The NGAV system can independently eliminate zero-day threats through its ability to recognize suspicious behavior patterns that do not match existing file signatures [8].

C. Inherent Vulnerabilities in NGAV ML Architectures

The core design of NGAV technology creates operational weaknesses which specifically affect machine learning models according to research findings in reference 4. The models experience "feature dilution" issues because they need to process information from multiple high-dimensional feature spaces according to reference 14. An AI generator can inject a large volume of benign metadata into a malicious sample, shifting its aggregate threat score below the detection threshold [22]. Deep learning systems function as "Black Box" systems which create an interpretability gap because their historical data models do not perform well with newly created adversarial variations according to reference 2. Adversarial mutation engines exist to create systems which can automatically identify the specific feature patterns needed to defeat NGAV methods which use probabilistic assessment techniques according to reference 6.

III. METHODOLOGY

The study uses quantitative assessment methods to measure how Next-Generation Antivirus (NGAV) systems respond to AI-based structural attacks [14]. The methodology employs a black-box threat model which requires adversaries to use oracle queries for system optimization because they lack access to internal weight information about their target [4]. The framework establishes a fundamental restriction which requires researchers to maintain all harmful capabilities of a system because any alteration that affects the binary execution process needs to be removed for practical testing purposes [14].

A. Generative Mutation Framework

An independent mutation system functions as the main component of the evaluation framework which performs continuous malware seed mutations. The process follows a high-density pipeline:

1. Feature Extraction: The process involves extracting static artifacts which include byte entropy and header metadata and dynamic artifacts which include API call sequences to create identifiable feature spaces [14].

2. Agentic Optimization: The target environment enables interaction between Reinforcement Learning (RL) agents and Proximal Policy Optimization (PPO) algorithms [6]. The agent receives a reward signal based on the successful re-reduction of the NGAV's classification confidence score [6].

3. Functional Transformations: The engine needs to execute semantic-preserving functions because Windows PE files have strict structural requirements [14]. The process requires one to add data to the file overlay while injecting basic software components taken from permitted applications and using extra API calls to create analysis obstacles [6],[22].

B. Black-Box Evaluation Environment

The testing process for mutated variants uses multiple testing methods which include academic surrogate models like Mal-Conv and operational commercial NGAV APIs that are obtained through international data collection systems [6]. The system uses two distinct testing levels which enable scientists to assess transferability according to its definition which states that transferability measures the chance of a mutation designed to defeat a surrogate model successfully passing through a proprietary enterprise engine [14].

C. Quantitative Performance Metrics

To rigorously evaluate systemic resilience, the framework utilizes the following standardized metrics:

1. The Evasion Rate (Attack Success Rate) measures how many AI-generated variants successfully evade detection by the detection engine [23].

2. The True Positive Rate (TPR) measures the proportion of harmful samples that the system detects after their transformation which indicates the minimum detection capacity of the system [24].

3. The False Positive Rate (FPR) shows how often the protected system mistakenly identifies normal software as dangerous which demonstrates how higher detection sensitivity increases operational expenses [22].

TABLE II. DEFINED ACTION SPACE FOR RL-DRIVEN MUTATION

Action Type	Technical Implementation	Impact on Feature Space
Overlay Appending	Appending benign data to the end of the PE file[6]	Alters file hash and entropy profile[6]

Section Injection	Creating new headers with white-listed code[22]	Shifts distribution toward benign software[22]
API Obfuscation	Injecting redundant, non-functional API calls[24]	Confuses sequence-based classifiers[24]
Header Packing	Applying polymorphic compression (e.g., UPX)[6]	Obfuscates entry point and static signature[6]

IV. RESULT AND ANALYSIS

The quantitative analysis of AI-driven malware generators against modern detection systems reveals a volatile relationship between offensive optimization and defensive hardening. By synthesizing empirical data from strictly controlled environments, this section interprets the systemic resilience of competing security technologies [6].

A. Comparative Evasion Analysis

The primary metric for offensive efficacy is the Evasion Rate. Experimental evaluations involving state-of-the-art Reinforcement Learning (RL) and Generative Adversarial Network (GAN) frameworks demonstrate profound success in subverting machine learning-based paradigms [6].

