DOI: https://doi.org/10.5281/zenodo.16944337

Attack and defense techniques for adversarial machine learning

Técnicas de ataque y defensa para el aprendizaje automático antagónico

Héctor Xavier Limón Riaño[1] *y Luis Fernando Rodríguez Hernández*[2]

Resumen: El panorama de la ciberseguridad ha evolucionado rápidamente, presentando riesgos para todos los sistemas informáticos. La inteligencia artificial, en particular el aprendizaje automático, ha surgido para mejorar la seguridad preservando la privacidad de los datos, detectando anomalías y malware, generando confianza y abordando los desafíos de la ciberseguridad. Sin embargo, los adversarios pueden explotar estas técnicas, lo que ha dado lugar al desarrollo del aprendizaje automático adversarial. Nuestro artículo analiza el estado actual del aprendizaje automático adversarial mediante una revisión de 68 estudios realizados entre 2016 y 2023, describiendo las técnicas de ataque y defensa, los desafíos y las consideraciones. Este estudio busca apoyar a los investigadores en la mejora de las medidas de seguridad basadas en IA y el fomento de avances en este campo para lograr soluciones más robustas.

Palabras clave: machine learning, systematic literature review, adversarial machine learning, cybersecurity

Abstract: The cybersecurity landscape has rapidly evolved, posing risks to all computer systems. Artificial Intelligence, particularly Machine Learning, has emerged to enhance security by preserving data privacy, detecting anomalies and malware, establishing trust, and tackling cybersecurity challenges. However, adversaries can exploit these techniques, leading to the development of Adversarial Machine Learning. Our paper analyzes the current state of Adversarial Machine Learning through a review of 68 studies from 2016 to 2023, outlining attack and defense techniques, challenges, and considerations. This study aims to support researchers in enhancing AI-based security measures and fostering advancements in the field for more robust solutions.

Keywords: machine learning, systematic literature review, adversarial machine learning, cybersecurity

1 Doctor en Inteligencia Artificial. Universidad Veracruzana. hlimon@uv.mx
2 Licenciado en Redes y Servicios de Cómputo. Universidad Veracruzana. fernandocontacto236@gmail.com

Introduction

In the age of cloud computing, security challenges like data privacy and system vulnerabilities are critical concerns (Aljumah & Ahanger, 2020). Artificial Intelligence, especially Machine Learning (ML), offers promising solutions for boosting security by detecting threats and analyzing anomalies (Badiger & Shyam, 2023). Despite advancements in ML security techniques for cloud computing, adversaries constantly evolve tactics to exploit these safeguards. The emerging field of Adversarial Machine Learning (AML) focuses on studying such threats and defenses against them. Major tech companies are investing in safeguarding ML systems against AML threats. AML traces back to the early 2000s when statistical (Frederickson, Moore, Dawson, & Polikar, 2018). classifiers identified spam. This paper aims to explore AML vulnerabilities, attack techniques, mitigation strategies, and evaluation metrics through a systematic literature review and thematic synthesis. The study provides a foundational understanding of AML for securing ML-based systems, discussing methods, results, challenges, and future research.

Research method

To examine AML challenges and considerations from attack and defense angles, we conducted a systematic literature review following the methodology outlined by Kitchenham et al. (2015), originally designed for Software Engineering but adaptable across computer science domains. We supplemented this approach with additional methods, including search, selection, snowballing, data extraction, and synthesis processes.

Search process

To conduct the literature review, we started by defining the following research questions:
RQ1. What are the prevailing attack techniques in AML?
RQ2. What are the methods and strategies to mitigate threats in AML?
RQ3. What are the key considerations and challenges involved in implementing AML?

To gather primary studies, our primary approach was automatic search using tailored search strings. We refined these strings by applying a Quasi-Gold standard method (H. Zhang, Babar, & Tell, 2011) and manually selecting studies that addressed our research questions to assess search performance. The final search string selected was:

- ("adversarial machine learning" OR "model poisoning") AND
- ("challenges" OR "opportunities" OR "issues" OR "problems") AND
- ("security" OR "security violation" OR "attack" OR "exploit" OR
- "filtration" OR "exfiltration" OR "defending" OR "mitigation")

Selection process

In order to select primary studies, we established specific criteria to include or exclude studies. To ensure the inclusion of up-to-date and relevant research, our study encompasses literature from 2016 onwards. By focusing on recent studies, we aim to capture the state-of-the-art advancements in the field. The inclusion criteria are as follows:

- IC-1: The study is in English
- IC-2: The year of the study is between 2016 and 2023
- IC-3: The title and abstract suggest that the study answers at least one research question
- IC-4: The full text of the study answers at least one research question

On the other hand, we established the exclusion criteria as follows:

- EC-1: The study is a presentation, book chapter, or opinion
- EC-2: The study is a duplicate from another database
- EC-3: If a study has been updated, we keep the most recent version

The selection process comprised three stages:

- Stage 1: We applied IC-1, IC-2, and EC-1 filters.
- Stage 2: We analyzed the titles and abstracts of the studies and applied the IC-3, and EC-2 filters.
- Stage 3: We read and analyzed the full text of the articles and applied the IC-4, and EC-3 filters.

Snowballing search process

After obtaining primary studies through automatic search, we expanded our collection via a snowballing process following Wohlin's guidelines (2014). This involved manual searches through references (backward snowballing) and citations (forward snowballing), with one iteration of backward snowballing. Each candidate identified underwent our selection process outlined in the previous section. The final selection yielded 68 primary studies.

Data extraction process

The data extraction process involves extracting pertinent information from each primary study that we selected. To facilitate this task, we employed an extraction template, which includes details such as study ID, title, source URL, authors, year, database source, keywords, references, and for each research question, the study’s corresponding answer.

Synthesis process

To effectively organize the extracted information, we employed a thematic synthesis process, as outlined by (Cruzes & Dyba, 2011). This method facilitated the identification, analysis, and reporting of patterns, or themes, within the collected data from the studies. The steps followed in this process are depicted in Figure 1. The outcome of this process is a thematic map that highlights the most significant topics in the literature concerning our research questions. Our thematic map is depicted in Figure 2.

