Adversarial Attacks on Machine Learning Models

Divya Jayant Sarode[1], Divya Prakash Surwade[2]

[1, 2] PG Students, , 🆔 *0009-0004-4750-0106*   🆔 *0009-0000-3368-1007*
1,2 Godavari Foundations, Godavari College of Engineering, Jalgaon, India, 425001,

Email of Authors: divyapatil2918@gmail.com, surwade2002divya@gmail.com

Abstract *–The rapid advancement of artificial Intelligence (AI) applications have brought security challenges, particularly in the form of adversarial machine learning (AML) attacks. As organizations worldwide invest in developing their own large language models and AI-driven applications, concerns over data security and model integrity have grown significantly. AML attacks pose a serious threat by manipulating machine learning models, often leading to a drastic decline in their accuracy and reliability. These attacks are especially alarming in critical domains such as healthcare and autonomous transportation, where compromised AI systems can have severe real-world consequences.*

This paper systematically explores various AML attack strategies, categorizing them based on adversarial techniques and tactics. It also examines their impact on machine learning models and highlights vulnerabilities that attackers exploit. Additionally, we review open-source tools designed to test AI and ML systems against adversarial threats, providing organizations with practical solutions for security assessment. By presenting a comprehensive analysis and actionable security recommendations, this study aims to assist organizations in safeguarding their machine learning models and ensuring robust AI deployment in real-world applications.

Keywords: Adversarial Machine Learning, AI Security, Cyber Resilience, Defense Mechanisms, Explainable AI.

I. INTRODUCTION

Adversarial machine learning (AML) has emerged as a major security challenge in the field of artificial intelligence. In addition to finance, component ntegration and other related fields, the work also applies to areas as technique-leading today as artificial intelligence itself. With adversarial attacks on these systems, machine learning (ML) models are used in more and more sectors. This has gained the concern of developers and security researchers.in self-driving cars, for example, if an enemy launches a cyberattack on the road sign arm and knee support framework it may result in a stop sign becoming speed limit signs; this could easily lead to accidents. Even while ML models are typically good at categorizing benign inputs, studies have shown that adversaries can quietly alter inputs to make the model predict the wrong thing. This weakness emphasizes how crucial it is to create strong defenses against AML attacks. A thorough analysis of the different kinds of assaults, their effects, accessible security testing tools, and practical mitigation techniques is still missing from the literature, despite the growing concern regarding AML.

A. Problem Statement

In the realm of cyber security, an attack on a security system is any attempt to disrupt its intended functionality, compromising its **confidentiality, integrity, or availability**. Similarly, adversarial attacks on artificial intelligence (AI)

Systems involve deliberately manipulating the model to behave unpredictably, often leading to incorrect or misleading outputs. Adversarial Machine Learning (AML) attacks are specifically designed by threat actors to exploit vulnerabilities in machine learning (ML) models, making them act in unexpected ways or generate

International Journal of Innovations in Engineering and Science,   www.ijies.net

erroneous findings. These attacks may take place during testing, prior to deployment (during training), or even After the model has been implemented in a real-world setting. Creating adversarial examples—inputs that have AML assaults can take many different forms, including privacy-based attacks, data poisoning, and evasion attempts. Additionally, they might be classified as either targeted or untargeted. An adversary manipulates the model in a targeted attack to cause a certain wrong result, like making an AI system install malware or stop functioning.

