Machine Learning in Offensive and Defensive Malware: A Comparative Analysis

Michael Murray
California Polytechnic State University, San Luis Obispo, California
michael.murray.iv@gmail.com

I. INTRODUCTION

The rapid evolution of machine learning (ML) technologies has significantly reshaped the malware space, offering new opportunities for both attackers and defenders. On the offensive side, machine learning has led to advancements in malware generation and evasive malware. Defensively, ML has important implications for both detection and mitigation of successful attacks. This research will compare offensive ML-driven malware technologies with defensive ones. It will also delve into commercial products currently available for use. The aim of this paper is to answer the following question:

Does the use of machine learning in the malware space have a greater positive or negative impact on the field, considering both offensive and defensive applications?

II. CONTEXT AND HISTORY

Machine learning has revolutionized the field of cybersecurity by enabling automation, pattern recognition, and advanced predictive analytics. Before the introduction of ML, primarily signature-based systems were used for malware detection. These systems searched for known attack patterns to detect malware, which was an effective way to stop basic attacks but could not detect unique or novel methods as accurately [1]. In recent years there has been a significant shift toward ML-based approaches, allowing security systems to detect malicious behavior dynamically rather than relying solely on predefined signatures. This has been accompanied by adversarial ML techniques, which can generate malware capable of bypassing detection mechanisms. Today, there is an ongoing battle between offensive and defensive ML applications, with both attackers and defenders constantly refining their techniques.

III. OFFENSIVE USES OF ML

Attackers can use ML to generate dynamic malware capable of evading detection systems. Various techniques have been employed to achieve these results, some of which are detailed below.

A. Model-Based Reinforcement Learning (MEME)

MEME is an algorithm which aims to combine malware evasion with model extraction [2]. The goal of malware evasion is to make the malware undetectable by a security system, so that it can perform its malicious activities unnoticed. Model extraction is a family of attacks which aims to retrieve a model's parameters so that a "surrogate model" can mimic its behavior. This allows malicious actors to probe the copied model for weaknesses.

There were two goals of this method, which MEME successfully accomplished: to make malware more resistant to detection by security models, and to extract knowledge about the security models themselves. MEME generated adversarial malware that successfully evades antivirus systems in the range of 32-73%. It also produces surrogate models which match the original models' predictions with 97-99% accuracy.

B. MalFox (Conv-GANs)

The MalFox framework explores the use of deep learning to generate adversarial malware. This technique employs convolutional generative adversarial networks (Conv-GANs) for generating malware designed to evade black-box detection models while retaining their malicious functionality [3]. Most of the programs fed into MalFox return with their original functionality intact – over 99% for a dataset of 15144 programs. Additionally, the average rate of

evasion has improved significantly – by about 56.2%. This means the MalFox framework is able to consistently generate code which has a higher chance of evading a threat detection system.

C. FGAM

Fast Gradient Sign Adversarial Malware (FGAM) is an attack technique that changes malware code using gradient sign adjustments [4]. The approach is inspired by adversarial attacks in image classification, where small changes to input data can trick ML models. The study highlights that FGAM can make malware more resilient against detection by introducing subtle modifications that do not alter the core functionality of the malicious software. The success rate of the malware deception model generated by FGAM was increased by ~84% compared with existing methods. This method attains similar results to MalFox – it can consistently generate code that has a higher chance of evasion. FGAM does this for a specific subset of attack types in image classification.

D. Adversarial Attacks on Windows Malware Detection

This study introduces RAMEN, a unifying framework that encapsulates various attacks against machine learning models. It also introduces three novel strategies (Full DOS, Extend, and Shift) which manipulate the DOS header, extend it, and shift the contents of the first section to increase the evasion rate [5]. These attacks can severely harm the performance of commercial antivirus solutions that rely on machine learning.

These methods demonstrate powerful ways that machine learning has been leveraged to improve the performance of adversarial malware. The key changes that ML-based adversarial malware focuses on are:

- Improved evasion rate against threat detection systems.
- Decreasing the number of changes necessary for successful evasion.
- Generation of surrogate models for analyzing security systems.

IV. DEFENSIVE USES OF ML

Cybersecurity professionals have integrated ML into threat detection systems to identify and mitigate malware more effectively. ML enables proactive defense mechanisms that evolve alongside emerging threats. Some of the benefits of ML in malware detection are:

Improved Accuracy: ML has enhanced threat detection by analyzing patterns and behaviors rather than relying solely on signatures.

Adaptability: ML models continuously learn from new data, improving detection of zero-day threats – unfamiliar attacks without precedents

Automation: The use of ML has reduced the need for manual intervention by partially or fully automating malware detection and response.