Figure 1. Thematic synthesis process followed, adapted from (Cruzes & Dyba, 2011)

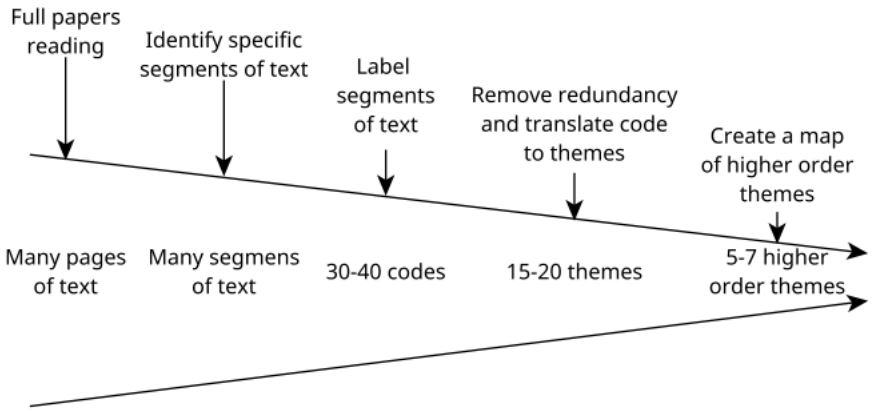
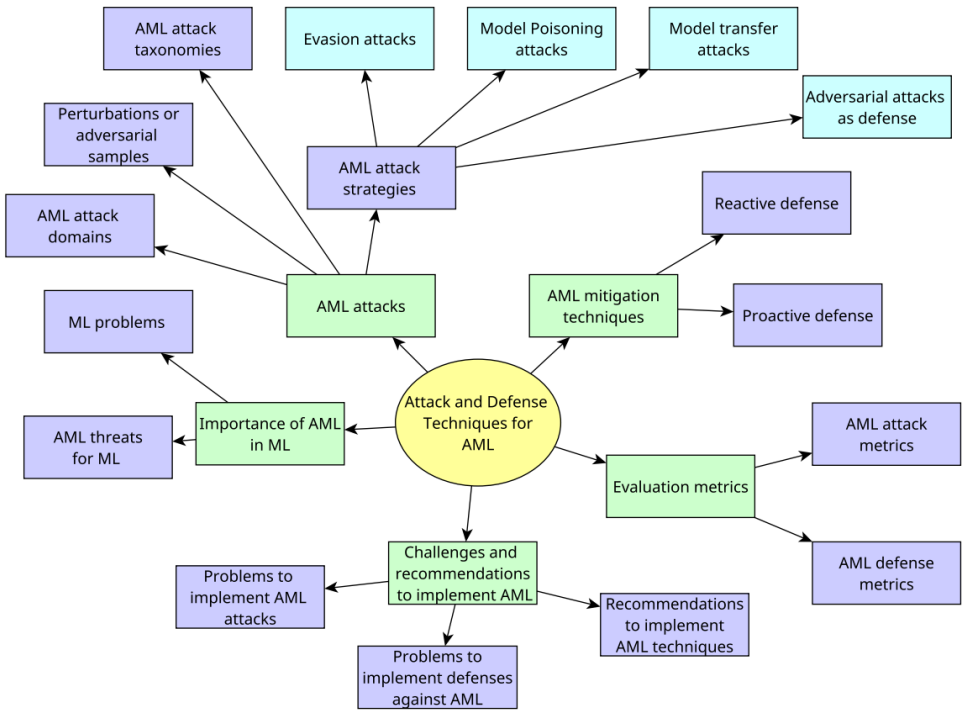


Figure 2. Thematic map. Different colors indicate topic level



Results

In this section, we present the results obtained from our systematic literature review. We begin by discussing the study demographics, including the number of studies found per year and the distribution of studies across different research questions. Following this, we will delve into the complete findings, which are based on the thematic map shown in Figure 2.

Study demographics

Figure 3 depicts an upward trend in studies pertaining to AML, particularly observing a significant surge since 2020. This trend illustrates the growing interest and recognition of AML's significance in data security, emphasizing its relevance and applicability across different domains. In Figure 4, we present the distribution of studies that address each of the research questions discussed in section 3.1. The results indicate a distinct inclination towards research focusing on AML attacks, whereas there is comparatively less emphasis on mitigation techniques for AML threats. Additionally, the challenges and considerations in implementing AML remains largely unexplored.

Figure 3. Primary studies per year

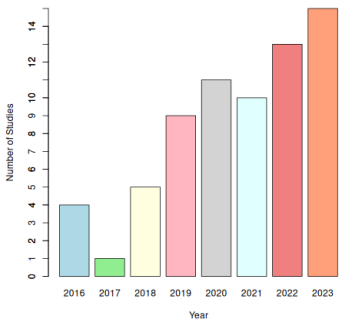
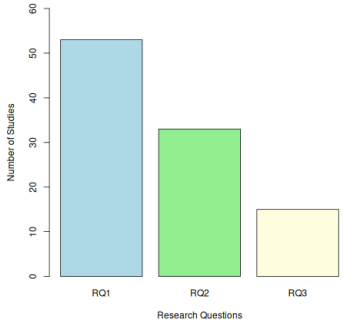


Figure 4. Number of studies by research question



ML problems

ML, widely used in various fields, faces security threats from malicious actors aiming to compromise model integrity. Attackers exploit vulnerabilities in ML systems for information theft, sabotage, or financial gains. Securing ML models remains challenging as attackers and defenders adapt their methods. With the growing use of ML, users and professionals must prioritize system protection (Mehta et al., 2022; Wilhjelm & Younis, 2020; Anthi et al., 2021; U. Verma et al., 2022; Usama et al., 2020).

AML threats for ML

AML poses a serious threat to ML applications by targeting model adaptability, security, and reliability. Adversarial perturbations, a common AML tactic, introduce minimal

changes to input data to mislead models and yield incorrect outputs, compromising performance and user trust. Understanding and mitigating such attacks are crucial for developing secure and reliable intelligent systems across applications (Ayub et al., 2020; Lagesse et al., 2016; Tang et al., 2019; Usama et al., 2019; Khalid et al., 2019; Guesmi et al., 2022).