In image processing, for example, an adversary can make two photos look almost the same to the human eye while forcing a machine learning model to identify them as entirely distinct objects because of subtle pixel differences. As seen in Figure 1, these modest modifications cause misclassification by taking advantage of flaws in the model's decision limits.
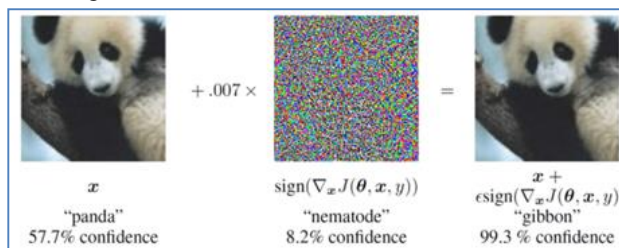

Fig 1: Adversarial input added to normal image; the classifier incorrectly identifies the image of a panda as a gibbon

B. Problem Description

A well-known adversarial assault scenario is depicted in Figure 1, where a convolutional neural network (ConvNet) incorrectly labels an image because of properly constructed adversarial perturbations. First, ConvNet successfully recognizes the unaltered image on the left as a panda. But when a small, well considered alteration is made to the picture, the model misidentifies it as a gibbon.The ideal direction for shifting all pixel values to trick the model is shown in the middle image of Figure 1. Although this alteration seems like random noise to human observers, it is calculated using ConvNet's parameters to take advantage of its flaws. The model maintains some degree of uncertainty even after the adversarial change, since it still gives the image a 58% chance of being a panda. However, the model's decision-making process is drastically changed when this subtle disturbance is introduced into the original image and fed into the model in 32-bit floating-point format. Consequently, the ConvNetmore confidently misclassifies the altered panda image as a gibbon than it did in its initial, albeit inaccurate, classification.

This occurrence draws attention to a critical flaw in machine learning models: even small input disturbances can produce wildly inaccurate results. These adversarial manipulations have been seen in actual AML assaults, when attackers strategically alter input data in minor ways to fool ML algorithms and cause major disruptions. The necessity for strong defenses to guarantee the dependability and security of AI systems is highlighted by this vulnerability.

C. Motivation of Research

Attacks using Adversarial Machine Learning (AML) pose a significant problem for the area of artificial intelligence (AI) and provide a rare chance to investigate one of the most urgent security issues that contemporary machine learning systems face. Adversarial assaults present serious threats to the confidentiality, integrity, and availability of machine learning algorithms, which are becoming the foundation of crucial applications in a variety of fields. The increasing complexity of these attacks emphasizes how urgent research into efficient defenses is needed. By revealing the various tactics adversaries employ to tamper with AI models and trick algorithms, this literature review seeks to offer a thorough analysis of AML. It also examines the most recent developments in security protocols intended to protect these systems.

D. Contributions

Adversarial Machine Learning (AML) attacks are thoroughly examined in this paper, along with their strategies, tactics, practical ramifications, and open-source tools for security testing and prevention. The following are the main contributions of this study: **Systematic Review & Taxonomy**: A thorough and organized analysis of AML attack techniques that places them into a comprehensible taxonomy to improve comprehension of their variances and effects. **Adversarial Tactics Analysis:** A thorough examination of the methods adversaries employs to carry out AML attacks, backed by actual case studies that illustrate the effects of such attacks. **Evaluation of Countermeasures:** Finding and evaluating countermeasures that are specific to various AML attack types gives ML researchers and cybersecurity experts fresh perspectives on protecting AI models. In addition to theoretical talks, this study offers open-source frameworks and tools that businesses may utilize to test and protect their machine learning models from hostile attacks. This research serves as a security reference for companies wishing to strengthen their AI-driven apps by outlining defensive tactics against AML

International Journal of Innovations in Engineering and Science,   www.ijies.net

threats. With practical suggestions for creating more robust AI systems, the paper mainly addresses AI developers, ML engineers, and cybersecurity experts. This study establishes groundwork for creating a secure ML testing environment, which will ultimately aid in the creation of reliable and secure AI applications in an increasingly intelligent digital world, even though protecting AI and ML models is still a challenge.

II.   RELATED WORKS

Attackers are now targeting AI and machine learning (ML) systems, causing serious risks for organizations through Adversarial Machine Learning (AML) attacks. These attacks can compromise the security of AI-driven applications, making it essential to build robust and secure AI models that can withstand such threats.