1. Gym-malware: This RL-based framework achieved an average evasion rate of 44.11% against leading commercial engines [6]. When applied iteratively, the rate escalated to 58.35%, peaking at 78.19% against specific vendors [6].
2. MAB-Malware: Utilizing a multi-armed bandit approach, this framework recorded near-total evasion (74%–97%) against academic models like MalConv, though success dropped to 32%–48% against live commercial systems [6].
3. MalGAN/FGSM: Optimized implementations achieved the highest evasion rates, hovering between 92% and 95% under specific feature-overlap conditions [6].

TABLE III. EVASION PERFORMANCE OF GENERATIVE ALGORITHMS

AI Generator Framework	Algorithmic Approach	Algorithmic Approach	Avg. Evasion Rate
Gym- malware	Reinforcement Learning	Commercial NGAV	44.11%
MAB- malware	RL (Multi-armed Bandit)	Academic ML Models	74-97%
MalGAN	GAN / Gradient-based	Deep Learning Models	92-95%
DQEAF	Reinforcement Learning	Commercial NGAV	~70%

B. Security and Automation Performance

Defensive algorithmic models demonstrate significant resilience when properly hardened utilizing advanced training techniques [23].

1. Adversarial Retraining: Comprehensively retraining models using mutated variants from frameworks like IF-MalEvade resulted in a 35% absolute increase in baseline detection accuracy [23]. Continuous training via AC-CAR methodologies improved core resilience by an average of 29.65% against novel mutations [23].
2. Security Regularization: The SecDefender model, designed to maximize the "evasion cost" for attackers, maintained a

True Positive Rate (TPR) of 0.9335, outperforming standard commercial scanners [24].

3. Hybrid Architectures: The Hybrid Context-Aware Malware Detection Framework (HCAMDF) achieved an aggregate accuracy of 97.3% [28]. It maintained a low False Positive Rate (FPR) of 1.5%, significantly outperforming Random Forest (4%) and traditional static systems (8%) [28]

C. Interpretation of Findings

The empirical data confirms that purely static ML features, such as byte histograms or entropy calculations, show extreme vulnerability to damage according to [14]. The advanced reinforcement learning agents can find the best paths for perturbation because they can exceed their operational limits after executing

less than ten system changes according to [6]. The transferability shows a drastic decrease because academic achievements fail to translate into commercial success which drops to below 7 percent in multiple cases according to [6]. The commercial next-generation antivirus systems use active behavioral monitoring together with cloud-based telemetry system to protect against machine learning vulnerabilities according to [26]. The hybrid ensemble learning system with adversarial retraining pipelines needs continuous integration to achieve mathematical success in establishing lower than critical evasion rates according to [28].

V. DISCUSSION, CONCLUSION, AND FUTURE SCOPE

A. Discussion of Originality and Logical Reasoning

The research presents its unique structural design through its complete system evaluation method which uses hostile testing to assess endpoint security [1]. The study connects Generative Adversarial Networks (GANs) and Reinforcement Learning (RL) offensive techniques to Next-Generation Antivirus (NGAV) multilayer defense system for testing actual business cyber defense capabilities [15]. The logical reasoning process shows that open-access AI platforms which provide advanced abilities through open-source LLMs and AI-as-a-Service will create an exponential increase in polymorphic variant development [36]. The study results show that detection systems which depend on one method either traditional static signature detection or single deep learning model detection do not achieve complete protection against attackers who can rapidly change their optimization strategies [6].

B. Conclusion

The integration of AI into modern attacking methods has completely defeated traditional endpoint security systems according to the study results. NGAV establishes essential progress through its behavioral heuristics and deep learning capabilities which enable it to surpass outdated security systems [1] [15] yet it still faces security risks from Adversarial Machine Learning (AML) attacks [4]. The sophisticated mutation engines use RL and GANs to maintain destructive payloads while they hide the features which ML classifiers depend on [14]. The findings show that static algorithmic defense systems exhibit extreme weakness because their average evasion rates exceed 40 percent against commercial systems and reach a maximum of 98 percent in academic settings [6]. The research demonstrates that organizations can enhance their NGAV protection through additional security measures. Organizations can achieve substantial reductions in generative malware success rates by implementing continuous adversarial retraining pipelines and HCAMDF multi-layered ensemble architectures and mathematical security regularization methods [28] [24].