AML Attacks

AML attacks vary, with tactics like input manipulation and loss function alteration compromising model performance. The challenge lies in distinguishing normal from malicious data, making attacks persistent and demanding evolving mitigation strategies. Understanding these threats is vital for proactive defense, leading to cost savings and improved system security (Lin & Biggio, 2021; Guihai & Sikdar, 2021; Mehta et al., 2022; Hao & Tao, 2022).

AML attacks on image processing

AML research has primarily focused on computer vision, particularly image processing and neural networks, with applications in biometrics, watermarking, and anomaly detection. Adversarial attacks aim to deceive models, especially in image classification where Deep Learning is highly effective but susceptible to misclassification due to adversarial examples. These attacks involve subtle image manipulations or introducing imperceptible noise to generate misleading outcomes. Noteworthy examples include deceiving watermark detectors with substitute models and using adversarial patches for 3D images. Researchers continuously explore new attack methods, such as real-time adversarial perturbation frameworks for object detection in autonomous vehicles or sophisticated techniques altering visual perception in camera-based object recognition systems. Understanding and defending against such attacks are essential for enhancing the security and reliability of computer vision systems (Hao & Tao, 2022; G. Zhang & Sikdar, 2022; Shinde & Shah, 2018; Khalid et al., 2019; Seo, Park, & Kang, 2022; Quiring & Rieck, 2018; Drenkow, Lennon, Wang, & Burlina, 2023; Yoon, Jafarnejadsani, & Voulgaris, 2023; Man, Li, & Gerdes, 2023).

AML attacks on natural language processing

Adversarial attacks in Natural Language Processing (NLP) pose distinct challenges in comparison to image processing. (Marulli et al., 2021) highlight that attacks developed for image systems are ineffective when applied to the vector representations commonly used in NLP. (Edwards & Rawat, 2020) note that adversarial examples in NLP can impact model performance by modifying semantics, spelling, or even entire phrases, although these alterations are often noticeable. For instance, (Cresci, Petrocchi, Spognardi, & Tognazzi, 2022) describe attacks on fake news detection systems, such as "TextBugger", which manipulate the content of news articles to deceive classifiers.

These attacks target the title, content, or source of the article to reverse the classifier's outcome.

AML attacks on Wireless Networks

The wireless domain faces emerging security challenges with the rise of ML applications. Research points to the lack of focus on AML attacks in network traffic classification due to the difficulty of altering data packets without changing their content. Examples include a Backdoor poisoning attack affecting ML model behavior in wireless networks and vulnerabilities of ML-based modulation models and Deep Learning techniques in autonomous cognitive networks and TOR traffic classification systems, respectively. Solutions like the "Fast Gradient Sign Method" (FGSM) enhance security by generating adversarial samples. Attacks on NextG networks leverage AML to disrupt resource allocation, countered by defenses such as Q-Protect, RandomOpt, RandomTop, and MisNACK to safeguard against AML threats (Shi, Zeng, & Nguyen, 2019; Davaslioglu & Sagduyu, 2019; Usama, Asim, et al., 2019; Usama, Qayyum, Qadir, & Al-Fuqaha, 2019; E. Catak, Catak, & Moldsvor, 2021; Shi, Sagduyu, Erpek, & Gursoy, 2023).

AML attack on cybersecurity defenses

AML attacks pose unique challenges in cybersecurity, especially in computer security where complex and evolving ML systems are more susceptible to breaches. Limited research exists in this area, with examples including DNS server poisoning, attacks on Intrusion Detection Systems (IDS), malware evasion strategies, SPAM detection system vulnerabilities, and phishing website detection challenges. Researchers emphasize defense techniques like adversarial training and the need for robust security mechanisms in combating such attacks (Shi et al., 2019; Yeboah-Ofori et al., 2021; Jin, Tomoishi, & Matsuura, 2019; Anthi et al., 2021; Ayub et al., 2020; Chen, Ye, & Bourlai, 2017; Yuan, Apruzzese, & Conti, 2023).

AML attacks on Quantum ML

Classification models based on Quantum ML provide efficiency in handling large volumes of data, though they face security challenges akin to traditional models. (Edwards & Rawat, 2020) warn about the negative impact that adversarial examples can have on quantum ML models, emphasizing the possibility of attacks like the Fast Gradient Sign Method (FGSM) and the risk of transfer attacks, although the latter has not been extensively explored yet.

AML attacks on audio processing

In this domain, adversarial examples have received less exploration but can produce a significant impact on model performance. (Lin & Biggio, 2021) present a case

of targeted evasion adversarial attack in audio, specifically aimed at the Mozilla DeepSpeech speech-to-text recognition system. The inclusion of barely perceptible noises compromised the accuracy of the model, showing the vulnerability within this specific context.

Perturbations or adversarial examples

Perturbations play a critical role in AML by injecting malicious data to deceive ML classifiers, leading to incorrect predictions. These imperceptible manipulations pose a significant security challenge, especially for Deep Learning systems. Research pioneers highlighted vulnerabilities in neural networks to such perturbations, impacting ML system accuracy. Perturbations are crafted to maximize prediction errors and their transferability between models is noteworthy. Techniques like the Fast Gradient Sign Method (FGSM) are used for optimal perturbation generation. Perturbations also aid in obfuscating ML models for security. Recent attack strategies include Cloak & Co-locate, involving co-locating attacker VMs with victims in the cloud, and manipulating fraud detection systems through perturbations in training datasets to deceive models and enable fraudulent activities (Guihai & Sikdar, 2021; McDaniel, Papernot, & Celik, 2016; Ntalampiras, 2023; Szegedy et al., 2013; Cresci et al., 2022; Khalid et al., 2019; Lin & Biggio, 2021; Davaslioglu & Sagduyu, 2019; G. Verma et al., 2018; Nazari et al., 2023; Paladini et al., 2023).

