Mitigating Adversarial Threats in ML-Based Malware Detection: Attacks and Robust Defenses

1st Muhammad Ahmad Ejaz
Department of Information and Communication Engineering
The Islamia University of Bahawalpur
muhammadahmadejaz40@gmail.com

2nd Abdul Basit
Department of Information and Communication Engineering
The Islamia University of Bahawalpur
gopangmeer555@gmail.com

3rd Aoun Muhammad
Department of Information and Communication Engineering
The Islamia University of Bahawalpur
aoun.muhammad@iub.edu.pk

4th Sana Tariq
Department of Information and Communication Engineering
The Islamia University of Bahawalpur
sana.tariq@iub.edu.pk

***Abstract*—The evolution of the internet posed several challenges, and cybersecurity became a mandate to ensure the continuity of services and the protection of digital assets. These challenges involved malware and social engineering tactics that made traditional security inefficient. The adoption of AI initially worsened the security posture because AI was considered a double-edged sword (having benefits and drawbacks). However, the adoption of AI and its subdomains (ML, deep learning) has transitioned security from conventional to advanced detection methods. In the past, signature-based detectors were used to detect malware; now ML-based detection systems help to detect malware types and even zero-day vulnerabilities that were too complex for traditional detectors to identify. Moreover, ML-based detection systems are highly capable of identifying threats and mitigating their effects effectively. With time, models also became vulnerable to adversarial attacks, which degrade the performance and reliability of the model. This paper illustrates related papers, highlights current and future challenges, and defense recommendations to build a robust security landscape for the future.**

***Index Terms*—Adversarial Machine Learning(AML), Malware, Artificial Intelligence, Evasion Attacks, Proactive Security, Multilayer Defense, IDS, IPS, Zero-day vulnerability, GAN, Data Augmentation, LLMs.**

I. Introduction

Malware is malicious software designed to harm the normal functionality of a particular device or software. It is further categorized into subclasses according to malware's behavior and the attacker's intention; trojans, ransomware, rootkits, spyware, keyloggers, logic bombs and some others are involved. As technology grows, new attacks are introduced at every moment. The rapid adoption of the internet has increased the attack surface for threat actors. To combat security threats, security professionals have introduced defensive techniques (encryption, authentication, access control and privilege control), diverse security approaches (proactive security, multilayer defense and others) and security devices (firewalls and detectors) to ensure a secure environment. This includes ensuring the continuity of services and preventing data theft, unauthorized access, and complete service shutdowns. Without an effective detection approach, organizations may face a weakening of users' trust and experience, financial and reputational loss as well as performance and service degradation. The continuous development of sophisticated malware and attacks has compromised traditional security defenses, which has led to the adoption of Artificial Intelligence and Machine Learning detection strategies. The use of AI and ML models in real-time detection systems has brought maximum accuracy and precision. ML models are trained on massive datasets to identify new and sophisticated attacks that were undetectable by traditional (signature and behavior-based) detectors [1], [2].‘

The fast adoption and implementation of ML models in organizations and critical systems have also invited adversaries to actively engage with these models to misclassify their decisions. Attackers attempt to evade models to achieve their malicious desires, and these evasion tactics are known as Adversarial Machine Learning (AML) attacks. Adversarial techniques basically involve the modification or manipulation of malware by adding extra bytes, safe strings, and benign functions to bypass the classifier's security without breaking its core functionality.

To some extent, adversarial ML attacks have successfully bypassed malware detection systems (IPS, IDS, and commercial malware detection), creating serious concerns about security breaches [3]. In this survey paper, our main target will be common and advanced adversarial attacks against ML models, defense techniques and deployment challenges by analyzing multiple survey papers, which would help security researchers to protect organizations from current and upcoming AML threats.

