International Journal of Computer Science Trends and Technology (IJCST) – Volume 8 Issue 4, Jul-Aug 2020

RESEARCH ARTICLE OPEN ACCESS

Adversarial Attacks on Machine Learning Cybersecurity Defenses in Cloud Systems

Alex Mathew
Department of Cyber Security Bethany College
USA

ABSTRACT
Adversarial attacks are implemented when an attacker introduces maliciously designed inputs to deceive or corrupt machine-learning (ML) models. In this study, the researcher examined the mechanisms used by two primary adversarial ML attacks, namely poisoning attacks, and evasion attacks, to undermine the cybersecurity defenses in the cloud system. A meta-analysis of previous articles was used as the methodology, and a block diagram is provided to give the search strategy and selection approach. The findings revealed that attackers' knowledge about the white-box and black-box models determine the success rate of adversarial attacks.
Keywords: —Adversarial attacks, machine-learning, cloud system, poisoning attacks, evasion attacks.

I. INTRODUCTION

Cloud computing has transformed every aspect of life because of advanced technologies. However, high-tech technologies have also led to a cybersecurity threat that requires strong defenses to avert attacks from cybercriminals. Cybercriminals use adversarial attack models to infiltrate the cloud system; therefore, learning about the various attacks and their mechanisms can build strong cybersecurity defenses. On the other hand, machine learning (ML) models are essential for various decision support systems with accurate, efficient, and faster output; however, the applications of ML in the cloud system encounter inconsistent danger of active adversarial attacks [4]. Experts explain that the challenges of ML applications in cloud systems, such as intrusion detection systems (IDS), malware detection, and spam filtering are adversarial [4]. Recent studies have shown that deep learning-oriented systems are susceptible to attacks by adversarial examples as technology advances [13]. The current research examines the adversarial attacks on ML cybersecurity defenses, concentrating on the Fast Gradient Sign Method (FGSM), evasion, and poisoning attacks.

II. BACKGROUND

Studies have demonstrated that the quantities of information attackers have on a particular ML system determine the sophistication of an attack and the effort needed to counter the adversaries [3]. For example, in a white-box framework, the attackers have Knowledge about the classification algorithms, outputs, and inputs. In a black-box model, the attackers only know the classification, making it the hardest to exploit, while in a gray-box model, they see the performance against inputs [3]. Fig. 3 provides an outlook of the models and the attack process.

Fig. 1: Adversarial Attack Process

The information about the three models forms a foundational concept for this paper. It is likely to comprehend the proliferation and application of ML oriented IDS having the background information. Use of ML-based IDS has facilitated a more efficiency and flexibility in the automated discovery of cyber invasions in cloud systems [1, 11]. Nonetheless, the onset of IDSs also generated an extra attack dimension, commonly referred to as adversarial ML (AML). The AML attacks are likely to cause severe consequences by bypassing the available IDS, which leads to financial loss, death, or infrastructure damages due to delayed detection. The attackers exploit the blind spots during the AML process and generate adversarial examples or samples from a more sophisticated database [1, 3]. The study builds on two types of AML attacks on cloud systems.

Poisoning attacks and evasion attacks are the two primary types of AML attacks [3]. These attacks depend on the aspect and phase of the targeted ML model, and the influence exploit of an attack affects the classifier's decision [1]. The attacks could also be causative, which happens during the training stage (poison attacks) or exploratory that occurs during testing stages (evasion attacks) [15]. The adversarial examples could cause a misclassification either through the target or indiscriminate approach [1, 15]. There are various methods for

adversarial sample generation; however, the current paper examined only one.

The Fast Gradient Sign Method (FGSM) is the preferred method for this study. The approach targets every feature of the input data to add specific perturbation amount [1, 12]. Equation (1) provides the formula for computing the perturbation noise based on the gradient of the cost function J based on input data [1]. Let θ depicts model parameters, x the inputs, y input data labels, € represents the value of applied noise, and J (θ,x,y) cost functions applicable to exploit the targeted neural network[1].

$$x^* = x + € \operatorname{sign}(\nabla x J(\theta, x, y)) \quad (1)$$

In the FGSM approach, adversarial examples are generated by determining the optimal direction of positive change in the loss [4, 12]. It is a faster method because only one-step gradient update is computed as the sign gradient per level [4, 12].