A. Machine Learning and Adversarial Attacks

ML models have transformed various fields by learning from data and making predictions. They play a crucial role in areas like medical diagnosis, autonomous vehicles, and data analysis, improving accuracy and

Fig 2. Structure of the review Article

efficiency.

However, adversarial attacks pose a major challenge by introducing deceptive inputs that trick ML models into making incorrect decisions.

Fig.3- A stop sign that underwent modification so that the ML model could identify it as a speed limit sign.
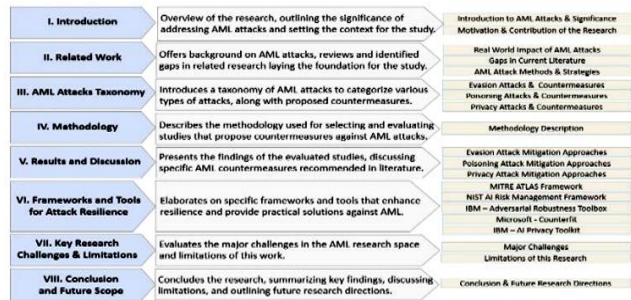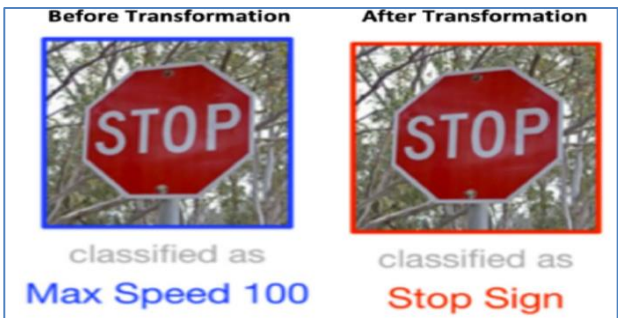
This section highlights the real-world risks of AML attacks and identifies existing research gaps. It lays the foundation for understanding different AML attack methods, defensive strategies, and available open-source tools. The insights from this discussion will help strengthen AI security in both academic research and real-world applications.

B. Real-World Impact of AML Attacks and Research Gaps

Adversarial Machine Learning (AML) attacks have had serious real-world consequences in cybersecurity, self-driving cars, healthcare, and even everyday AI applications. Below are some key examples of how attackers have exploited vulnerabilities in ML models:

Tesla Autopilot Attack – Ethical hackers at Keen Labs tricked Tesla's self-driving system using small stickers on the road, causing the car to swerve off course. This attack combined evasion and data poisoning techniques to fool the ML model.

1. **Email Spam Detection Bypass** – Attackers found a way to **manipulate spam filters** in Proofpoint's Email Protection system by reverse-engineering how emails were classified as spam. This allowed them to send undetected spam emails.

2. **Healthcare Privacy Leaks** – Hackers used **membership inference attacks** to predict whether a person had **HIV** by exploiting weaknesses in ML-based healthcare systems, leading to severe privacy risks.

These incidents highlight the urgent need to improve ML security.

Identified Gaps in Research and ML Security: Several studies have pointed out major challenges in securing ML models:

Lack of Awareness Among Developers: Many ML developers do not consider security during model development. Studies found that companies often **overlook AML risks**: and some ML engineers do not know how to defend against adversarial attacks.

Limited Research on Non-Visual Domains: Most research focuses on image-based attacks, leaving speech, text, and other ML applications relatively unexplored.

Unrealistic Security Solutions: Many past studies proposed security measures that reduce ML model accuracy or are too complex to implement in real-world settings.

Inadequate Defensive Strategies: Some studies outline threats without offering concrete solutions, making it

International Journal of Innovations in Engineering and Science, www.ijies.net

difficult for cybersecurity professionals to apply defenses effectively.

How This Research Fills the Gap

Unlike previous works that mainly focus on highlighting AML threats, this research goes a step further by: Providing practical countermeasures to defend against AML attacks. Connecting attack types to specific security solutions that organizations can implement. Offering a clear roadmap for securing ML models while maintaining performance.

