Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

Securing the CyberML Ecosystem: A Comprehensive Review of Adversarial Attack and Defenses Across the Machine Learning Lifecycle

Reena Cheema
Dept. of Computer Science and
Applications Lovely Professional University
Phagwara, India

***Abstract* - With the advancement of Machine Learning (ML), there has been huge development in our daily lives particularly in how we protect our computer network and systems. Not only the use of ML has reached beyond imagination in the areas of understanding the issues like botnet detection, intrusion detection and user authentication but also in making intelligent decisions through malware analysis, anomaly detection, spam filtering and phishing detection [2]. Once implemented, these ML algorithms assist in user behavior detection through human interaction verification and keystrokes dynamics detection [10]. In this paper, we use the term 'CyberML' as an implementation of ML in cyber security. The implementation of ML thus helps to uncover the advancement of adversarial machine learning (AML) to help in building more reliable and trustworthy security systems. This paper does so by discussing the attacks against malware detection like stingray, text perturbation attack, and attacks against network anomaly detection like slack attack, jamming attack and IDSGAN.**

Keywords* - *CyberML, Adversarial Machine Learning (AML), Malware Detection, Network Anomaly Detection, Machine Learning Lifecycle.

I. INTRODUCTION

Machine Learning (ML) impacts our everyday lives because it allows machines to learn how to process data and make intelligent decisions automatically, without needing to be directly programmed for every specific task. There are statistical models, and algorithms that help to act like an intelligent machine who is capable of making decisions without any external coding from the programmer. It has been found that over the years, machine learning has been widely deployed for identification of cyber attacks, and also act as a preventative measure for intrusion detection, malware analysis and botnet detection. These applications belong to different areas like network protection, application security, user behavior, and protection for end points.

In all types of protections, the threats on different levels such as host level, application level are checked for access control along with processing the behavior to detect the fraud and anomalies. On the other hand machine learning has helped to identify human interaction, verification, and key dynamics detection at the next level under the modern cyber security. In this paper, we use the term "CyberML" to describe any use of ML within the cybersecurity domain.

While CyberML has huge potential to improve automated threat detection, it also introduces brand-new vulnerabilities. Attackers can exploit these vulnerabilities by manipulating the data that the ML model relies on, causing the system to make incorrect predictions or fail entirely. The attackers also evolve with the growth of technology that fall under the umbrella of Adversarial Machine Learning (AML). This paper looks closely at how AML impacts two critical areas of cybersecurity: finding malicious software (malware detection) and spotting unusual network traffic (network anomaly detection).

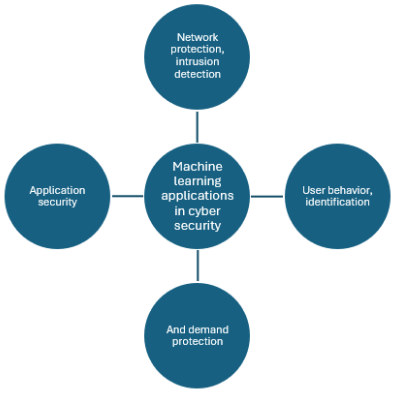

Fig. 1. Visualization of various machine learning applications in cyber security

II. BACKGROUND

The huge rise in the growth of use of machine learning has also given growth to the attackers that use adversarial machine learning (AML) to attack the machine learning models. They target models by modifying the data samples to commit misclassification and image perturbation. The deployment of cyber ML is the core target of the attackers and it is done in every step of the machine learning lifecycle. To understand how an ML model can be attacked, we first need to understand how it is built. The ML lifecycle generally consists of several key stages. For simplicity, we focus on the four main stages where attacks are most likely to happen:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

- Data Gathering and Preparation: Collecting raw data from various sources and cleaning it up so the model can learn from it.
- Train Model: Building the ML model using the prepared data and teaching it to identify patterns or rules.
- Test Model: Giving the model new data it has never seen before to test if it can accurately make predictions in the real world.
- Deployment: Putting the finalized model into a real-world, live system.

The main idea of malicious or compromised machine learning data or model is to degrade the system or lead to the failure of the application. In the national vulnerability database, more than 370 vulnerabilities are reported in the machine learning model development where the models fall prey to these attacks. For example, any kind of modification to one or two pixels of an image in the data processing stage leads to validation problems in the images that provide a stable ground to perform AML in image recognition cases. Thus, there is a strong need to modify the parameters of the data effectively to reduce the chances of attack.

When attackers try to break these systems, they generally use four main strategies:

- Evasion: Tweaking malicious data just enough so the model thinks it is harmless.
- Poisoning: Injecting bad data into the training set so the model learns the wrong rules.
- Model Extraction: Stealing the underlying logic of a model by repeatedly asking it questions.
- Inference: Guessing what sensitive data was used to train the model in the first place.