AML attack taxonomies

We identified various approaches for classifying AML attack techniques, including NIST Taxonomy, Barreño's Proposal, Papernot's Categorization, Khalid et al.'s Taxonomy, and Olney and Karam's Classification. Different ways to categorize AML attacks include considering the type of security violation such as reliability, integrity, and availability attacks. Attacks can also be categorized based on specificity (targeted, indiscriminate, combined attacks), learning phase (testing, training phase attacks), and adversary's knowledge (white-box, black-box, and gray-box attacks) (Tabassi et al., 2019; Barreno et al., 2006; Papernot et al., 2016; Khalid et al., 2019; Olney & Karam, 2022; Ma et al., 2020; Anthi et al., 2021; Mehta et al., 2022; Liu et al., 2020; Khalid et al., 2020; U. Verma et al., 2022; Ayub et al., 2020).

AML attack strategies

Evasion attacks are a critical focus in AML that manipulate data during the testing phase to avoid detection without impacting the training process. These attacks involve introducing noise to test data to create adversarial perturbations, evading detection and potentially leading to misclassifications. Various methods, such as the Fast Gradient Sign Method (FGSM), DeepFool, L-BFGS, C&W attack, EnvAttack, JSMA, Mutual Information, and SIFA, are employed to generate optimal perturbations, evade detection

systems, and impact machine learning models in different scenarios. Notable examples include using JSMA in malware classification systems and FGSM in targeted and untargeted attacks. These evasion attacks pose a significant threat due to their ability to induce misclassifications without the need to modify the model's structure, ultimately compromising system security and accuracy (Anthi et al., 2021; E. Catak et al., 2021; Chen et al., 2017; Ebrahimabadi et al., 2021; Khalid et al., 2019; Ma et al., 2020; Olney & Karam, 2022; Venkatesan et al., 2021).

Model poisoning attacks, such as causal attacks, aim to introduce noise and modify labels during model training to degrade performance across various algorithms like SVM and Deep Learning (Baracaldo et al., 2018; Chiba et al., 2020; Davaslioglu & Sagduyu, 2019; Lin & Biggio, 2021; Ma et al., 2020; Marulli et al., 2021; Olney & Karam, 2022; Tian et al., 2022; Venkatesan et al., 2021; U. Verma et al., 2022). Backdoor attacks and Neural Trojans are integral to such poisoning tactics, posing significant security risks and manipulating models during both training and testing phases (Baracaldo et al., 2018; Davaslioglu & Sagduyu, 2019; Lin & Biggio, 2021; Marulli et al., 2021; Olney & Karam, 2022). Transfer attacks leverage the transferability of adversarial perturbations across models, emphasizing the need for robust defenses against these insidious threats (Lin & Biggio, 2021; Olney & Karam, 2022; Usama et al., 2020; U. Verma et al., 2022; Papernot et al., 2016; Edwards & Rawat, 2020; Quiring & Rieck, 2018; G. Zhang & Sikdar, 2022).

Adversarial perturbations not only pose a threat to ML models but can also be used to enhance their robustness (Cresci et al., 2022). Researchers have identified AML attacks as potential defensive techniques (Ebrahimabadi et al., 2021). For example, (Yilmaz & Siraj, 2021) present AMLODA, a model that seeks to hide patterns of electricity consumption through minimal and imperceptible data perturbations. (Ebrahimabadi et al., 2021) provide another example where poisoning Challenge-Response transmissions between IoT devices and a server is suggested as a means to prevent replay or spoofing attacks. (G. Verma et al., 2018) propose a method for a defender to achieve their obfuscation objectives using the L-BFGS attack, generating perturbations that make a sample virtually indistinguishable from the original in terms of packet size.

AML mitigation techniques

To defend ML models against adversarial attacks, strategies like gradient masking, robust optimization, and adversarial detection are crucial (Ma et al., 2020; U.Verma et al., 2022). Detection and prevention defenses, reactive defenses, and proactive defenses are implemented to mitigate the impact of AML attacks (Jin et al., 2019; Usama et al., 2020). Techniques like Hiding the Probability Vector and data obfuscation are effective in countering black-box attacks and concealing network patterns for enhanced security (Khalid et al., 2020; G. Verma et al., 2018). Threat analysis plays a pivotal

role in understanding threats and implementing appropriate security measures (Lin & Biggio, 2021).

To defend against adversarial attacks, proactive strategies in AML like distillation, adversarial training, feature extraction, and incorporating valid samples into datasets are crucial (Lin & Biggio, 2021; Usama et al., 2020; Ma et al., 2020). Adversarial training, introduced by Goodfellow et al. (2014) and enhanced by Huang et al. (2015), enhances model robustness against perturbations and is effective across various applications (Anthi et al., 2021; Liu et al., 2020; Tian et al., 2021; Yilmaz, Siraj, & Ulybyshev, 2020; U. Verma et al., 2022; E. Catak et al., 2021; Tang et al., 2019). Defensive distillation counters adversarial perturbations in neural networks, offering advantages such as reduced network size and computational costs (Papernot et al., 2015; U. Verma et al., 2022; F. O. Catak et al., 2022). Data filtering by Baracaldo et al. (2018) segregates malicious data, while detection methods aim to identify adversarial perturbations in training data, highlighting the importance of robust defenses in ML systems (W. Li et al., 2022; Anthi et al., 2021). Other techniques such as feature compression, noise reduction, depolarization, ROSA, Dual Model Divergence, and Hierarchical Clustering further enhance model security and resilience against attacks (Edwards & Rawat, 2020; Zhao, Yue, & Wang, 2023; Aboutalebi, Shafiee, Tai, & Wong, 2023; McCarthy, Ghadafi, Andriotis, & Legg, 2023).

Traditional security techniques against AML

There are traditional techniques that can enhance the robustness and resistance to attacks in ML systems, even though they were not explicitly designed to counter AML. These techniques provide an additional layer of defense against AML attacks. However, it is crucial to recognize that in practical scenarios, these techniques may be inadequate (U. Verma et al., 2022). One strategy within this category is the analysis of training data, where potentially malicious data is identified by examining its origin and associated metadata (Baracaldo et al., 2018). Another strategy is the refinement of the training process, which aims to smoothen the decision boundaries of the model or estimate the probability of an input being an adversarial sample based on its characteristics (McDaniel et al., 2016). Moreover, in their work, (G. Verma et al., 2018) discuss the use of encryption as a means to eliminate identifiable data in a network traffic classifier while retaining recognizable features such as packet size and arrival intervals.