By addressing these gaps, this study helps bridge the knowledge divide and ensures that ML models are better protected against evolving adversarial threats.

C. AML Attack Methods

Adversarial Machine Learning (AML) attacks generate deceptive inputs, known as adversarial examples, which trick ML models into making incorrect predictions. Below are three commonly used AML attack methods that hackers exploit:

1) Fast Gradient Sign Method (FGSM)

Invented by: Google researchers Ian J. Goodfellow, Jonathan Shlens, and Christian Szegedy

How it works:

FGSM modifies an input (e.g., an image) by adding a small, imperceptible noise in the direction of the model's gradient. The goal is to increase the prediction error, making it misclassify the input. The amount of noise is controlled by **epsilon (ε)**—higher values increase the attack's strength but also make it more noticeable.

2) Projected Gradient Descent (PGD) Method

Type: White-box attack (attacker has full access to the ML model)

How it works:

PGD is an enhanced version of FGSM that applies small FGSM-like perturbations iteratively over multiple steps. The attacker repeatedly refines the adversarial example while ensuring the perturbation remains within a defined limit (epsilon constraint). This makes it harder to detect than FGSM-generated adversarial samples.

D. Attack Strategies

Adversarial Machine Learning (AML) attacks exploit ML models at various stages, affecting both traditional machine learning (ML) and deep learning systems. These attacks can be categorized into different strategies based on the attacker's knowledge, intent, and timing.

1) Black-Box Attacks

The attacker has no knowledge of the model's internal structure (architecture, parameters, or training data). The attack is based on observing only the model's output for specific inputs.

How it works:

The adversary generates adversarial examples using a different model (a substitute model) or by trial and error. The goal is to create inputs that fool the original model without needing direct access to it.

2) White-Box Attacks

The attacker has full access to the model's parameters, architecture, and training data. This allows for highly optimized adversarial attacks.

How it works:

The attacker exploits the model's transparency to craft adversarial inputs. These inputs are slightly modified but force incorrect predictions, effectively bypassing the model's security.

3) Training-Time Attacks (Poisoning Attacks)

The attack happens during the training phase, where the adversary injects malicious data into the training set. The goal is to corrupt the model from the start.

How it works:

The attacker modifies training data so that the model learns incorrect patterns. Infederated learning, attackers can poison model updates before they are sent to the central server.

4) Targeted Attacks

Attackers focus on a specific individual, company, or organization with a clear goal (e.g., stealing data, espionage). These attacks are long-term and carefully planned.

How it works:

Attackers use a combination of malware, phishing, and exploit vulnerabilities to penetrate a specific target. Often seen in Advanced Persistent Threats (APTs), which involve continuous and adaptive hacking over time.
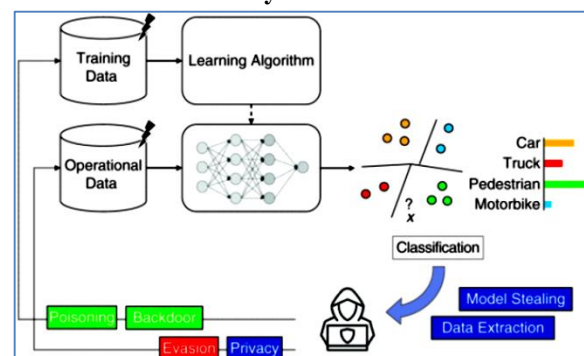
AML Attacks Taxonomy

Fig.4- Major AML Attacks

International Journal of Innovations in Engineering and Science,   www.ijies.net

5) Untargeted Attacks

Attackers aim to infect as many victims as possible, without caring about specific targets. These attacks rely on broad, automated techniques.

Common strategies:

Phishing – Mass emails tricking users into revealing personal information.

Watering Hole Attacks – Hackers infect popular websites to spread malware.

