Journal of Research in Science and Engineering (JRSE)
ISSN: 1656-1996 Volume-7, Issue-8, August 2025
DOI: 10.53469/jrse.2025.07(08).14

Adversarial Robustness in AI Authentication Systems: Mitigating Threat Vectors Through Gradient Masking and Ensemble Defense

Abdel Hady

Identity & Fraud Management / Digital Software Engineer Senior Manager / Lead Analyst
Leading Banking Organization, Dallas, Texas, United States

Abstract: ***Adversarial attacks on Artificial Intelligence (AI) and authentication systems have become significant cybersecurity threats, exploiting vulnerabilities in machine learning models and biometric protocols. This paper explores adversarial techniques such as spoofing and perturbation attacks, their impact on AI decision-making, and effective defense strategies like adversarial training and input transformation. Addressing these challenges is critical [1], as there is a growing reliance on AI in sensitive areas like finance and healthcare. The study underscores the need for continued innovation in defensive strategies to mitigate evolving adversarial threats. It also highlights the importance of addressing adversarial attacks to ensure the security and reliability of AI-driven systems in sensitive sectors.***

Keywords: Adversarial attacks, Artificial Intelligence, Authentication systems, Cybersecurity, AI vulnerability

1. Introduction

Developing and designing reliable, secure, and usable user authentication systems has become increasingly important in recent years. One of the main reasons is the implementation and widespread use of online services, such as bank operations, online shopping, or access to personal and professional information. Since this information is sensitive, it has to be protected against unauthorized subjects, and user authentication protocols are commonly applied. Artificial Intelligence (AI) algorithms and Biometrics are tools usually combined to construct these protocols. While the first one allows for the extraction of patterns and structures hidden in the data, the second one represents the characteristics that define how each person is. Therefore, these tools can describe how a specific subject behaves and differentiate it from the rest of the users.

An adversarial attack is a deceiving technique that "fools" machine learning models using a defective input. Adversarial machine learning can cause ML models to malfunction. Its intricate approach sets adversarial AI apart from conventional cybersecurity [2][3] threats. Instead of directly attacking the system infrastructure or exploiting known software vulnerabilities, adversarial AI operates more abstractly. It capitalizes on the very essence of AI, which is to learn and adapt from data, by introducing subtle perturbations that appear innocuous to human observers but confound the AI's decision-making process. Adversarial AI, also known as adversarial attacks or AI attacks, is a facet of machine learning that involves malicious actors deliberately attempting to subvert the functionality of AI systems. Adversarial attacks deceive AI systems, causing them to make incorrect or unintended predictions or decisions. Threat actors introduce attacks in the input data, altering the original data or the AI model by changing the parameters or architecture. Although deep neural networks (DNNs) have succeeded in many tasks, adversarial examples generated by adding slight but purposeful distortions to natural examples deceive them. As AI becomes integral to authentication, used in Biometrics, password validation, CAPTCHA, and behavioral analysis, the implications of these attacks are profound, potentially compromising the security of sensitive data and critical infrastructure. The potential risks of adversarial attacks are significant, and robust defenses are urgently needed.

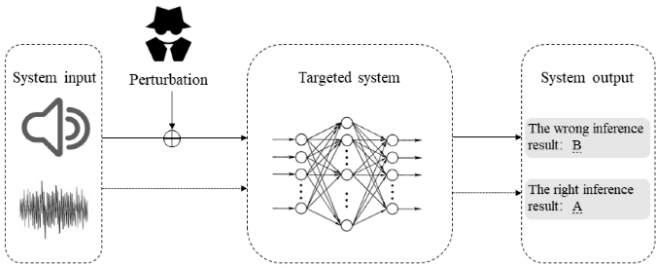

Figure 1: Adversarial attack [4]