AML attack metrics

We identified various attack metrics such as Perturbation Success Rate, ASR Success Rate, Classification Confidence, Recall, Precision, Specificity, MSE, Perturbation Norm D, SSIM, CC, F1 Score, Accuracy, False Detection Probability, False Alarm Probability, Decision Boundary Distance, Minimum Cost for a Successful Attack, Inference Stability, and Misclassification Rate. These metrics play a crucial role in evaluating the success

and impact of different adversarial attacks on machine learning models (Liu et al., 2020; G. Zhang & Sikdar, 2022; Ntalampiras, 2023; Guihai & Sikdar, 2021; Usama et al., 2019; Marulli et al., 2021; F. O. Catak et al., 2022; Khalid et al., 2020; Yilmaz & Siraj, 2021; Shi et al., 2019; Ma et al., 2020).

AML defense metrics

We identified False Positive Rate (FPR) and False Negative Rate (FNR) (Yilmaz & Siraj, 2021): Evaluate AMLODA model performance by measuring incorrect classifications. Model Resilience (McDaniel et al., 2016): Measures the model's ability to withstand input perturbations. Robustness (W. Li et al., 2022): Compares the classifier's robustness under different parameters. Root Mean Square Error (RMSE) (Tang et al., 2019): Assesses the quality of adversarial training in Neural Networks by calculating prediction errors.

Challenges and recommendations to implement AML

The practical implementation of AML faces challenges due to lack of actionable research, evolving adversarial strategies, and discrepancies in security expectations. Limited interaction between statistical and ML communities complicates the process. Overcoming these hurdles requires a deep analysis of security relations, which is complex for organizations (Kumar et al., 2020; W. Li et al., 2022; Yeboah-Ofori et al., 2021; Zizzo et al., 2019). Implementing AML attacks is hindered by attackers' varying knowledge levels, particularly in black-box attacks, where understanding the model is challenging (Ma et al., 2020; Khalid et al., 2020). The intricacies of ML implementations demand time and expertise, limiting exploration across various practical domains, like autonomous networks (Usama et al., 2020). Restricted access to model information impedes security research involving ML models (Wilhjelm & Younis, 2020). Adversarial perturbations in attacks must remain feasible to avoid irreversibly corrupting samples (Venkatesan et al., 2021).

Existing literature on defense strategies against AML often lacks adequate mitigation measures despite growing awareness of the issue (Anthi et al., 2021). Addressing this challenge is an ongoing battle as attackers constantly create new ways to bypass defenses, perpetuating a cycle (Usama et al., 2020). Implementing effective solutions is complex and requires deep understanding of security measures and model operations (Anthi et al., 2021). For example, excessive use of samples in Adversarial Training can impact model accuracy, aligning classifications with adversarial patterns (X. Li et al., 2020; Simion, Gavrilut, & Luchian, 2019). Despite the hurdles, developing impactful defenses is feasible as attackers face limitations in targeting models, and the field has seen significant advancements (Usama et al., 2020).

Conclusions and future work

ML systems' significance, especially in cloud computing, has led to increased vulnerability to advanced threats that traditional mitigation techniques may not address adequately. AML is now crucial in cybersecurity, safeguarding ML systems at scale in cloud infrastructures. AML attacks can severely affect ML systems, evolving rapidly to include subtle data manipulation and nearly undetectable perturbations.

While AML research offers mitigation measures, defense strategies are limited compared to attack techniques, indicating a need for more attention to developing innovative defenses. Common defense strategies like Adversarial Training are prevalent, highlighting the necessity for both proven and innovative defense approaches to combat evolving AML threats.

Future plans involve a more extensive study on AML, exploring industry standards and grey literature like arXiv to gain insights into variations between academic and industry AML approaches.

Bibliografía

Aboutalebi, H., Shafiee, M. J., Tai, C.-e. A., & Wong, A. (2023). Knowing is half the battle: Enhancing clean data accuracy of adversarial robust deep neural networks via dual-model bounded divergence gating. IEEE Access, 1–1. http://dx.doi.org/10.1109/ACCESS.2023.3347498

Aljumah, A., & Ahanger, T. A. (2020). Cyber security threats, challenges and defence mechanisms in cloud computing. IET Communications, 14(7), 1185-1191. http://dx.doi.org/10.1049/iet-com.2019.0040

Anthi, E., Williams, L., Rhode, M., Burnap, P., & Wedgbury, A. (2021). Adversarial attacks on machine learning cybersecurity defences in industrial control systems. Journal of Information

Security and Applications, 58(nil), 102717. http://dx.doi.org/10.1016/j.jisa.2020.102717

Ayub, M. A., Johnson, W. A., Talbert, D. A., & Siraj, A. (2020, March). Model evasion attack on intrusion detection systems using adversarial machine learning. In 2020 54th annual conference on information sciences and systems (ciss). IEEE.http://dx.doi.org/10.1109/CISS48834.2020.1570617116

Badiger, V. S., & Shyam, D. K. (2023, 1). A survey on cloud security threats using deep learning algorithms. In 2023 international conference on intelligent and innovative technologies in computing, electrical and electronics (iitcee) (p. 696-701). http://dx.doi.org/10.1109/IITCEE57236.2023.10090981

Baracaldo, N., Chen, B., Ludwig, H., Safavi, A., & Zhang, R. (2018, July). Detecting poisoning attacks on machine learning in iot environments. In 2018 ieee international congress on internet of things (iciot). IEEE. http://dx.doi.org/10.1109/ICIOT.2018.00015

Barreno, M., Nelson, B., Sears, R., Joseph, A. D., & Tygar, J. D. (2006, 3). Can machine learning be secure? In Proceedings of the 2006 acm symposium on information, computer and communications security (p. nil). http://dx.doi.org/10.1145/1128817.1128824

Catak, E., Catak, F. O., & Moldsvor, A. (2021, May). Adversarial machine learning security problems for 6g: mmwave beam prediction use-case. In 2021 ieee international black sea conference on communications and networking (blackseacom).