ML Lifecycle and AML Attack Categories

The lifecycle of an ML model consists of two main phases:

Training Phase – The model learns from training data and configurations to produce a trained model.

Operational Phase – The trained model is deployed and actively used. In cases like online learning, where user feedback continuously updates the model, the operational phase loops back into training.

Based on this lifecycle, five key categories of AML attacks exist:

Poisoning Attacks – Attackers manipulate training data to degrade ML performance, either disrupting the model entirely or enabling targeted misclassifications. Such attacks exploit ML's reliance on high-quality data, impacting applications like spam filtering, malware detection, and intrusion detection.

Backdoor Attacks – Attackers embed hidden triggers in training data, causing the model to behave maliciously when activated. For example, a tampered road sign classifier may misidentify stop signs when a specific sticker (trigger) is present

Evasion Attacks – Attackers craft adversarial examples to mislead models into incorrect predictions. Examples include image manipulations that trick facial recognition systems or ML-based security tools. These attacks exploit model limitations and often transfer across similar models.

Model Stealing Attacks – By querying a deployed model, attackers approximate or extract its parameters, leading to intellectual property theft and enabling stronger evasion attacks. Cloud-based ML services are at higher risk.

Data Extraction Attacks – Attackers attempt to reconstruct or identify sensitive training data, such as biometric or medical records, violating data privacy. For instance, a facial recognition system could be exploited to generate an approximate facial image from a name.

III. METHODOLOGY

This section outlines the approach used to evaluate countermeasures against AML attacks and their practical applications in ML environments. The study primarily relied on peer-reviewed academic sources, including journals from IEEE, Elsevier, Springer, ACM, and Taylor & Francis. The research process began with keyword-based searches using terms such as "Evasion," "Poisoning," "Privacy," "Adversarial Machine Learning," and "Machine Learning Attacks. "An initial screening of abstracts and introductions led to the identification of over 500 papers related to ML security. This selection was further refined by focusing on papers that proposed actionable countermeasures for detecting, preventing, or mitigating AML attacks, reducing the count to 363 relevant papers. A final filtering process was conducted based on specific criteria.
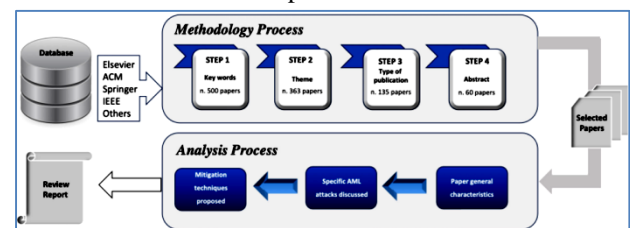

Fig.5- Methodology Process Flow

The Criteria such as practical implementation feasibility and real-world applicability—resulting in 135 shortlisted papers. After a detailed review, only 60 papers that demonstrated tested and implementable countermeasures were included in the final analysis. These countermeasures were categorized, assessed for strengths and weaknesses, and examined for their potential in enhancing ML security.

Additionally, beyond academic literature, the study also reviewed open-source AI/ML frameworks and security tools released by leading Standard Developing Organizations (SDOs) to strengthen AML defense mechanisms.

IV. RESULTS AND DISCUSSION

Results and Discussion: A Synthesis of AML Attack Countermeasures Section V of this review delves into the countermeasures proposed in academic studies to combat adversarial machine learning (AML) attacks. The analysis reveals a range of strategies, each targeting different aspects of AML vulnerabilities. A comprehensive grasp of these countermeasures is of utmost importance to any company or organization looking to enhance the security of the ML models they use.

Countermeasures Against Evasion Attacks

Evasion attacks manipulate input data at the testing phase to cause misclassification. Prominent

International Journal of Innovations in Engineering and Science,   www.ijies.net

countermeasures include: - Adversarial Training**:** This involves training models using both clean & adversarial

Countermeasures Against Poisoning Attacks

Poisoning attacks compromise the training data to degrade model performance. Key strategies include; Data Sanitization: This focuses on cleaning and validating training data to remove potentially malicious samples, ensuring data integrity.