This paper provides an in-depth examination of the intersection between adversarial attacks, AI, and authentication systems. It investigates the rise of attacks on behavioral Biometrics and CAPTCHA systems, where AI models fail under adversarial conditions, allowing attackers to automate bypasses or mimic legitimate users. The study reviews existing defenses against adversarial attacks in AI-driven authentication systems, such as adversarial training, which enhances model robustness by including adversarial examples during training, and defensive distillation, which smooths models' decision boundaries to reduce their susceptibility to adversarial inputs. Given adversarial attacks' dynamic and rapidly evolving nature, the paper emphasizes the critical need for adaptive security measures that evolve with AI advancements. Additionally, the research underscores the importance of interdisciplinary approaches, combining insights from machine learning, cryptography, and human-computer interaction to build more secure authentication systems. The paper concludes by outlining future directions for research, such as the development of AI models with intrinsic robustness to adversarial perturbations, the creation of real-time adversarial detection systems, and the potential of

quantum-resistant algorithms to counteract advanced attack strategies.

2. Problem Statement

As AI continues to evolve, so do the tactics of those seeking to exploit it. Adversarial attacks, with their potential to cause significant real-world consequences, exploit the vulnerabilities and limitations inherent in machine learning models and intense neural networks. Popular Adversarial Attack Techniques are [5] [6] [12]:

a) **Fast Gradient Sign Method (FGSM):** Developed by Goodfellow et al., FGSM adds perturbations to the input data in the direction of the gradient of the loss function to create adversarial examples.
b) **Projected Gradient Descent (PGD):** An iterative version of FGSM, PGD applies small perturbations repeatedly to craft more robust adversarial examples.
c) **Carlini & Wagner (C&W) Attack:** This formidable attack optimizes a specific objective function to generate adversarial examples that are harder to detect and defend against, underscoring the need for advanced defense strategies.

Types of Adversarial Attacks that are particularly concerning for Large Language Models (LLMs) [7]:

- **White-Box Attacks:** White-box attackers have complete knowledge of the AI model, including its architecture, parameters, and training data. This knowledge allows for the precise crafting of adversarial examples.
- **Black-Box Attacks:** Black-box attackers have limited information; they rely on querying the model and observing its outputs to create adversarial examples.
- **Evasion Attacks:** Evasion attacks take advantage of characteristics or patterns an AI model has picked up during training. These attacks usually employ optimization techniques to identify the most efficient changes that can mislead the model while still being undetectable by human onlookers. Evasion attacks focus on the model itself. They involve modifying data to seem legitimate but lead to an incorrect prediction. Below are two subtypes of Evasion attacks:
- **Targeted Attacks:** The attacker aims to mislead the model into making a specific incorrect prediction. In targeted evasion attacks, the attacker seeks to force the AI model to produce a particular, predefined, incorrect output. For instance, they might want the model to classify a benign object as harmful, leading to potential security breaches or false alarms.
- **Untargeted Attacks:** The goal is to cause the model to make incorrect predictions. In nontargeted evasion attacks, the goal is to make the AI model produce any incorrect output, regardless of the production.
- **Poisoning Attacks:** In a poisoning attack, the model learns incorrectly because of the training data. Here, the attacker manipulates the training data used to create the model to subtly distort the model's understanding of the underlying patterns in the data.
- **Model extraction attacks:** With model stealing or extraction attacks, attackers aim to learn about the model architecture and parameters. The goal is to replicate the model exactly. This information may lead to a direct financial gain. Model extraction attacks are done by querying the model repeatedly and comparing the input to the corresponding output.
- **Inference attacks:** Inference attacks focus on the data used to train the model. The goal is to extract confidential data from the model. This confidential data can be released directly or inferred from the model's output through carefully crafted queries.

3. Solution

The ways we can defend are as diverse as those that can be attacked. We can adjust training data, the training process, or even the network itself [8]:

Adversarial training: This first approach focuses on the training data. Adversarial training augments the training dataset with adversarial examples, improving the model's robustness.

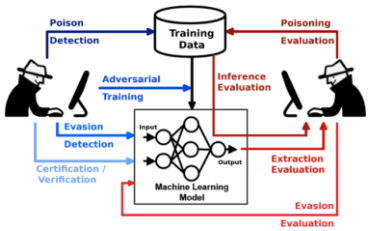

Figure 2: Adversarial attack and Adversarial Training [9]

Adversarial Machine Learning (AML) is an intriguing and rapidly growing field focusing on understanding and defending against adversarial attacks on machine learning models. AML explores how adversaries can exploit machine learning models by introducing subtle perturbations to input data, causing the model to make incorrect predictions. AML addresses the development of attack methods and the creation of robust defense mechanisms. AML is a brute-force solution where we generate a lot of adversarial examples and explicitly train the model not to be fooled by each of them.