III. PROPOSED METHODOLOGY

The methodology used in this study is a meta-analysis of previously conducted experiments with reputable outcomes. An adversarial attack is a topic that has been broadly studied to understand invasion techniques and methods under cybersecurity defenses and cloud systems [4]. The summary of the conceptual findings, which include equations, algorithms, and flow charts, has been implemented using viable selection criteria and search strategy.

IV. BLOCK DIAGRAM

The search strategy is presented in a block diagram of six steps (Fig. 2). The initial step is an overview of available sources based on the selection criteria of a recent publication (2016 to 2020) then extraction of reputable articles with viable sources. The key terms were the search elements, which led to the present a comprehensible report [2].

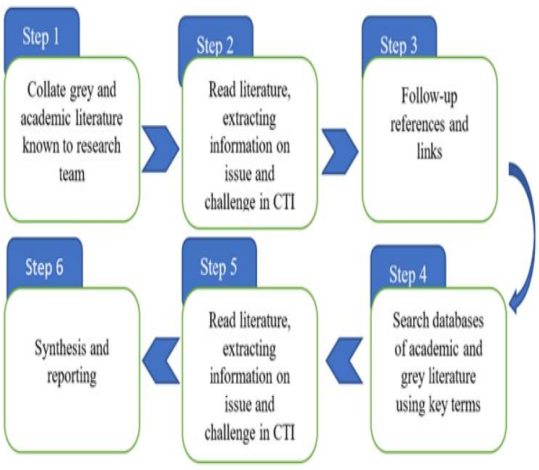

Fig. 2 Methodology Block

V. ALGORITHM

Understanding the AML attacks requires Knowledge about the ML techniques based on various algorithms by tasks. Fig. 3 illustrates that the classification, rule learning, regression, clustering, dimensionality reduction, and generating modeling are critical aspects of algorithms to consider [4, 6]. The target model's training is required for better outcomes.

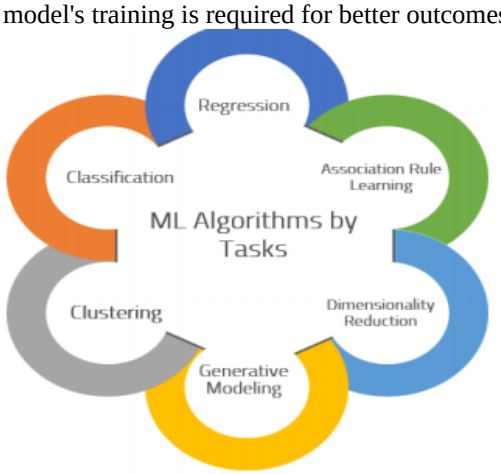

Fig. 3: Machine Learning Tasks

Algorithm 1 outlines the general estimated gradient-ascent algorithm applicable to resolve both poisoning and evasion attacks [9].

Algorithm 1 Gradient-based Evasion and Poisoning Attacks

Input: $\mathbf{x}, y$: the input sample and its label; $\mathcal{A}(\mathbf{x}, y, \kappa)$: the attacker's objective; $\kappa = (\mathcal{D}, \mathcal{X}, f, \mathbf{w})$: the attacker's knowledge parameter vector; $\Phi(\mathbf{x})$: the feasible set of manipulations that can be made on $\mathbf{x}$; $t > 0$: a small number.

Output: $\mathbf{x}'$: the adversarial example.

1: Initialize the attack sample: $\mathbf{x}' \leftarrow \mathbf{x}$
2: **repeat**
3: Store attack from previous iteration: $\mathbf{x} \leftarrow \mathbf{x}'$
4: Update step: $\mathbf{x}' \leftarrow \Pi_\Phi(\mathbf{x} + \eta \nabla_{\mathbf{x}} \mathcal{A}(\mathbf{x}, y, \kappa))$, where the step size η is chosen with line search (bisection method), and Π_Φ ensures projection on the feasible domain Φ.
5: **until** $|\mathcal{A}(\mathbf{x}', y, \kappa) - \mathcal{A}(\mathbf{x}, y, \kappa)| \leq t$
6: **return** $\mathbf{x}'$

Fig. 4 Gradient-based Evasion and Poisoning Attacks [9]

VI. FLOW CHART

The flow charts below present the samples of AML attacks, evasion, and poisoning attacks, respectively. Figure 5 illustrates a flow chart for an exploratory attack where attackers confuse ML model decisions after learning the target information.