A large variety of ML models have been tried to improve malware detection capabilities. A study published in *MDPI* evaluated various classification techniques including Decision Trees (DT), Convolutional Neural Networks (CNN), and Support Vector Machines (SVM), for their effectiveness in malware detection [6]. All these methods achieved high detection accuracies – DT with 99%, CNN with 98.76%, and SVM with 96.41%. These metrics surpass the accuracy of traditional signature-based detection methods, demonstrating the value of ML techniques.

The application of deep learning (DL) models has also proved valuable for malware detection. The paper “Deep Learning Models for Detecting Malware Attacks” explores the capabilities of DL algorithms in this space [7]. The study found that DL models are capable of handling large datasets and performing automatic feature extraction, which makes them more effective in detecting diverse malware variants. This means that DL models are more adaptable to zero-day threats. Another study, “ML with

Feature/Platform	SentinelOne Singularity Platform	CrowdStrike Falcon	Microsoft Defender for Endpoint	Trellix Endpoint Security Suite	Bitdefender GravityZone
Overall Rating	4.7 (2,500+ reviews)	4.7 (2,000+ reviews)	4.6 (2,500+ reviews)	4.5 (1,867 reviews)	4.6 (1,200+ reviews)
Ease of Use	4.6	4.5	4.4	4.4	4.5
Malware Detection	Excellent	Excellent	Strong	Excellent	Excellent
Pricing Flexibility	4.5	4.4	4.3	4.5	4.4
Quality of Technical Support	4.4	4.5	4.6	4.4	4.5
Integration Ease	High (API Integration)	High (API Integration)	Moderate	Moderate	High (API Integration)
False Positive Rate	Low	Low	Moderate	Moderate	Low
Training & Onboarding	Comprehensive	Comprehensive	Moderate	Moderate	Comprehensive

Table 1. A comparison of various metrics for leading Malware Detection products [9].

Cuckoo Sandbox", compares the effectiveness of DL techniques with conventional ML methods such as SVM, k-nearest neighbor, and Random Forest [8]. All the tested methods achieved accuracy of over 95%. The Recursive Neural Network (RNN) tied for the highest accuracy with some of the conventional methods, at 99%. DL algorithms are powerful tools for developing advanced malware detection systems, which have high detection accuracy and can adapt to the ever-evolving threat landscape.

V. PRODUCTS IN INDUSTRY

Many leading cybersecurity companies offer threat detection services that employ machine learning techniques similar to those described in the section IV. Following is a comparison of the strengths and weaknesses of five leading products, based on user reviews: SentinelOne Singularity Platform, CrowdStrike Falcon, Microsoft Defender for Endpoint, Trellix Endpoint Security Suite, and Bitdefender GravityZone.

- **SentinelOne Singularity Platform**: Users praise its ease of use, time-to-value, deep visibility into endpoint activity, and ease of deployment. Complaints include some bugs related to VPN misclassification.

- **CrowdStrike Falcon**: Users appreciate its ease of use, excellent time-to-value, and strong endpoint visibility. Complaints include issues with sensor repair and ticket support.

- **Microsoft Defender for Endpoint**: Users like its real-time protection, integration with Microsoft tools, and strong protection, especially for Windows environments. Complaints include KQL limitations and device grouping issues.

- **Trellix Endpoint Security Suite**: Users like its strong malware and ransomware detection capabilities, flexible deployment options (cloud and on-premise), and low performance impact on endpoints. Some dislike its deployment complexity and integration issues, particularly with SIEM solutions.

- **Bitdefender GravityZone**: Users like its endpoint tagging, sandboxing, and granular control. Complaints include performance issues during full scans.

Various metrics for these products have also been compiled in Table 1. It seems that SentinelOne, Crowdstrike, and Bitdefender all have a slight edge on Trellix and Microsoft's offerings, particularly in the Ease of Use, Integration Ease, False Positive Rate, and Training & Onboarding categories. These differences are worth considering for potential customers.

Overall, while each of these industry solutions have their respective pros and cons, they are all well-regarded for their malware detection rates and reliability. Choosing the right solution depends on the specific needs and infrastructure of an organization.

VI. CONCLUSION

The integration of machine learning into the malware space has driven significant advancements in both offensive and defensive strategies. On the offensive side, attackers have leveraged ML techniques to create adversarial malware with advanced capabilities. This malware is more evasive, requires fewer changes to bypass systems, and can generate surrogate models to analyze and exploit black box security systems. Defensively, ML has led to improved malware detection rates and response mechanisms, which are able to adapt to zero-day threats.

Although there have been important developments on both sides of the malware space, I think that the offensive side has benefitted more from ML. The ability for adversarial malware to evade security systems makes it increasingly difficult to detect and may allow attackers to stay one step ahead of cybersecurity professionals. Looking forward, continued research is necessary to mitigate the risks posed by ML-enhanced malware.