IEEE. http://dx.doi.org/10.1109/BlackSeaCom52164.2021.9527756

Catak, F. O., Kuzlu, M., Tang, H., Catak, E., & Zhao, Y. (2022). Security hardening of intelligent reflecting surfaces against adversarial machine learning attacks. IEEE Access, 10, 100267–100275. http://dx.doi.org/10.1109/ACCESS.2022.3206012

Chen, L., Ye, Y., & Bourlai, T. (2017, September). Adversarial machine learning in malware detection: Arms race between evasion attack and defense. In 2017 european intelligence and security informatics conference (eisic). IEEE. http://dx.doi.org/10.1109/EISIC.2017.21

Chiba, T., Sei, Y., Tahara, Y., & Ohsuga, A. (2020, December). A defense method against poisoning attacks on iot machine learning using poisonous data. In 2020 ieee third international conference on artificial intelligence and knowledge engineering (aike). IEEE. http://dx.doi.org/10.1109/AIKE48582.2020.00022

Cresci, S., Petrocchi, M., Spognardi, A., & Tognazzi, S. (2022, March). Adversarial machine learning for protecting against online manipulation. IEEE Internet Computing, 26(2), 47–52. http://dx.doi.org/10.1109/MIC.2021.3130380

Cruzes, D. S., & Dyba, T. (2011, 9). Recommended steps for thematic synthesis in software engineering. In 2011 international symposium on empirical software engineering and measurement (p. 275-284). http://dx.doi.org/10.1109/ESEM.2011.36

Davaslioglu, K., & Sagduyu, Y. E. (2019, November). Trojan attacks on wireless signal classification with adversarial machine learning. In 2019 ieee international symposium on dynamic spectrum access networks (dyspan). IEEE. http://dx.doi.org/10.1109/DySPAN.2019.8935782

Drenkow, N., Lennon, M., Wang, I.-J., & Burlina, P. (2023, January). Do adaptive active attacks pose greater risk than static attacks? In 2023 ieee/cvf winter conference on applications of computer vision (wacv). IEEE. http://dx.doi.org/10.1109/WACV56688.2023.00143

Ebrahimabadi, M., Lalouani, W., Younis, M., & Karimi, N. (2021, 7). Countering puf modeling attacks through adversarial machine learning. In 2021 ieee computer society annual symposium on vlsi (isvlsi) (p. 356-361). http://dx.doi.org/10.1109/ISVLSI51109.2021.00071

Edwards, D., & Rawat, D. B. (2020, October). Quantum adversarial machine learning: Status, challenges and perspectives. In 2020 second ieee international conference on trust, privacy and security in intelligent systems and applications (tps-isa). IEEE. http://dx.doi.org/10.1109/TPS-ISA50397.2020.00026

Frederickson, C., Moore, M., Dawson, G., & Polikar, R. (2018, 7). Attack strength vs. detectability dilemma in adversarial machine learning. In 2018 international joint conference on neural networks (ijcnn) (p. 1-8). http://dx.doi.org/ 10.1109/IJCNN.2018.8489495

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. https://arxiv.org/abs/1412.6572

Guesmi, A., Khasawneh, K. N., Abu-Ghazaleh, N., & Alouani, I. (2022, July). Room: Adversarial machine learning attacks under real-time constraints. In 2022 international joint conference on neural networks (ijcnn). IEEE. http://dx.doi.org/ 10.1109/IJCNN55064.2022.9892437

Guihai, Z., & Sikdar, B. (2021, 10). Adversarial machine learning against false data injection attack detection for smart grid demand response. In 2021 ieee international conference on communications, control, and computing technologies for smart grids (smartgridcomm) (p. 352-357). http://dx.doi.org/10.1109/SmartGridComm51999.2021 .9632316

Hao, J., & Tao, Y. (2022, May). Adversarial attacks on deep learning models in smart

grids. Energy Reports, 8, 123–129. http://dx.doi.org/10.1016/j.egyr.2021.11.026
Huang, R., Xu, B., Schuurmans, D., & Szepesvári, C. (2015). Learning with a strong adversary. arXiv preprint arXiv:1511.03034.
Jin, Y., Tomoishi, M., & Matsuura, S. (2019, September). A detection method against dns cache poisoning attacks using machine learning techniques: Work in progress. In 2019 ieee 18th international symposium on network computing and applications (nca). IEEE. http://dx.doi.org/10.1109/NCA.2019.8935025
Khalid, F., Hanif, M. A., Rehman, S., Qadir, J., & Shafique, M. (2019, March). Fademl: Understanding the impact of pre- processing noise filtering on adversarial machine learning. In 2019 design, automation and test in europe conference and exhibition (date). IEEE. http://dx.doi.org/10.23919/DATE.2019.8715141
Kitchenham, B. A., Budgen, D., & Brereton, P. (2015). Evidence-based software engineering and systematic reviews (Vol. 4). CRC press.
Kumar, R. S. S., Nystrom, M., Lambert, J., Marshall, A., Goertzel, M., Comissoneru, A., . . . Xia, S. (2020, 5). Adversarial machine learning-industry perspectives. In 2020 ieee security and privacy workshops (spw) (p. 69-75). http:// dx.doi.org/10.1109/SPW50608.2020.00028
Lagesse, B., Burkard, C., & Perez, J. (2016, March). Securing pervasive systems against adversarial machine learning. In 2016 ieee international conference on pervasive computing and communication workshops (percom workshops). IEEE. http://dx.doi.org/10.1109/PERCOMW.2016.7457061
Li, W., Liu, X., Yan, A., & Yang, J. (2022, August). Kernel-based adversarial attacks and defenses on support vector classification. Digital Communications and Networks, 8(4), 492–497. http://dx.doi.org/10.1016/j.dcan .2021.12.003
Li, X., Qiu, K., Qian, C., & Zhao, G. (2020, July). An adversarial machine learning method based on opcode n-grams feature in malware detection. In 2020 ieee fifth international conference on data science in cyberspace (dsc). IEEE. http://dx.doi.org/10.1109/DSC50466.2020.00066
Lin, H.-Y., & Biggio, B. (2021). Adversarial machine learning: Attacks from laboratories to the real world. Computer, 54(5), 56-60. http://dx.doi.org/10.1109/MC.2021.3057686
Liu, K., Yang, H., Ma, Y., Tan, B., Yu, B., Young, E. F. Y., . . . Garg, S. (2020, August). Adversarial perturbation attacks on ml-based cad: A case study on cnn-based lithographic hotspot detection. ACM Transactions on Design Automation of Electronic Systems, 25(5), 1–31. http://dx.doi.org/10.1145/3408288
Ma, Y., Xie, T., Li, J., & Maciejewski, R. (2020, January). Explaining vulnerabilities to adversarial machine learning through visual analytics. IEEE Transactions on Visualization and Computer Graphics, 26(1), 1075–1085. http:// dx.doi.org/10.1109/TVCG.2019.2934631
Man, Y., Li, M., & Gerdes, R. (2023, May). Remote perception attacks against camera-based object recognition systems and countermeasures. ACM Transactions on Cyber-Physical Systems. http://dx.doi.org/10.1145/3596221
Marulli, F., Verde, L., & Campanile, L. (2021). Exploring data and model poisoning

