Defending Against Adversarial Attacks in 6G: Practical Mitigation Approach

Sogo Pierre Sanon[†], Akshay Sant [‡] and Hans D. Schotten[*†]
[†]Intelligent Networks Research Group, DFKI, D-67663 Kaiserslautern, Email:sogo_pierre.sanon@dfki.de
[*]Institute for Wireless Communication and Navigation, RPTU, D-67663 Kaiserslautern, Email:schotten@rptu.de
[‡]University of Rochester, Rochester, NY, USA, Email: asant2@ur.rochester.edu

Abstract

The integration of Artificial Intelligence (AI) into mobile networks has led to significant improvements in operational efficiency, resource optimization, and security monitoring. However, the increasing reliance on AI has also introduced vulnerabilities to Adversarial Machine Learning (AML) threats, which are expected to become even more critical with the advent of 6G networks. While numerous AML techniques have been identified, not all pose substantial risks to mobile communication systems. Implementing defenses against all possible attacks can be computationally expensive and may degrade system performance. This study critically examines adversarial threats in AI-driven mobile networks, and attacks are categorized based on their feasibility and real-world impact. A risk-based framework is presented to assist in efficiently prioritizing security investments. The most critical AML threats to mobile networks are identified, and targeted mitigation strategies that ensure a balance between security, computational efficiency, and system reliability are provided. The findings of this research serve as a guideline for AI security implementation in future networks, promoting a strategic approach to adversarial defense while maintaining high network performance.

This is a preprint of the publication which has been presented at the Eighth International Balkan Conference on Communications and Networking

Index Terms

Adversarial Machine Learning, Mobile Networks, 6G security, AI risk mitigation, 5G, 6G

I. Introduction

Artificial Intelligence (AI) is a transformative technology revolutionizing a myriad of domains, including healthcare, finance, and transportation. Mobile communication systems have also embraced this trend, leveraging AI to enhance operational efficiency, optimize resource allocation, and enable advanced features such as real-time anomaly detection and predictive maintenance. With the advent of 6G networks, the role of AI in mobile networks is poised to become even more critical, enabling ultrafast data transfer, seamless connectivity, and intelligent automation across applications such as augmented reality, autonomous systems, and interconnected smart infrastructures [1].

However, the integration of AI into mobile networks introduces significant security challenges. These risks can be categorized into two broad areas: inherent AI risks and adversarial risks. Inherent risks stem from issues like model hallucinations, biases, and flawed implementations, which can lead to erroneous decision-making, inaccurate outputs, or exploitation vulnerabilities. Addressing these requires secure development practices and adherence to frameworks such as the Secure Software Development Framework (SSDF), emphasizing robust and ethical AI deployment across organizational infrastructures.

The second type of risk is adversarial risks, which arise from deliberate, malicious actions by adversaries aiming to compromise, disrupt, or manipulate AI systems for their gain. Adversarial machine learning (AML) is a key concern in this context. AML exploits vulnerabilities in AI models by subtly altering input data to produce incorrect or misleading predictions. In mobile networks, adversarial attacks could compromise authentication systems, disrupt security monitoring, or undermine anomaly detection, potentially causing widespread operational failures and security breaches.

One of the fundamental challenges in addressing these adversarial risks is the lack of information-theoretic security guarantees for AI systems. Unlike cryptographic algorithms that can theoretically offer unbreakable security under ideal conditions, AI systems inherently lack such guarantees. In fact, theoretical results indicate that achieving information-theoretic security for existing AI paradigms is impossible [2]. This highlights a critical gap in the security landscape of AI systems, where vulnerabilities are not just a consequence of flawed implementation but an intrinsic limitation of the technology itself.

Despite advancements in adversarial defenses, current mitigation strategies—such as adversarial training, gradient masking, and robust optimization—often come with trade-offs. These include increased computational costs, reduced model accuracy, and limited scalability in real-world applications. In the high-demand environment of 6G mobile networks, where performance and efficiency are paramount, these trade-offs are particularly challenging to navigate.

This paper aims to differentiate between adversarial attacks that pose real threats to mobile networks and those that have limited practical impact. Instead of treating all AML threats as equally harmful, we conduct a critical assessment of attack

feasibility and impact in the context of mobile networks. The key contributions of this study are:

- Identification of High-Risk Adversarial Attacks: We evaluate the wide range of adversarial attacks in AI-driven mobile networks and determine which ones present realistic, high-impact threats.
- Risk-Based Prioritization for Defenses: By analyzing attack feasibility, we provide a framework to help organizations focus on cost-effective security investments, avoiding unnecessary mitigation strategies for low-risk threats.
- Guidelines for AI Security in Mobile Networks: We outline practical recommendations to ensure robust yet efficient AML defenses, balancing security, computational efficiency, and system performance in 6G networks.

The structure of this paper is as follows: Section II provides an overview of the related work on adversarial machine learning and its application to mobile networks. Section III delves into the foundational concepts of adversarial machine learning, describing attack vectors and vulnerabilities in ML models. In Section IV, we discuss the role and challenges of integrating ML into mobile network systems, focusing on 6G technologies. Section V analyzes adversarial attacks specifically targeting ML-powered systems within mobile networks, offering an in-depth understanding of the threats. Section VI presents practical mitigation techniques, evaluating their effectiveness and feasibility for real-world deployment, and Section VII provides a discussion and recommendations for addressing adversarial risks in mobile networks. Finally, Section VIII concludes the paper with a discussion on implications, key takeaways, and future research directions.