REFERENCES

[1] The Evolution of Intrusion Detection & Prevention, SecureWorks. [Online]. Available: https://www.secureworks.com/blog/the-evolution-of-intrusion-detection-prevention.

[2] K. Zhou, et al., "The Power of MEME: Adversarial Malware Creation with Model-Based Reinforcement Learning," *arXiv*, 2023. [Online]. Available: https://arxiv.org/pdf/2308.16562.

[3] Y. Liu, et al., "MalFox: Camouflaged Adversarial Malware Example Generation Based on Conv-GANs Against Black-Box Detectors," *arXiv*, 2020. [Online]. Available: https://arxiv.org/pdf/2011.01509.

[4] Z. Jiang, et al., "FGAM: Fast Adversarial Malware Generation Method Based on Gradient Sign," *arXiv*, 2023. [Online]. Available: https://arxiv.org/pdf/2305.12770.

[5] H. S. Anderson, et al., "Adversarial EXEmples: A Survey and Experimental Evaluation of Practical Attacks on Machine Learning for Windows Malware Detection," *arXiv*, 2020. [Online]. Available: https://arxiv.org/abs/2008.07125.

[6] P. Radanliev, et al., "Malware Analysis and Detection Using Machine Learning Algorithms," *Symmetry*, vol. 14, no. 11, p. 2304, Nov. 2022. [Online]. Available: https://www.mdpi.com/2073-8994/14/11/2304.

[7] Y. Huang, et al., "Deep Learning Models for Detecting Malware Attacks," *arXiv*, 2022. [Online]. Available: https://arxiv.org/pdf/2209.03622.

[8] A. F. Alshmarni and M. A. Alliheedi, "Enhancing Malware Detection by Integrating Machine Learning with Cuckoo Sandbox," *Journal of Information Security and Cybercrimes Research*, vol. 7, no. 1, pp. 85–92, Jun. 2024. [Online]. Available: https://journals.su.edu.sa/jiscr/article/view/1384.

[9] "Compare Bitdefender GravityZone, CrowdStrike Falcon, Microsoft Defender, SentinelOne, and Trellix Endpoint Security," *Gartner*, [Online]. Available: https://www.gartner.com/reviews/market/endpoint-protection-platforms/compare/product/bitdefender-gravityzone-vs-crowdstrike-falcon-vs-microsoft-defender-for-endpoint-vs-sentinelone-singularity-platform-vs-trellix-endpoint-security-suite. [Accessed: Feb. 26, 2025].

FURTHER READING

S. Kumar and R. Kumar, "Machine Learning in Malware Analysis: Current Trends and Future Directions," International Journal of Advanced Computer Science and Applications, vol. 15, no. 1, 2024. [Online]. Available: https://thesai.org/Downloads/Volume15No1/Paper_124-Machine_Learning_in_Malware_Analysis.pdf.

P. Radanliev, et al., "Malware Analysis and Detection Using Machine Learning Algorithms," Symmetry, vol. 14, no. 11, 2022. doi: 10.3390/sym14112304.

D. Ucci, et al., "Adversarial Machine Learning in Malware Detection," National Science Foundation Public Access Repository. [Online]. Available: https://par.nsf.gov/servlets/purl/10053527.

B. Biggio and F. Roli, "Adversarial Machine Learning in Malware Detection: Arms Race between Evasion Attack and Defense," IEEE Transactions on Cybernetics, vol. 48, no. 10, pp. 3158–3170, 2018. doi: 10.1109/TCYB.2018.2800792.

W. Chen, et al., "A Comparison of Adversarial Learning Techniques for Malware Detection," arXiv, 2023. [Online]. Available: https://arxiv.org/pdf/2308.09958.

L. Demetrio, et al., "A Malware Classification Survey on Adversarial Attacks and Defences," arXiv, 2023. [Online]. Available: https://arxiv.org/abs/2312.09636.

Endpoint Security, Trellix. [Online]. Available: https://www.trellix.com/platform/endpoint-security/.

Singularity XDR, SentinelOne. [Online]. Available: https://www.sentinelone.com/resources/singularity-xdr/.

CrowdStrike Platform, CrowdStrike. [Online]. Available: https://www.crowdstrike.com/platform/.

Microsoft Defender for Endpoint, Microsoft Learn. [Online]. Available: https://learn.microsoft.com/en-us/defender-endpoint/microsoft-defender-endpoint.

Welcome to GravityZone, Bitdefender. [Online]. Available: https://www.bitdefender.com/business/support/en/77209-79436-welcome-to-gravityzone.html.