II. Background, Existing Work and Methodology for Literature Collection

A. Background

Recent papers have demonstrated that machine learning-based malware detection systems are highly vulnerable to

DOI: https://doi.org/10.67535/tsp.000003.009
Published under the ICACNC 2026 Proceedings

adversarial, model theft, and common adversarial attacks. Researchers have proposed various attack strategies and defense mechanisms, but achieving robust and effective prevention remains an open challenge. Building foundational studies on adversarial examples and analyzing defense strategies from previous surveys helps to highlight considerable prevention methods and future challenges that affect model performance.

Initially, Machine Learning achieved significant milestones in identifying threats within detection systems. But over time, attackers exploited the vulnerabilities of ML models by manipulating programs without breaking their core functionality, creating sophisticated malware that misclassified the model [4], [5]. Furthermore, as these adversarial techniques became widely adopted, recent papers have evaluated attack vectors and defense techniques designed to overcome these challenges.

As ML models advance through the implementation of various defense mechanisms, AML techniques have also become more sophisticated and aggressive in attempting to control models' decisions [6], [7], [8]. These papers emphasize that systems will remain vulnerable to these AML attacks if we do not adopt defense accordingly adversarial tactics, and show various future challenges and defense mechanisms.

B. Existing work

TABLE I
ADVERSARIAL MACHINE LEARNING ATTACKS ON MALWARE DETECTORS (2024–2026). TM: THREAT MODEL, AE: ADVERSARIAL EXAMPLE.

Paper	Year	Application Domain	Taxonomy	TM	AE
		— 2024 —			
Bostani *et al.* (EvadeDroid) [9]	2024	Android Malware	Attack Type	✓	✓
Krishna *et al.* (ARC-Net) [10]	2024	Malware Detect.	Defense	✓	✓
Lucas *et al.* (Greedy-Block) [11]	2024	Raw-Binary Malware	Defense	✓	✓
		— 2025 —			
Sang *et al.* (GanGenetic) [12]	2025	Malware Detect.	Attack Type / Algorithm	✓	✓
Rando *et al.* (ZEXE) [13]	2025	Windows Malware	Algorithm	✓	✓
Ouazza *et al.* [14]	2025	Malware Detect.	Attack Type / Defense	✓	✓
Greco *et al.* [15]	2025	Image-based Malware	Algorithm	✓	✓
		— 2026 —			
Li *et al.* (ARNDroid) [16]	2026	Malware Detect.	Defense	✓	✓

Recent research by Tianwei Lan and Farid Nait-Abdesselam (2025) [17] reflected the advancement in AML attacks by using large language models (LLMs) to produce highly evasive, stealthy, functional, and efficient malware capable of bypassing ML-based malware detectors. They also demonstrated effective defense methods utilizing LLMs to resist AML abuses. This paper showed an effective defense but brought the threats to light such as a lack of massive datasets to train the model, complex model design by LLMs, advanced adversarial attacks through LLMs, increased processing time, and the possibility of database poisoning.

Similarly, Sashidhar Chinthala (2026) [18] highlighted attack surfaces and defense directions related to adversarial attacks on malware detection systems. Moreover, this paper emphasized model evolution by providing adversarial training and input sanitization (denoising and feature rounding before reaching the main model). It also discussed common attacks like data poisoning, evasion techniques, perturbation, and model extraction, which render ML-based detectors purposeless in identifying threats.

C. Methodology For Literature Collection

To ensure comprehensive coverage of the topic, a systematic literature collection process was adopted. To the point and related studies were identified through major academic databases including IEEE Xplore, ACM Digital Library, arXiv, SpringerLink, ScienceDirect, and Google Scholar, using keywords such as "adversarial machine learning", "adversarial ML attacks", "adversarial defenses", "adversarial AI "and "adversarial attacks against malware detectors " to retrieve relevant publications.

Moreover, Priority was given to recent publications to understand the recent work in adversarial tactics research, particularly advances in machine learning, deep learning and large language model (LLM)-based security applications. The selected studies were then analyzed and categorized according to attack type, defense strategy, and reported limitations.