Defensive distillation: Defensive distillation involves training a model to mimic the softened output probabilities of another model. We first train a standard model (teacher model) on the original dataset. The teacher model generates the training data's soft labels (probability distributions over classes). Soft labels train student models. The result is a model with smoother decision boundaries more resilient to small perturbations. Distilling the Knowledge in a Neural Network introduced distillation as a technique for model compression, where a small model is created by training to imitate a large one.

Gradient masking: Gradient masking includes a variety of techniques that obscure or hide the gradients of the model. Model switching is a rudimentary gradient masking approach that involves using multiple models within the system to make predictions that are changed randomly. Gradient masking creates a moving target, as an attacker would not know the current model. They will also have to compromise all the models for an attack to be successful.

Input Transformation: Techniques like feature squeezing, input normalization, and randomization can make it more

difficult for adversarial perturbations to affect the model's predictions.

4. Application of the solution in various organization processes

Adversarial attacks have significant implications in various domains. Here's how some organizations can benefit from adversarial learning integration [10]:

Cybersecurity Enhancement: Adversarial machine learning helps improve intrusion detection systems by training models to recognize and counteract sophisticated cyber-attacks that might evade traditional defenses.

Fraud Detection in Finance: Financial institutions apply adversarial machine learning to refine fraud detection algorithms, making them more resistant to tactics employed by fraudsters to evade detection.

Medical Diagnostics: In healthcare [11], adversarial methods enhance diagnostic algorithms, ensuring that medical imaging and data interpretation remain accurate despite potential manipulative inputs.

Retail Recommendation Systems: Retailers leverage adversarial machine learning to secure recommendation engines against manipulative tactics that could skew product suggestions or customer preferences.

Autonomous Driving Safety: Adversarial techniques are employed to improve the reliability of autonomous vehicle perception systems, making them more resilient to attempts to deceive sensors.

Model Integrity Verification: Adversarial methods test the integrity of machine learning models, ensuring they perform consistently and accurately across different scenarios and inputs.

Legal and Compliance Checks: In the legal sector, adversarial machine learning aids in evaluating AI-based decision-making systems to ensure they operate fairly and comply with regulations.

Academic Integrity: Educational institutions use adversarial techniques to enhance plagiarism detection systems and maintain scholastic integrity by identifying and mitigating attempts to game the system.

Manufacturing Quality Control: Adversarial machine learning helps detect anomalies in quality control processes, ensuring that manufacturing systems can effectively identify defects or irregularities.

Government Surveillance: Government agencies apply adversarial methods to secure AI-driven surveillance systems, safeguarding against misuse and ensuring the systems operate as intended.

Fraud Prevention in E-commerce: E-commerce platforms use adversarial learning to enhance security against fraudulent activities, including fake reviews and deceptive practices.

Financial Risk Management: Adversarial machine learning assesses and manages financial risks by testing models against manipulated inputs that could impact financial stability.

Voice Recognition Systems: Adversarial techniques are applied to improve the robustness of voice recognition systems, ensuring they can accurately process speech even when inputs are intentionally altered.

Smart Home Devices: Adversarial machine learning helps secure smart home devices, ensuring that systems such as voice assistants and security cameras are resilient to attempts to deceive or disrupt their functionality.

5. Benefits of solutions

Some of the benefits [12][13] of using different strategies:

Enhanced Model Robustness: Adversarial machine learning helps develop models more resistant to manipulative or deceptive inputs, improving their reliability and stability.

Improved Security: By training models on adversarial examples, organizations can better safeguard against cyber-attacks and malicious attempts to compromise AI systems.

Higher Accuracy in Anomalous Conditions: Adversarial training can make models more accurate in detecting and handling unusual or edge-case scenarios that might otherwise lead to errors.

Increased Fraud Detection: In financial sectors, adversarial methods enhance the ability to detect and prevent fraudulent activities by making fraud detection systems more resilient to evasion tactics.