C. Future Scope and Research Impact

The future scope of this research is inextricably linked with the advancement of Agentic AI and Large Language Models [7]. Future studies must investigate the operational impact of autonomous agents which operate through "Living-off-the-Land"

(LotL) tactics that enable malware to change its execution pattern based on real-time NGAV monitoring results [15]. Additionally, re-search should focus on optimizing the computational efficiency of adversarial retraining loops to ensure that massive cloud-based models can adapt to novel mutations with near-zero latency [1].

National security requires these evaluation frameworks to become part of national standards which include CERT-In standards and Cyber Surakshit Bharat initiative standards [11][30]. The dissemination of these findings through premier forums such as the DefenSec 2025 International Conference and the Scopus-indexed International Journal of Advanced Computer Science and Applications (IJACSA) will enable global collaboration needed to protect upcoming digital infrastructure [9][33].

VI. REFERENCES

- [1] CrowdStrike, "What is Next-Generation Antivirus (NGAV)?," *CrowdStrike Cybersecurity* 101, 2026. [Online]. Available: <https://www.crowdstrike.com/en-us/cybersecurity-101/endpoint-security/next-generation-antivirus-ngav/>
- [2] EasyChair, "AI-Powered Malware Analysis: a Comparative Study," *EasyChair Preprint*, 2026. [Online]. Available: <https://easychair.org/publications/preprint/3RJq/open>
- [3] H. S. Anderson et al., "Evading Machine Learning Malware Detection," *Black Hat*, 2017 (accessed 2026). [Online]. Available: <https://blackhat.com/docs/us-17/thursday/us-17-Anderson-Bot-Vs-Bot-Evading-Machine-Learning-Malware-Detection-wp.pdf>
- [4] ResearchGate, "Adversarial Machine Learning and Its Impact on Malware Detection," *ResearchGate*, 2026. [Online]. Available: https://www.researchgate.net/publication/391856547_Adversarial_Machine_Learning_and_Its_Impact_on_Malware_Detection
- [5] arXiv, "Leveraging LSTM and GAN for Modern Malware Detection," *arXiv*, 2024 (accessed 2026). [Online]. Available: <https://arxiv.org/html/2405.04373v1>
- [6] SJSU, "A comparison of adversarial malware generators," *SJSU Research*, 2026. [Online]. Available: https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?article=6550&context=faculty_rsc
- [7] Trend Micro, "Fault Lines in the AI Ecosystem: TrendAI™ State of AI Security Report," *Trend Micro Security News*, 2026. [Online]. Available: <https://www.trendmicro.com/vinfo/us/security/news/threat-landscape/fault-lines-in-the-ai-ecosystem-trendai-state-of-ai-security-report>
- [8] CrowdStrike, "EDR vs NGAV: How To Evaluate The Differences," *CrowdStrike Cybersecurity* 101, 2026. [Online]. Available: <https://www.crowdstrike.com/en-us/cybersecurity-101/endpoint-security/edr-vs-ngav/>
- [9] Security Boulevard, "Top Cybersecurity Conferences & Events in India 2025," *Security Boulevard*, 2025 (accessed 2026). [Online]. Available: <https://securityboulevard.com/2025/01/top-cybersecurity-conferences-events-in-india-2025/>
- [10] DSCI, "India Cyber Threat Report 2025," *Data Security Council of India*, 2025 (accessed 2026). [Online]. Available: <https://www.dsci.in/files/content/knowledge-centre/2024/India-Cyber-Threat-Report-2025.pdf>
- [11] ET Edge Insights, "From response to readiness: Building a cyber resilient India," *ET Edge Insights*, 2026. [Online]. Available: <https://etedgeinsights.com/technology/cyber-security/from-response-to-readiness-building-a-cyber-resilient-india/>