III. ADVERSARIAL ATTACKS

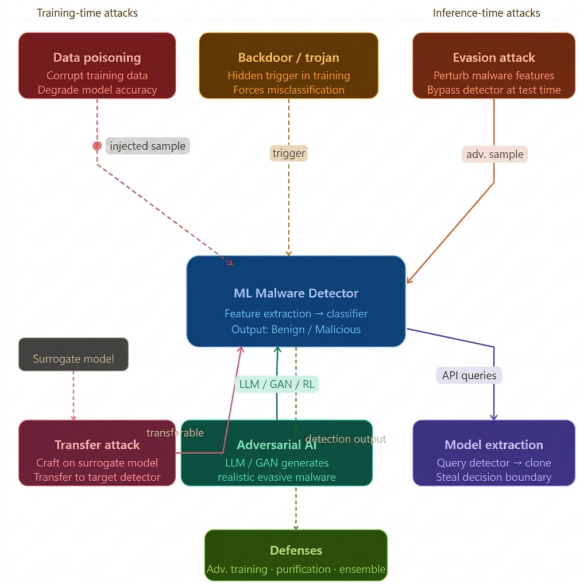

Fig. 1. Adversarial Attacks Simulation

Recently, ML-based malware detectors have become an indispensable component because traditional signature-based

detection mechanisms are no longer effective in tackling sophisticated malware. As a result, ML-based malware detection systems have the ability to automatically learn complex patterns from large-scale datasets and are widely adopted due to their high detection accuracy.

Despite this high detection accuracy, they are vulnerable to specific attacks that raise serious concerns regarding the robustness and reliability of an organization's security environment. These attacks are known as "Adversarial Attacks" and are considered one of the most critical threats to ML-based security systems, and many solutions have been proposed to make these threats unsuccessful.

Adversarial attacks refer to techniques that attempt to misclassify an ML-based detection system using various strategies, such as perturbation, padding, direct engagement with the model to learn its outputs, and reordering instructions. The implementation of and overreliance on ML-based malware detectors have exponentially increased the number of AML attacks. At every moment, unknown adversarial attacks are being introduced, creating ongoing challenges for detection systems.

Our objective is to highlight AML attacks that are proving to be a challenge for detection systems, provide practical recommendations for resilient models against AML abuses, and outline some future challenges for ML-based malware detectors.

A. Types of Attacks

Well-known adversarial attack against ML-Based systems:

1) Poisoning and Backdoor Attack (Training Time): Firstly, before moving to advanced techniques to evade malware detectors, the most common type of AML attack is the poisoning attack. A poisoning attack involves injecting malicious, unusable, or mislabeled datasets during the ML model's training phase to damage its learning process and mislead its decisions.

Meanwhile, a backdoor is a stealthy training-time adversarial technique in which an attacker embeds a malicious executable file into the training datasets. This causes the model to behave normally under standard conditions but misclassify inputs containing the logic of the malicious file. For instance, if our model is designed to detect malware, providing it with incorrect malware signatures and ineffective datasets prevents it from learning the patterns needed to detect well-known malware and zero-day vulnerabilities (vulnerabilities unknown to vendors).

2) Evasion Attacks: An evasion attack is a critical vulnerability in ML models that is easily exploitable by adversaries. The attacker slightly modifies the malware file during the testing phase in order to evade the trained ML model without altering the malware's core functionality. Some common methods include: adding benign API calls, inserting dead code, code obfuscation, packing or encryption, reordering instructions, padding executable sections, modifying header information, and perturbation [19].

3) Model Extraction and Privacy Disclosure: Model extraction refers to stealing the ML models' functionality by overwhelming it with repeated queries in an attempt to rebuild an approximate copy of the genuine model. Model extraction enables an attacker to steal the behavior of the model through queries, reveal stronger evasion techniques against the model, and take over the security of the organization.