III. LITERATURE REVIEW

There have been extensive study on the vulnerabilities of machine learning where most survey look only at the threat or focus on the one stage of the lifecycle. The realworld attackers do not consider only one stage instead they identify every possible options from end to end to attack. Moreover, academic research is more concentrated on highly complex attacking methods, but real attackers look for faster implementation of attacks that create a dramatic impact. As an example, the attackers perform attacks on the internal operations and make them time consuming to pretend as if they malfunctioned. And these attacks are more focused on Malwa and anomaly detection across different stages. Network anomaly detection systems are important for the network security and they identify different type of patterns that are created by the threats. Respective of their capability to alter intrusion detection systems, these frameworks face a lot of disruptions [3].

- Generative Adversarial Network (GAN) Exploits: Generative modeling is one of the most effective tool for bypassing the network defenses. The framework like IDSGAN make use of traffic features that look normal but target the feature extraction mechanism to create issues [14]. Conversely, more studies has identified that category one attacks without causing any kind of disruption in the normal network throughput, which is more difficult to be spotted [4].
- Protocol and Structural Obfuscation: Attackers make a huge change in the connection Properties to create illusions for the classifiers that consider statistical values as their primary mode for evaluation. By making any kind of changes in the properties of the network, techniques like TCP Obfuscation create illusions, where the malicious data and code is transferred from one position to another without triggering any alerts [6].
- Physical and Data Link Layer Disruption: Adversarial ML techniques impact the existing model, not only through the software, but through the communication. In the radio networks, machine learning rely on scan frequencies, channels, and route data through the best parts to save power and computation. However, the category to adverse framework can cause problems in such roots and bigger negative impact wherever wireless communication is involved at the radio layer [5].
- Slack Space and Byte-Level Manipulation: Models that process white based convolutional neural networks often malfunction in the changes that preserve the structure. Slack FGM attack targets the regions, where padding or space is given between the payloads or any files [1]. By making changes to the repetitive content in the original code, the attacks create a dramatic change in the signature, allowing the malicious information to pass through [16].

IV. STAGES OF ADVERSARIAL MACHINE LEARNING

Designing robust malware and anomaly detectors is incredibly challenging because attackers constantly mutate their methods, making the ML detectors highly unstable. Network anomaly detection usually monitors network traffic for unexpected patterns. However, if an attacker carefully manipulates the data at any point in the ML lifecycle, these detectors will fail. Here is how weaknesses are exposed at each stage:

4.1 Data Gathering and Preparation Stage

The quality of a model's predictions depends entirely on the quality of its training data.

- Data Poisoning: The most common attack here is data poisoning, where the attacker changes the data or its labels to slowly ruin the learning process. Because security systems often pull data from untrusted sources (like the open web), attackers can easily sneak in corrupted data [11].
- GAN-Based Attacks: For network anomaly detection, attackers use Generative Adversarial Networks (GANs)—like the "IDSGAN" framework —to generate fake, malicious network traffic that looks perfectly normal to a detector [7].

4.2 Train Model Stage

During training, an attacker's goal is usually to observe the process or actively change the data distribution.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

- Feature Manipulation: Attackers can inject malicious data to shift the ML model's decision boundaries. For instance, by modifying just 20% of the input features, attackers have successfully increased misclassification rates to 98% in some models.
- Data Leakage: Sometimes, the training data itself leaks sensitive information by accidentally "memorizing" specific inputs, leading to privacy risks.

4.3 Test Model Stage

The testing stage is heavily targeted by evasion attacks. Attackers analyze the model to figure out exactly what minor changes will trick it.

- Malware Evasion: Attackers can generate an attack simply by modifying a single API call (a line of source code) in a malware file [6]. This is cheap, easy, and can produce an infinite number of undetectable malware variants.
- Network Anomaly Evasion: Attackers modify network connection properties to hide their TCP communication, successfully escaping many types of intrusion detection classifiers.

4.4 Deployment Stage

After a model goes live into production, the nature of the threat shifts. The attacker tried to impact the deployed system directly. This is done in two ways:

- Model Extraction and Evasion: where the attacker identify the spots making it undetectable as the system is flooded with redundant inputs. In this type of attack, there is a clone generation of the model itself, making it very difficult to stop the attack.
- The Single-Model Vulnerability: In this case, there is a negative impact due to over line on single algorithmic framework. For example, the random forest algorithms are highly brittle where the botnet detection system take huge support of it, making it easily targeted by the adversarial machine learning models to impact them negatively.

V. FUTURE RESEARCH DIRECTIONS

Due to the high-speed evolution of the cyber security issues, the machine learning models are failing against the attacks. To identify the gaps and build more strong models that identify and stop these attacks there is a need to work on the following factors:

- Overcoming Data Bottlenecks: it is very difficult to get good data for testing the cyber security tools because the companies keep their data locked and away from the reach of public. There is a need to get this data to develop smarter tools.
- Minimizing False Positives: There is also a need to minimize the fake alarms that made lead to reduce the confidence level on the security system. This is a negative point because in reality there may be ignorance of actual alarm due to these false alarms.
- Real-Time, Proactive Defenses: There is also a need of live and instant defense system that protect the user data and security programs whenever something relevant is detected that may be a potential threat [9].
- Hardware-Assisted Detection: To identify the modern attack and virus activities. The use of software is a slow option, and there is a need to implement computer hardware that act as a deep level, malicious activities scanners [8].