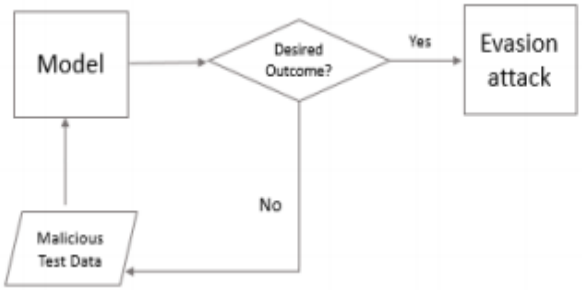

Fig. 5: Evasion Attack [4]

Figure 6 presents a flowchart of poisoning attacks, which entails adversarial manipulation of the guidance information before training to induce an erroneous forecast from the ML model [4].

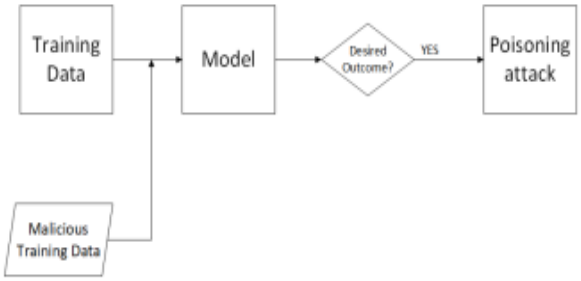

Fig. 6: Poisoning Attacks [4]

VII. RESULT ANALYSIS

The results from various studies and sources illustrate that the adversarial attacks of ML cybersecurity defenses are poisoning attacks and evasion attacks. Poisoning attacks are frequent in online instructional frameworks and learn the input data's vulnerability as they come. The cybercriminal using this method offers AML samples that change the decision predictions or boundaries for their benefits [4]. For example, the first graph in Fig. 7 depicts that the cybersecurity model has learned a vivid decision limit or prediction between malicious (red) and benign (blue) samples, while the second graph has an adversary input represented by the dots. The confusion in the second graph leads to the classification of malicious examples as benign and facilitate attacks.

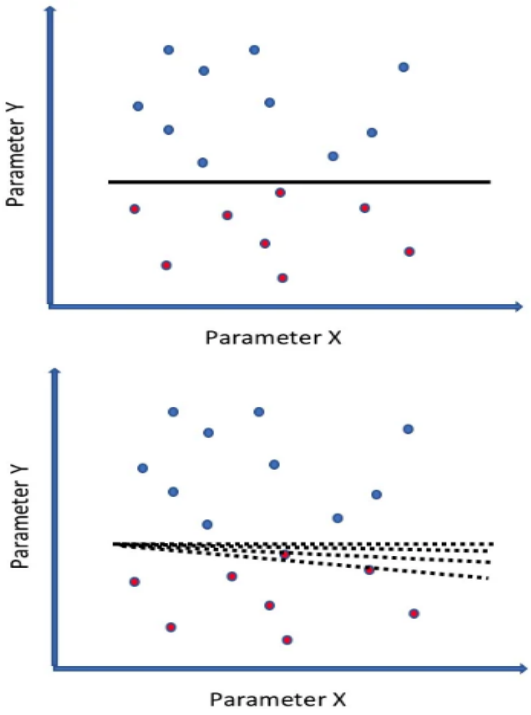

Fig. 7: Poisoning Attacks

The poisoning attacks exploit the cyber system by corrupting the decision-making and misclassify samples. It forces an irregularity detection algorithm to facilitate an attack use beyond the standard data instances. Additional adversarial examples added by the attacker confuse the system and promote an attack. Initiated poisoning attacks exploit the online blind spots for anomaly detection and adversarial noise [5, 6]. The IDS is acclimated to confusing samples and fails or delays to detect the large number of adversarial examples resulting in uppermost degradation in categorization accuracy [5].