Better Generalization: Models trained with adversarial examples often generalize better across a broader range of inputs, making them more effective in real-world applications.

Improved Diagnostic Tools: In healthcare, adversarial techniques refine diagnostic algorithms, leading to more accurate medical imaging and better patient care.

Strengthened Privacy Protections: Organizations can implement more robust measures to protect sensitive data by understanding how adversarial attacks can exploit privacy vulnerabilities.

Robustness in Autonomous Systems: Adversarial methods contribute to the safety and reliability of autonomous systems, such as self-driving cars, by making them less susceptible to deceptive inputs.

Resilience in Natural Language Processing (NLP): Adversarial techniques enhance the robustness of NLP models, improving their ability to handle ambiguous or misleading text inputs.

Enhanced Anomaly Detection: Adversarial machine learning improves the ability to detect anomalies and deviations in various systems, from industrial processes to financial transactions.

Better Model Evaluation: Adversarial methods provide a rigorous framework for evaluating model performance, revealing potential weaknesses and areas for improvement.

Greater Adaptability: Models trained with adversarial examples can adapt more effectively to evolving threats and changing input conditions, maintaining their performance over time.

More robust Ethical Safeguards: Adversarial machine learning helps identify and address ethical concerns by ensuring that AI systems operate pretty and are not easily manipulated for harmful purposes.

6. Conclusion

- Adversarial attacks expose critical weaknesses in AI models, demonstrating how minor, intentional perturbations can cause significant misclassifications.
- Utilizing adversarial machine learning helps identify these vulnerabilities by generating adversarial examples to test model robustness.
- Adversarial training, where models are exposed to adversarial examples during training, is a proven technique to enhance model resilience against attacks.
- Other defensive strategies include input preprocessing, which can filter or transform data to mitigate the impact of adversarial inputs.
- Robust optimization methods aim to improve model stability and performance under adversarial conditions, making them less susceptible to manipulation.
- Ensemble methods, which combine multiple models, can reduce the effectiveness of adversarial attacks by diversifying the decision-making process.
- Continuous research and development in adversarial machine learning and defensive techniques are essential for staying ahead of evolving attack strategies.
- By integrating adversarial machine learning with these additional solutions, AI systems can achieve higher security, reliability, and trustworthiness.

In summary, while adversarial attacks expose AI vulnerabilities, employing adversarial machine learning alongside other strategies like adversarial training, input preprocessing, and ensemble methods helps enhance model robustness and security.

References

[1] K. Ren, T. Zheng, Z. Qin, X. Liu, (2020) Adversarial attacks and defenses in deep learning
[2] https://medium.com/@kombib/adversarial-machine-learning-a-comprehensive-guide-for-beginners-58042d12d284
[3] https://www.paloaltonetworks.com/cyberpedia/what-are-adversarial-attacks-on-AI-Machine-Learning
[4] Applied Sciences | Free Full-Text | Adversarial Attack and Defense on Deep Neural Network-Based Voice Processing Systems: An Overview (mdpi.com)
[5] N. Carlini, D. Wagner, (2021). Adversarial Examples Are Not Easily Detected: A Study of Black-Box Attacks and Defenses
[6] H. Zhang, X. Wang, X. Liu, (2024). Advanced Adversarial Attacks and Defenses: A Comprehensive Review
[7] https://www.datacamp.com/blog/adversarial-machine-learning
[8] https://openai.com/index/attacking-machine-learning-with-adversarial-examples/
[9] Adversarial Example Generation with PyTorch | by Akriti Upadhyay | Medium
[10] Y. Liu, J. Ma, Y. Zheng, (2023). Adversarial Machine Learning: Techniques, Applications, and Challenges
[11] L. Wei, K. Ding, H. Hu, (2020). Automatic skin cancer detection in dermoscopy images based on ensemble lightweight deep learning network
[12] B. Goodfellow, I. Shlens, S. Szegedy, (2023). Explaining and Improving the Robustness of Neural Networks Against Adversarial Attacks
[13] C. Chen, S. Zhang, X. Chen, (2023). Practical Adversarial Machine Learning: Techniques and Tools for Defense and Attack