VI. CONCLUSION

The machine learning has become one of the most important asset in identifying the threats like malware, classification or network anomaly detection in the modern cyber security system due to its ability to understand the situation effectively and act on the same time. However, the modernization and constant evolution of the attackers they utilize adversarial machine learning to create a greater negative impact. By examining the entire lifecycle of a machine learning model, we can exactly identify why, and where these systems easily breakdown through making changes into the data, inputs and processing techniques or manipulating API end points. There is also a huge need to incorporate hardware that act in the real time to achieve substantial benefits. Finally, the use of machine learning can help to identify the data leakage and feature manipulation that can be protected to avoid poisoning of the models to acquire more stable models against the attacks.

REFERANCES

[1] Mahesh B. Machine learning algorithms-a review. Int J Sci Res 2020;9:381–6.

[2] Patra C, Giri D, Kundu B, Maitra T, Wazid M. Rhetorical structure theory-based machine intelligence-driven deceptive phishing attack detection scheme. J Inf Secur Appl 2025;94:104184.

[3] Ahmad Z, Shahid Khan A, Wai Shiang C, Abdullah J, Ahmad F. Network intrusion detection system: A systematic study of machine learning and deep learning approaches. Trans Emerg Telecommun Technol 2021;32(1):e4150.

[4] Jadidi Z, Foo E, Hussain M, Fidge C. Automated detection-in-depth in industrial control systems. Int J Adv Manuf Technol 2022;118(7):2467–79.

[5] Yadav G, Paul K. Global monitor using SpatioTemporally correlated local monitors. In: 2021 IEEE 20th international symposium on network computing and applications. NCA, 2021, p. 1–10.

[6] Chowdhury M, Rahman A, Islam R. Malware analysis and detection using data mining and machine learning classification. In: Abawajy J, Choo K-KR, Islam R, editors. International conference on applications and techniques in cyber security and intelligence. Cham: Springer International Publishing; 2018, p. 266–74.

[7] Kim J-Y, Bu S-J, Cho S-B. Malware detection using deep transferred generative adversarial networks. In: Liu D, Xie S, Li Y, Zhao D, El-Alfy E-SM, editors. Neural information processing. Cham: Springer International Publishing; 2017, p. 556–64.

[8] Rathore H, Sasan A, Sahay SK, Sewak M. Defending malware detection models against evasion based adversarial attacks. Pattern Recognit Lett 2022;164:119–25.

[9] Khan RU, Zhang X, Kumar R, Sharif A, Golilarz NA, Alazab M. An adaptive multi-layer botnet detection technique using machine learning classifiers. Appl Sci 2019;9(11):2375.

[10] Raul N, Shankarmani R, Joshi P. A comprehensive review of keystroke dynamics-based authentication mechanism. In: Khanna A, Gupta D, Bhattacharyya S, Snasel V, Platos J, Hassanien AE, editors. International conference on innovative computing and communications. Singapore: Springer Singapore; 2020, p. 149–62.

[11] Pasupathi S, Kumar R, Pavithra L. Proactive DDoS detection: Integrating packet marking, traffic analysis, and machine learning for enhanced network security. Clust Comput 2025;28(3):210.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

[12] Paudice A, Muñoz-González L, Lupu EC. Label extraction/sanitization against label flipping poisoning attacks. In: Joint European conference on machine learning and knowledge discovery in databases. Springer; 2018, p. 5–15.

[13] Lin Z, Shi Y, Xue Z. Idsgan: Generative adversarial networks for attack generation against intrusion detection. 2018, arXiv preprint arXiv:1809.02077.

[14] Apruzzese G, Colajanni M. Evading botnet detectors based on flows and random forest with adversarial samples. In: 2018 IEEE 17th international symposium on network computing and applications. NCA, IEEE; 2018, p. 1–8.

[15] Apruzzese G, Colajanni M, Marchetti M. Evaluating the effectiveness of adversarial attacks against botnet detectors. In: 2019 IEEE 18th international symposium on network computing and

[16] Usama M, Asim M, Latif S, Qadir J, et al. Generative adversarial networks for launching and thwarting adversarial attacks on network intrusion detection systems. In: 2019 15th international wireless communications & mobile computing conference. IWCMC, IEEE; 2019, p. 78–83.

[17] Raff E, Barker J, Sylvester J, Brandon R, Catanzaro B, Nicholas CK. Malware detection by eating a whole exe. In: Workshops at the thirty-second AAAI conference on artificial intelligence. 2018.

[18] Suciu O, Coull SE, Johns J. Exploring adversarial examples in malware detection. In: 2019 IEEE security and privacy workshops. SPW, IEEE; 2019, p. 8–14.

[19] Demme J, Maycock M, Schmitz J, Tang A, Waksman A, Sethumadhavan S, et al. On the feasibility of online malware detection with performance counters. ACM SIGARCH Comput Archit News 2013;41(3):559–70.