Poisoning attacks: Let (X, Y, μ, H, C) depicts a classification with learning algorithm L. While A for (L, X, Y, μ, H, C) is a poisoning adversary exploiting the input of a training set $T \leftarrow (\mu, c(\mu))m$ and then the outputs a transformed training sample $T\ 0 = A(T)$ of similar weight [4, 8]. The T values could also be interpreted as the m vectors, while hamming distance (HD) has vectors of m coordinates [8].

The following properties of A determine the occurrence of an attack for any $c \in C$ [8].

• "A plausible, if $y = c(x)$ for all $(x, y) \in T\ 0$.

• A tampering budget $b \in [m]$ if for all $T \leftarrow (\mu, c(\mu))m$, $T\ 0 \leftarrow A(T)$, then have $HD(T\ 0 , T) \leq b$.

• A mean tampering budget b, the outcome: $E/_{T \leftarrow (\mu, c(\mu))}{}^{m, T\ 0 \leftarrow A(T)} [HD(T\ 0, T))] \leq b$." [8].

The result shows that the poisoning attack is influenced by the capability of the adversarial examples in confusing the system between the malicious and benign samples of input data.

Fig. 8 presents a result of evasion attack, in which the attacker causes the system to misclassify an AML example. The graph

shows that the IDS determines if a given sample is normal or an intrusion of traffic based on the weighted calculation of parameters [3].

Fig. 8: Intrusion Detection System Occurs [3]

In a white-box model, Fig. 9 shows that the attacker will recognize that parameter B is responsible for classifying intrusion as usual; therefore, he or she will increase the parameter's weight to attain the invasion goal [3].

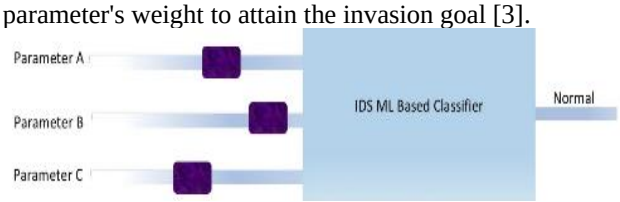

Fig. 9: Normal Setting [3]

Evasion attacks avoid the ML system model by introducing an adversarial example to misclassifies based on a test time procedure. For example, the objective is to determine a sample x' so that the gap between the target malicious sample x0 is reduced.

$$x' = \arg\min g(x) \text{ s.t. } d(x, x0) \leq d_{max} \quad (2)$$

The outcome depicts that the effectiveness of avoiding the ML system IDS rate is by introducing adversarial examples to manipulate header systems [5]. In addition, the threshold for classifying the malicious or benign samples depends on the security of the learning model [14]. FGSM algorithm is cited as a method to compute insertion, removal, and modification of text [5, 14]. Therefore, it is possible to prove the existence of evasion attacks by determining the robustness or risk of a cloud system. The equation and computation process could determine the adversarial risk and robustness.

Let (X, Y, μ, C, H, d) depicts an ideal classification problem.

For c ∈ C, and h ∈ H

Let E = {x ∈ X | h(x) 6= c(x)} denotes error area of h to c. [8]

Adversarial risk: For b ∈ R+, error-area and adversarial risk under b-perturbation is

$$Risk_b(h, c) = Pr\ x \leftarrow \mu\ [\exists\ x\ 0 \in Ball_b(x) \cap E] = \mu(\xi_b).$$

b = budget of adversary that perturbs x into x0 while b=0 then standard risk is

$$Risk\ (h, c) = Risk0\ (h, c) = \mu\ (E)$$ [8].

The results have shown that adversarial attack risks are likely to be established considering the attacker's Knowledge, attack falsification, attack frequency, attack goals, and attack timing. Attacker's Knowledge about the white box, gray box, and black-box attacks with targeted and reliability goals make an attacker successful [4, 6]. Attack transferability was also considered based on the input gradients considering loss function, and the gradient alignments [9, 10]. The adversarial attacks on ML models continue to be a challenge that requires more learning.