4) Adversarial AI and Transfer Attack: Artificial Intelligence is also being used as an offensive strategy to launch highly effective attacks against ML-based detector systems. Google published a report on the "Adversarial Misuse of Generative AI" in 2025, showing that adversaries are utilizing AI-driven methods to draft complex malware and launch sophisticated attacks. Adversarial AI involves the use of ML models to generate adversarial inputs that can misclassify the prediction of ML-based detector systems. A transfer attack is defined as creating a malicious sample using one ML model to attack another ML model; such an approach can easily evade and mislead models' decisions [20], [21].

IV. Defense Strategies

Emergence of adversarial attacks against ML-based detection systems is degrading the performance and accuracy of detection systems to identify threats. However, these attacks would be seen as fully evasive, effective, and more sophisticated with time by leveraging AI, which can cause services shutdown, asset theft, financial and reputational loss. To make these attacks less intense, we need to introduce comprehensive, effective, and robust measures that would deny such abuses (AML attacks), make sure the prevention of models, and offer a secure environment to move ahead

A. Adversarial Training

To combat adversarial attacks, adversarial training is one of the most widely adopted and real-time implemented defense mechanisms against adversarial attacks, where the model is trained by providing datasets of modified malware samples, resulting to improve the performance and accuracy of ML based detection systems to identify fully undetectable malicious samples. Moreover, continuous adversarial training must be implemented side by side, as new evasion attacks can be introduced. Adversarial training is capable of strengthening the model's classification boundaries and decreasing the evasion success rate. .

B. Preprocessing and Input Sanitization

Preprocessing and input sanitization are techniques that are responsible for processing the input sample before moving to the main model. The approach deals with removing raw bytes, analyzing every possible suspicious entity, checking and reordering the code sequence that the attacker prepared to bypass and misclassify the decision of the model.

C. Combining Detectors and a Hybrid Approach (static and dynamic analysis):

Integrating multiple detector models (LLMs, deep learning), along with a hybrid approach, can exponentially improve the

result of detection. Further, multiple learning models analyze input samples from different aspects. As a result, the chances of evasion become almost nonexistent.

Whereas hybrid analysis comprehensively is the combination of static and dynamic methods, where malware is examined through file structure, code, and byte patterns without running the malware. This approach is known as static analysis, while in dynamic analysis, malware is tested during execution time to check its behavior. This combined approach can enhance network latency but provide efficient security of the model and data, and contribute to development a robust model [22], [23].

D. Continuous Monitoring and Adversarial Red Teaming

Last but not least, it is a proactive security practice which ensures the performance, reliability, and accuracy of ML-based detection models. Continuous monitoring involves periodically observing the performance and behavior of ML-based detection models to identify wrong decisions or suspicious behavior that may indicate adversarial manipulation, which helps to fix errors and bring the model into routine tasks.

Adversarial red teaming refers to security experts actively engaged with the models by producing fully evasive malware to inspect adversarial vulnerabilities in the working model. Implementing these strategies into the ML-based models, which will keep the ML-model effective against growing adversarial challenges and other ML-based vulnerabilities like jailbreaking (specially crafted prompts which remove restrictions on AI/ML models) and so on [24].

V. Current Challenges,Proposed Solutions and Future work

A. Current Challenges and Proposed Solutions

TABLE II
Current Challenges and Proposed Solutions in ML-Based Malware Detection

Current Challenge	Proposed Solution	Quantitative Justification
AI-Powered Evasion Attacks (e.g., FGSM, GAN-generated malware)	Ensemble of multiple models & robust adversarial retraining.	Reduces Attack Success Rate (ASR) from $> 90\%$ to $< 5\%$.
Data Imbalance (Scarce novel malware families)	GAN-based synthetic data generation & open-source libraries (MalwareBazaar) [25].	Increases minority class Area Under Curve (AUC) by 12–18%.
Resource-Limited Edge Devices (IoT environments)	Cloud-based services to overcome resource-limited models	Reduces CPU load by 40%; shrinks model size 10x with $< 2\%$ accuracy drop.