attacks to deep learning-based nlp systems. Procedia Computer Science, 192, 3570–3579. http://dx.doi.org/10.1016/j.procs.2021.09.130

McCarthy, A., Ghadafi, E., Andriotis, P., & Legg, P. (2023, February). Defending against adversarial machine learning attacks using hierarchical learning: A case study on network traffic attack classification. Journal of Information Security and Applications, 72, 103398. http://dx.doi.org/10.1016/j.jisa.2022.103398

McDaniel, P., Papernot, N., & Celik, Z. B. (2016, May). Machine learning in adversarial settings. IEEE Security and Privacy, 14(3), 68–72. http://dx.doi.org/10.1109/MSP.2016.51

Mehta, C., Harniya, P., & Kamat, S. (2022, 2). Comprehending and detecting vulnerabilities using adversarial machine learning attacks. In 2022 2nd international conference on artificial intelligence and signal processing (aisp) (p. 1-5). http://dx.doi.org/10.1109/AISP53593.2022.9760580

Nazari, N., Makrani, H. M., Fang, C., Omidi, B., Rafatirad, S., Sayadi, H., . . . Homayoun, H. (2023, September). Adversarial attacks against machine learning-based resource provisioning systems. IEEE Micro, 43(5), 35–44. http://dx.doi .org/10.1109/MM.2023.3267481

Nicolae, M.-I., Sinn, M., Tran, M. N., Buesser, B., Rawat, A., Wistuba, M., . . . others (2018). Adversarial robustness toolbox v1. 0.0. arXiv preprint arXiv:1807.01069.

Ntalampiras, S. (2023, February). Adversarial attacks against acoustic monitoring of industrial machines. IEEE Internet of Things Journal, 10(4), 2832–2839. http://dx.doi.org/10.1109/JIOT.2022.3194703

Olney, B., & Karam, R. (2022). Diverse, neural trojan resilient ecosystem of neural network ip. ACM Journal on Emerging Technologies in Computing Systems, 18(3), 1-23. http://dx.doi.org/10.1145/3471189

P, L. M. D., & Gunasekaran, M. (2023, November). Generating and defending against adversarial examples for loan eligibility prediction. In 2023 international conference on system, computation, automation and networking (icscan). IEEE. http://dx.doi.org/10.1109/ICSCAN58655.2023.10395585

Paladini, T., Monti, F., Polino, M., Carminati, M., & Zanero, S. (2023, October). Fraud detection under siege: Practical poisoning attacks and defense strategies. ACM Transactions on Privacy and Security, 26(4), 1–35. http://dx.doi.org/10.1145/3613244

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016, 3). The limitations of deep learning in adversarial settings. In 2016 ieee european symposium on security and privacy (p. nil). http://dx.doi.org/ 10.1109/EuroSP.2016.36

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2015). Distillation as a defense to adversarial perturbations against deep neural networks. nil, nil(nil), nil. https://arxiv.org/abs/1511.04508

Qayyum, A., Usama, M., Qadir, J., & Al-Fuqaha, A. (2020). Securing connected and autonomous vehicles: Challenges posed by adversarial machine learning and the way forward. IEEE Communications Surveys and Tutorials, 22(2), 998-

1026. http://dx.doi.org/10.1109/COMST.2020.2975048

Quiring, E., & Rieck, K. (2018, September). Adversarial machine learning against digital watermarking. In 2018 26th european signal processing conference (eusipco). IEEE. http://dx.doi.org/10.23919/EUSIPCO.2018.8553343

Seo, J., Park, S., & Kang, J. (2022, June). Secure wireless communication via adversarial machine learning: A priori vs. a posteriori. ICT Express, 8(2), 220–224. http://dx.doi.org/10.1016/j.icte.2021.06.005

Shi, Y., Sagduyu, Y. E., Erpek, T., & Gursoy, M. C. (2023). How to attack and defend nextg radio access network slicing with reinforcement learning. IEEE Open Journal of Vehicular Technology, 4, 181–192. http://dx.doi.org/10.1109/OJVT.2022.3229229

Shi, Y., Zeng, H., & Nguyen, T. T. (2019, November). Adversarial machine learning for network security. In 2019 ieee international symposium on technologies for homeland security (hst). IEEE. http://dx.doi.org/10.1109/HST47167.2019.9032936

Shinde, P. P., & Shah, S. (2018, 8). A review of machine learning and deep learning applications. In 2018 fourth international conference on computing communication control and automation (iccubea) (p. nil). http://dx.doi.org/10.1109/ ICCUBEA.2018.8697857

Simion, C.-A., Gavrilut, D. T., & Luchian, H. (2019, September). An adversarial machine learning approach to evaluate the robustness of a security solution. In 2019 21st international symposium on symbolic and numeric algorithms for scientific computing (synasc). IEEE. http://dx.doi.org/10.1109/SYNASC49474.2019.00028

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. nil, nil(nil), nil. https://arxiv.org/abs/1312.6199

Tabassi, E., Burns, K. J., Hadjimichael, M., Molina-Markham, A. D., & Sexton, J. T. (2019). A taxonomy and terminology of adversarial machine learning. http://dx.doi.org/10.6028/nist.ir.8269-draft