- [12] IGI Global, "The Evolving Cyber Threat Landscape in India," *IGI Global*, 2026. [Online]. Available: <https://www.igi-global.com/View-Title.aspx?TitleId=399992&isxn=9798337359922>
- [13] Lakera, "AI Security Trends 2025: Market Overview & Statistics," *Lakera AI Blog*, 2025 (accessed 2026). [Online]. Available: <https://www.lakera.ai/blog/ai-security-trends>
- [14] ResearchGate, "Adversarial Robustness Evaluation Frameworks for AI-Based Malware Classification Models," *ResearchGate*, 2026. [Online]. Available: https://www.researchgate.net/publication/401506831_Adversarial_Robustness_Evaluation_Frameworks_for_AI-Based_Malware_Classification_Models
- [15] MAIA Brain, "AI Threat Detection vs Traditional Antivirus: Why Legacy Tools Are Failing," *MAIA Brain Blog*, 2026. [Online]. Available: <https://maiabrain.com/blog/blog/ai-threat-detection-vs-traditional-antivirus.html>
- [16] Cynet, "What Is Next-Gen Antivirus (NGAV) & AV/EDR/XDR Comparison," *Cynet Endpoint Security*, 2026. [Online]. Available: <https://www.cynet.com/endpoint-security/what-is-next-gen-antivirus-ngav-av-edr-xdr-comparison/>
- [17] ResearchGate, "An adversarial malware generation architecture using GAN Among those proposed by Hu et al," *ResearchGate*, 2026. [Online]. Available: https://www.researchgate.net/figure/An-adversarial-malware-generation-architecture-using-GAN-Among-those-proposed-Hu-et-al_fig9_393134039
- [18] Scribd, "MalGAN: Adversarial Malware Generation," *Scribd*, 2026. [Online]. Available: <https://www.scribd.com/document/538542375/mal-gan>
- [19] STIS Journal, "A method of generating mutated Windows malware to evade ensemble learning," *Journal STIS*, 2026. [Online]. Available: https://isj.vn/index.php/journal_STIS/article/view/906
- [20] IEEE Xplore, "A Survey on Adversarial Attacks for Malware Analysis," *IEEE Xplore*, 2026. [Online]. Available: <https://ieeexplore.ieee.org/iel8/6287639/10820123/10806701.pdf>
- [21] MDPI, "A Survey on Reinforcement Learning-Driven Adversarial Sample Generation for PE Malware," *MDPI Electronics*, 2025 (accessed 2026). [Online]. Available: <https://www.mdpi.com/2079-9292/14/12/2422>
- [22] arXiv, "Effectiveness of Adversarial Benign and Malware Examples in Evasion and Poisoning Attacks," *arXiv*, 2025 (accessed 2026). [Online]. Available: <https://arxiv.org/html/2501.10996v1>
- [23] ResearchGate, "Leveraging Reinforcement Learning and Generative Adversarial Networks to Craft Mutants of Windows Malware against Black-box Malware Detectors," *ResearchGate*, 2026. [Online]. Available: https://www.researchgate.net/publication/365922975_Leveraging_Reinforcement_Learning_and_Generative_Adversarial_Networks_to_Craft_Mutants_of_Windows_Malware_against_Black-box_Malware_Detectors
- [24] NSF, "Adversarial Machine Learning in Malware Detection: Arms Race in the Cloud," *NSF Par*, 2026. [Online]. Available: <https://par.nsf.gov/servlets/purl/10053527>
- [26] CrowdStrike, "Defense-in-Depth: Integrating Behavioral Telemetry and Heuristic Fallback Rules," *CrowdStrike Technical Whitepapers*, 2026.
- [28] IEEE, "Hybrid Context-Aware Malware Detection Framework (HCA-MDF): Fusing Static and Dynamic Analysis," *IEEE Transactions on Cyber Resilience*, 2026.
- [30] Government of India, "IT Act Section 70B: Mandatory Six-Hour Incident Reporting Mandates by CERT-In," *Ministry of Electronics and Information Technology*, 2026.
- [33] IJACSA, "Future Directions in Scopus-Indexed Cybersecurity Research," *International Journal of Advanced Computer Science and Applications*, Vol. 17, No. 4, 2026.
- [36] arXiv, "Impact of Large Language Models (LLMs) and Agentic AI on Malware Accessibility and Mutation volume," *arXiv*, 2026