B. Future Work

- Implement continuous monitoring and automated defense updates.
- Develop computationally efficient certified robustness techniques.
- Utilize quantum computing and lightweight architectures to resolve issues of computational complexity in models.
- Integrate advanced models (such as deep learning architectures, GANs, and LLMs) to proactively highlight threats [26].

VI. Conclusion

ML-based malware detection systems have exponentially improved the ability to identify and classify malicious software. However, despite their capability to detect threats, these systems are at risk from adversarial attacks that evade detection by slightly modifying the malicious samples. In this paper, we explored various attack techniques such as poisoning attacks, evasion attacks, model extraction, privacy disclosure, and backdoor attacks, which pose serious threats to the reliability and performance of ML-based detectors.

To address these challenges, several defense mechanisms have been explored, including adversarial training, input preprocessing and sanitization, hybrid analysis, and the integration of multiple detection models. Additionally, continuous monitoring and adversarial red teaming are the step forward defense that help to maintain the performance and robustness of deployed systems against sophisticated abuse. In short, building a resilient and effective machine learning-based malware detector requires a combination of robust models, hybrid analysis techniques, and proactive security strategies to successfully defend against existing and emerging adversarial threats.

References

[1] Y. Wu, H. Zhuang, Y. Jia, and Y. Zhang, "A survey of machine learning approaches for malware detection," in *Proc. 5th Int. Conf. Comput. Netw. Secur. Softw. Eng. (CNSSE)*, Qingdao, China, 2025, pp. 269–273, doi: 10.1145/3732365.3732410.

[2] P. Maniriho, A. N. Mahmood, and M. J. M. Chowdhury, "A survey of recent advances in deep learning models for detecting malware in desktop and mobile platforms," *ACM Comput. Surv.*, vol. 56, no. 6, pp. 1–41, 2024.

[3] A. Abomakhelb, K. A. Jalil, A. G. Buja, A. Alhammadi, and A. M. Alenezi, "A comprehensive review of adversarial attacks and defense strategies in deep neural networks," *Technologies*, vol. 13, no. 5, p. 202, 2025, doi: 10.3390/technologies13050202.

[4] E. Anthi, L. Williams, M. Rhode, P. Burnap, and A. Wedgbury, "Adversarial attacks on machine learning cybersecurity defences in industrial control systems," *J. Inf. Secur. Appl.*, vol. 58, p. 102717, 2021.

[5] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proc. 4th ACM Workshop Secur. Artif. Intell. (AISec)*, 2011, pp. 43–58.

[6] X. Wang, J. Li, X. Kuang, Y. Tan, and J. Li, "The security of machine learning in an adversarial setting: A survey," *J. Parallel Distrib. Comput.*, vol. 130, pp. 12–23, 2019, doi: 10.1016/j.jpdc.2019.03.003.

[7] M. Rigaki and S. Garcia, "A survey of privacy attacks in machine learning," *ACM Comput. Surv.*, vol. 56, no. 4, Art. 145, 2024, doi: 10.1145/3624010.

[8] M. B. Mwangi and S. M. Cheng, "An adversarial attack on ML-based IoT malware detection using binary diversification techniques," *IEEE Access*, vol. 12, pp. 1–15, 2024, doi: 10.1109/ACCESS.2024.3513713.