VIII. CONCLUSION

The study analyzed the adversarial attacks for cloud systems. The study findings have shown that the primary AML attacks are evasion and poisoning attacks. These attacks occur because cybercriminals can exploit blind spots on input, outputs, or classification of samples in a system. Attackers' Knowledge about the black-box, white-box, and gray-box models was critical to adversarial attacks' success. Poisoning attacks are frequent in online learning platforms, and their success relies on the modification of samples to corrupt the system's decision-making. In contrast, the evasion attacks evade the ML by introducing an adversarial example, which misclassifies the samples to confuse the network between the malicious and benign data.

REFERENCES

[1] Anthi, E, Williams, L, Rhode, M, Burnap, P, and Wedgbury, A, 2020, "Adversarial attacks on machine learning cybersecurity defences in industrial control systems," *ArXiv.org e-Print archive*,1, pp. 1-9.

[2] Abu, S , Selamat, SR, Ariffin, A, and Yusof, R, 2018,"Cyber threat intelligence – issue and challenges," *Indonesian Journal of Electrical Engineering and Computer Science*, 10(1), pp. 371-379 DOI: 10.11591/ijeecs.v10.i1

[3] Khan, M 2018, "Adversarial machine learning attacks in the age of AI?" *Security Intelligence*, [Online] Available at: https://securityintelligence.com/how-can-companies-defend-against-adversarial-machine-learning-attacks-in-the-age-of-ai/ (accessed 1 July 2020).

[4] Ibitoye, O, Abou-Khamis, R, Matrawy, A, and M. Omair Shafiq, MO, 2019, "The Threat of adversarial attacks on machine learning in network security - a survey." *ArXiv:1911.02621v1*, pp. 1-14.

[5] Duddu, V, 2018, "A survey of adversarial machine learning in cyber warfare." *Defence Science Journal*, 68(4), pp. 356-366, DOI : 10.14429/dsj.68.12731

[6] Shokri, R, Stronati, M, Song, C, and Shmatikov, V, 2017, "Membership inference attacks against machine learning models." *In IEEE Symposium on Security and Privacy (S&P) -- Oakland*, pp. 3-18. doi: 10.1109/SP.2017.41

[7] Abraham, T, de Vel, O, and Montague, P, 2018, "Adversarial machine learning for cyber-security: NGTF project scoping study. *Cyber and Electronic Warfare Division*, pp. 1-44.

[8] Mahloujifar, S, Diochnos, ID, and Mahmoody, M 2019, the curse of concentration in robust learning: evasion and poisoning attacks from concentration of measure. *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, pp. 4536-4543.

[9] Demontis, A, Melis, M, Pintor, M, Jagielski, M, Biggio, B, Oprea, A, Nita-Rotaru, C, and Roli, F 2019, "Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks." *ArXiv:1809.02861v4*, pp. 1-19.

[10] Tramèr, F, Papernot, N, Goodfellow, I, Boneh, D., and McDaniel, P 2017, "The space of transferable adversarial examples." *ArXiv:1704.03453v2*, pp. 1-15.

[11] Li, J, Zhao, Z, Li, R, Zhang, H, and Zhang, T 2018, "AI-based two-stage intrusion detection for software defined IoT networks." *IEEE Internet of Things Journal*, pp. 1–13. doi:10.1109/jiot.2018.2883344

[12] Yuan, X, He, P, Zhu, Q, and Xiaolin Li, X, 2018, "Adversarial examples: Attacks and defenses for deep learning." *ArXiv:1712.07107v3*, pp. 1-20.

[13] Zhang, J, Zhang, B, and Zhang, B 2019, "Defending adversarial attacks on cloud-aided automatic speech recognition systems." Workshop Presentation SCC'19, pp. 23-31.

[14] Xu, W, Qi, Y, and Evans, D, 2016, "Automatically evading classifiers: A case study on PDF malware classifiers." *In Network and Distributed System Security Symposium 2016 (NDSS)*, pp. 1-15.

[15] Shi, Y, and Sagduyu, EY, 2017, "Evasion and causative attacks with adversarial deep learning." *Conference Paper*, pp. 1-6.