Tang, Z., Jiao, J., Zhang, P., Yue, M., Chen, C., & Yan, J. (2019, August). Enabling cyberattack-resilient load forecasting through adversarial machine learning. In 2019 ieee power and energy society general meeting (pesgm). IEEE. http://dx.doi.org/10.1109/PESGM40551.2019.8974076

Tian, J., Wang, B., Li, J., & Konstantinou, C. (2021, November). Adversarial attack and defense methods for neural network based state estimation in smart grid. IET Renewable Power Generation, 16(16), 3507–3518. http://dx.doi.org/10.1049/rpg2.12334

Tian, J., Wang, B., Li, J., Wang, Z., Ma, B., & Ozay, M. (2022, August). Exploring targeted and stealthy false data injection attacks via adversarial machine learning. IEEE Internet of Things Journal, 9(15), 14116–14125. http://dx.doi.org/ 10.1109/JIOT.2022.3147040

Usama, M., Asim, M., Qadir, J., Al-Fuqaha, A., & Imran, M. A. (2019, August).

Adversarial machine learning attack on modulation classification. In 2019 uk/ china emerging technologies (ucet). IEEE. http://dx.doi.org/10.1109/UCET.2019.8881843

Usama, M., Qadir, J., & Al-Fuqaha, A. (2018, 10). Adversarial attacks on cognitive self-organizing networks: The challenge and the way forward. In 2018 ieee 43rd conference on local computer networks workshops (lcn workshops) (p. nil). http://dx.doi.org/10.1109/LCNW.2018.8628538

Usama, M., Qadir, J., Al-Fuqaha, A., & Hamdi, M. (2020, January). The adversarial machine learning conundrum: Can the insecurity of ml become the achilles' heel of cognitive networks? IEEE Network, 34(1), 196–203. http:// dx.doi.org/10.1109/MNET.001.1900197

Usama, M., Qayyum, A., Qadir, J., & Al-Fuqaha, A. (2019, June). Black-box adversarial machine learning attack on network traffic classification. In 2019 15th international wireless communications and mobile computing conference (iwcmc). IEEE. http://dx.doi.org/10.1109/IWCMC.2019.8766505

Venkatesan, S., Sikka, H., Izmailov, R., Chadha, R., Oprea, A., & de Lucia, M. J. (2021, 11). Poisoning attacks and data sanitization mitigations for machine learning models in network intrusion detection systems. In Milcom 2021 - 2021 ieee military communications conference (milcom) (p. 874-879). http://dx.doi.org/10.1109/MILCOM52596 .2021.9652916

Verma, G., Ciftcioglu, E., Sheatsley, R., Chan, K., & Scott, L. (2018, October). Network traffic obfuscation: An adversarial machine learning approach. In Milcom 2018 - 2018 ieee military communications conference (milcom). IEEE. http:// dx.doi.org/10.1109/MILCOM.2018.8599680

Verma, U., Huang, Y., Woodward, C., Schmugar, C., Ramagopal, P. P., & Fralick, C. (2022, August). Attacking malware detection using adversarial machine learning. In 2022 4th international conference on data intelligence and security (icdis). IEEE. http://dx.doi.org/10.1109/ICDIS55630.2022.00014

Wilhjelm, C., & Younis, A. A. (2020, December). A threat analysis methodology for security requirements elicitation in machine learning based systems. In 2020 ieee 20th international conference on software quality, reliability and security companion (qrs-c). IEEE. http://dx.doi.org/10.1109/QRS-C51114.2020.00078

Wohlin, C. (2014, 5). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In Proceedings of the 18th international conference on evaluation and assessment in software engineering (p. 1-10). http://dx.doi.org/10.1145/2601248.2601268

Yeboah-Ofori, A., Ismail, U. M., Swidurski, T., & Opoku-Boateng, F. (2021, July). Cyber threat ontology and adversarial machine learning attacks: Analysis and prediction perturbance. In 2021 international conference on computing, computational modelling and applications (iccma). IEEE. http://dx.doi.org/10.1109/ICCMA53594.2021.00020

Yilmaz, I., & Siraj, A. (2021). Avoiding occupancy detection from smart meter using adversarial machine learning. IEEE Access, 9, 35411–35430. http://dx.doi.

org/10.1109/ACCESS.2021.3057525

Yilmaz, I., Siraj, A., & Ulybyshev, D. (2020, December). Improving dga-based malicious domain classifiers for malware defense with adversarial machine learning. In 2020 ieee 4th conference on information and communication technology (cict). IEEE. http://dx.doi.org/10.1109/CICT51604.2020.9311925

Yoon, H.-J., Jafarnejadsani, H., & Voulgaris, P. (2023, July). Learning when to use adaptive adversarial image perturbations against autonomous vehicles. IEEE Robotics and Automation Letters, 8(7), 4179–4186. http://dx.doi.org/ 10.1109/ LRA.2023.3280813

Yuan, Y., Apruzzese, G., & Conti, M. (2023, December). Multi-spacephish: Extending the evasion-space of adversarial attacks against phishing website detectors using machine learning. Digital Threats: Research and Practice. http:// dx.doi. org/10.1145/3638253

Zhang, G., & Sikdar, B. (2022, 10). Ensemble and transfer adversarial attack on smart grid demand-response mechanisms. In 2022 ieee international conference on communications, control, and computing technologies for smart grids (smartgridcomm) (p. 53-58). http://dx.doi.org/10.1109/ SmartGridComm52983.2022.9960966

Zhang, H., Babar, M. A., & Tell, P. (2011). Identifying relevant studies in software engineering. Information and Software Technology, 53(6), 625-637. http:// dx.doi.org/10.1016/j.infsof.2010.12.010

Zhao, T., Yue, M., & Wang, J. (2023, November). Robust power system stability assessment against adversarial machine learning-based cyberattacks via online purification. IEEE Transactions on Power Systems, 38(6), 5613–5622. http:// dx.doi.org/10.1109/TPWRS.2022.3233735

Zizzo, G., Hankin, C., Maffeis, S., & Jones, K. (2019, June). Adversarial machine learning beyond the image domain. In Proceedings of the 56th annual design automation conference 2019. ACM. http://dx.doi.org/10.1145/ 3316781.3323470