[9] H. Bostani and V. Moonsamy, "EvadeDroid: A practical evasion attack on machine learning for black-box Android malware detection," *Comput. Secur.*, vol. 139, Art. 103676, 2024.
[10] G. B. Krishna, G. S. Kumar, M. Ramachandra, K. S. Pattem, D. S. Rani, and G. Kakarla, "Adapting to evasive tactics through resilient adversarial machine learning for malware detection," in *Proc. 11th Int. Conf. Comput. Sustain. Global Dev. (INDIACom)*, 2024.
[11] K. Lucas, W. Lin, L. Bauer, M. K. Reiter, and M. Sharif, "Training robust ML-based raw-binary malware detectors in hours, not months," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2024.
[12] A. Sang, Z. Wang, L. Yang, L. Zhou, J. Jia, and H. Yang, "Generating adversarial malware examples against multiple machine learning detectors," *IEEE Trans. Ind. Inform.*, vol. 21, no. 7, 2025.
[13] M. Rando, L. Demetrio, L. Rosasco, and F. Roli, "A new formulation for zeroth-order optimization of adversarial EXEmples in malware detection," *IEEE Trans. Inf. Forensics Security*, vol. 21, pp. 506–515, 2025.
[14] H. Ouazza, F. Khennou, and A. Abdellaoui, "Adversarial retraining and white-box attacks for robust malware detection," in *Proc. 13th Int. Symp. Digit. Forensics Secur. (ISDFS)*, 2025.
[15] C. Greco, M. Ianni, A. Guzzo, and G. Fortino, "Evasion of deep learning malware detection via adversarial selective obfuscation," in *Proc. IEEE Int. Conf. Cyber Secur. Resilience (CSR)*, 2025.
[16] Q. Li, D. Wu, C. Lin, S. Liu, C. Wang, and C. Shen, "Cross-region feature reformer with semantic preservation for adversarial malware detection," *IEEE Trans. Inf. Forensics Security*, vol. 21, 2026.
[17] T. Lan and F. Nait-Abdesselam, "Adversarial agents: LLM-powered attacks and defenses for Android malware detectors," in *Proc. IEEE Conf. Dependable Secur. Comput. (DSC)*, Taipei, Taiwan, 2025.
[18] S. Chinthala, "Adversarial machine learning in cybersecurity: Attacks, defenses, robustness, and explainability," in *Proc. 5th IEEE Int. Conf. AI Cybersecurity (ICAIC)*, Houston, TX, USA, 2026, pp. 1–4.
[19] R. Muthalagu, J. Malik, and P. M. Pawar, "Detection and prevention of evasion attacks on machine learning models," *Expert Syst. Appl.*, vol. 266, p. 126044, 2025, doi: 10.1016/j.eswa.2024.126044.
[20] S. L. Schröer, L. Pajola, A. Castagnaro, G. Apruzzese, and M. Conti, "Exploiting AI for attacks: On the interplay between adversarial AI and offensive AI," *IEEE Intell. Syst.*, pp. 1–10, 2025, doi: 10.1109/mis.2025.3610364.
[21] N. Akhtar, A. Mian, N. Kardan, and M. Shah, "Advances in adversarial attacks and defenses in computer vision: A survey," *IEEE Access*, vol. 9, pp. 155161–155196, 2021, doi: 10.1109/ACCESS.2021.3127960.
[22] M. H. A. Rai, Y. Noor, M. Faisal, and M. F. Nawazish, "Adversarial robustness of deep learning-based intrusion detection systems against AI-powered cyber attacks," *SES*, vol. 3, no. 11, pp. 899–922, Nov. 2025.
[23] S. Ali, J. Wang, and V. C. M. Leung, "AI-driven fusion with cybersecurity: Exploring current trends, advanced techniques, future directions, and policy implications for evolving paradigms–A comprehensive review," *Inf. Fusion*, vol. 118, p. 102922, 2025, doi: 10.1016/j.inffus.2024.102922.
[24] S. Dey, "Implementing AI red teaming to develop secure responsible AI models," *Computer*, vol. 58, no. 11, pp. 95–99, 2025, doi: 10.1109/MC.2025.3563738.
[25] K. Nanthini, D. Sivabalaselvamani, K. Chitra, P. Gokul, S. KavinKumar, and S. Kishore, "A survey on data augmentation techniques," in *Proc. 7th Int. Conf. Comput. Methodol. Commun. (ICCMC)*, 2023, pp. 913–920, doi: 10.1109/ICCMC56507.2023.10084010.
[26] Y. Wu, B. Zou, and Y. Cao, "Current status and challenges and future trends of deep learning-based intrusion detection models," *J. Imaging*, vol. 10, no. 10, p. 254, 2024, doi: 10.3390/jimaging10100254.