V. DISCUSSION

The countermeasures described in this review demonstrate the proactive measures being developed to defend against AML attacks. The results of a comprehensive review of AML research and mitigation techniques will be of utmost importance as the threat of AI security increases. Future work should focus on real-world implementations and scalability of these countermeasures, as well as improving the balance between robustness, utility, and privacy.Recent Frameworks and Tools for Improving Cyber Resilience Against AML Attacks- These tools provide practical solutions for testing, evaluating, and mitigating vulnerabilities in machine learning models.

VI. CONCLUSION

Adversarial Machine Learning (AML) attacks pose significant risks to AI/ML systems, threatening their integrity, reliability, and safety in critical sectors like healthcare and autonomous transportation. This systematic review identified key attack vectors—evasion, poisoning, and privacy-based attacks—and analyzed their real-world implications, such as manipulated traffic sign detection in autonomous vehicles and misdiagnoses in healthcare systems. The study also highlighted countermeasures like adversarial training, differential privacy, and federated learning, along with open-source tools (e.g., CleverHans, ART, Foolbox) for testing and mitigating vulnerabilities.

Organizations deploying AI systems must prioritize security by integrating these defensive strategies into their ML development lifecycle. The findings underscore the urgent need for robust, adaptive defenses to safeguard against evolving adversarial tactics.

VII. FUTURE SCOPE OF RESEARCH

1. Enhanced Defense Mechanisms:
2. Intersection of AI Security and Privacy:
3. Standardized Benchmarks:
4. Quantum-Resistant Defenses:

Final Remarks

As AI adoption grows, so does the sophistication of AML attacks. Future research must focus on proactive defense paradigms, interdisciplinary collaboration, and standardized tools to ensure resilient AI systems.

REFERENCES

[1] *Adversarial Machine Learning: Attacks from Laboratories to the Real World. Accessed: Dec. 22, 2023. [Online]. Available: https://ieeexplore. ieee.org/document/9426997*

[2] *(2019). Adversarial Machine Learning Against Tesla's Autopilot. [Online]. Available: https://www.schneier.com/blog/archives/2019/04/ adversarial_mac.html*

[3] *G. Apruzzese, H. S. Anderson, S. Dambra, D. Freeman, F. Pierazzi, and K. A. Roundy, ''Real attackers don't compute gradients': Bridging the gap between adversarial ML research and practice,'' 2022, arXiv:2212.14315*

[4] *A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, ''Towards deep learning models resistant to adversarial attacks,'' 2017, arXiv:1706.06083.*

[5] *A. Mitra, A. Jain, A. Kishore, and P. Kumar, ''A comparative study of demand forecasting models for a multi-channel retail company: A novel hybrid machine learning approach,'' Operations Res. Forum, vol. 3, no. 4, p. 58, Sep. 2022.*

[6] *A. Kurakin et al., ''Adversarial attacks and defences competition,'' in Adversarial Machine Learning. Cham, Switzerland: Springer, 2018, pp. 195–231.*

[7] *D. Pavithra, ''A study on machine learning algorithm in medical diagnosis,'' Int. J. Adv. Res. Comput. Sci., vol. 9, no. 4, pp. 42–46, Aug. 2018.*

[8] *M. Elbattah and O. Molloy, Analytics Using Machine Learning-Guided Simulations With Application to Healthcare Scenarios. New York, NY, USA: Taylor & Francis, 2018.*

[9] *M. Hashem Eiza and Q. Ni, ''Driving with sharks: Rethinking connected vehicles with vehicle cybersecurity,'' IEEE Veh. Technol. Mag., vol. 12, no. 2, pp. 45–51, Jun. 2017.*

[10] *K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, ''Robust physical-world attacks on deep learning visual classification,'' in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1625–1634.*