II. Related Work

Recent research on adversarial attacks in mobile and wireless networks has expanded as ML integration within 5G and emerging 6G infrastructures grows. Aminov [3] highlights vulnerabilities in ML models used for 5G network slicing, showing how adversarial attacks—such as FGSM, CW, BIM, and PGD—target models like CNNs, LSTMs, and MLPs. They propose a three-phase defense strategy that incorporates technical robustness and considers societal impacts, underscoring ecological and ethical factors. Likewise, Rifa-Pous et al. [4] discuss AI-driven 6G networks' heightened risks from trust and privacy vulnerabilities due to the disaggregated nature of 6G, calling for specialized defenses to secure these systems.

Sun et al. [5] examine privacy challenges within adversarial machine learning for 6G, noting that while ML can enhance privacy, it also introduces risks of adversarial misuse. They advocate for privacy-centered defenses, such as differential privacy and federated learning, as essential safeguards for 6G networks. Meanwhile, Flowers et al. [6] and Sagduyu et al. [7] explore adversarial attacks on RFML systems and cognitive radio applications. Flowers et al. focus on evasion attacks in spectrum sensing, while Sagduyu et al. investigate interference attacks in 5G functionalities like spectrum sharing, showing that current protocols are vulnerable to spectrum manipulation. Ajayi et al. [8] provide a broader analysis across wireless communication systems, emphasizing that adversarial attacks exploit ML parameters in adaptive networks, stressing the need for robust defenses in self-organizing environments.

This paper proposes an approach to assessing and prioritizing AML threats in AI-driven mobile networks. Unlike previous work, it evaluates adversarial threats based on their real-world feasibility and potential impact.

III. Adversarial Machine Learning

Adversarial attacks in Machine Learning (ML) involve deliberately manipulating input data to cause a model to produce incorrect predictions. The core idea revolves around perturbing an input feature vector x within a constrained norm ε such that the model $f(\cdot)$ produces a misclassification or more generally, an undesired output y'. Formally, the adversary aims to find a perturbation $A(x)$ such that:

$$\|A(x) - x\|_p \leq \varepsilon \quad \text{and} \quad f(A(x)) = y' \neq y,$$

where $\|\cdot\|_p$ represents the p-norm distance metric (typically L_1, L_2, or L_∞), ε is the maximum allowable perturbation magnitude, y is the true class or label, and y' is the adversarially targeted class or label.

The implications of such attacks are far-reaching. For instance, adversarial examples can enable spam or phishing emails to bypass ML-based detection systems, exacerbating cybersecurity risks. Similarly, financial systems relying on ML for fraud detection could be manipulated to classify fraudulent transactions as legitimate, resulting in significant financial losses. Other scenarios include the leakage of personally identifiable information (PII) or the theft of proprietary data, such as labeled datasets or trained models, which represent substantial investments for organizations.

Adversarial attacks can be broadly categorized into three types, each posing distinct risks, namely Poisoning, Inversion and Evasion.

The effectiveness of adversarial attacks depends on the threat model, which defines the attacker's knowledge and capabilities. White-box attacks, the most severe but least practical, assume full access to the model, while grey-box attacks involve partial knowledge, making them a more realistic threat. Black-box attacks, despite limited access, remain effective by refining inputs through iterative queries.

IV. Machine Learning in Mobile Networks

The rollout of 5G mobile networks has marked a significant technological milestone, offering enhanced bandwidth, ultra-low latency, and connectivity for billions of devices worldwide. As the development of 6G networks gains momentum, the nearly 50-year evolution of mobile networks provides invaluable engineering insights. These include advancements in infrastructure design, protocol standards, and algorithmic efficiency, which form a solid foundation for the next generation of mobile technology. Historically, mobile networks have relied on human-designed algorithms which are based

on explicit models of the physics governing network behavior, ensuring optimal performance under predictable conditions.

A. The Shift Towards ML in Mobile Networks

ML and AI represent a paradigm shift in algorithmic design. Unlike traditional methods, which rely on human-crafted logic, ML enables networks to autonomously learn input-output relationships from data. This adaptive approach is particularly valuable in complex, data-driven systems like mobile networks, where conventional methods may face limitations. The adoption of AI is driven by two key motivations:

1) Model Deficiencies: Detecting patterns in noisy, incomplete, or highly variable data is challenging for human-designed algorithms. ML can adapt to evolving network behaviors and diverse user interactions, improving adaptability.

2) Algorithmic Deficiencies: Traditional algorithms often struggle to balance computational speed and decision accuracy. ML enhances real-time decision-making and performance under dynamic conditions, offering a more efficient alternative.

B. Applications and Benefits of ML in 6G Networks

ML introduces new possibilities for optimization, adaptability, and resilience in 6G networks by enhancing performance across various aspects. Hardware acceleration allows for universal hardware designs that support multiple neural networks, reducing the need for reconfiguration and increasing system flexibility. Anomaly detection enables ML algorithms to identify network irregularities, such as security breaches or sudden traffic changes, allowing for proactive responses to maintain network integrity. Additionally, multi-vendor network optimization facilitates seamless cooperation between components from different vendors, improving operational flexibility and integration in increasingly complex telecom ecosystems.

C. Challenges and Risks of ML in Mobile Networks

While ML offers significant potential, several challenges and risks must be addressed to ensure robust and reliable performance. Technical debt arises as AI and ML solutions, while providing immediate benefits, can introduce hidden costs and vulnerabilities over time, including unexpected dependencies that complicate network maintenance and stability. Extrapolation risks occur because ML models rely on historical data and may fail in novel situations, leading to unpredictable behavior. Data collection and processing overhead presents another challenge, as effective ML models require vast amounts of data, necessitating robust infrastructure for seamless collection and real-time processing. Additionally, adversarial machine learning poses a growing threat, where malicious actors manipulate inputs to deceive ML models, potentially causing network disruptions or security breaches. Defenses against such threats include training on diverse datasets, real-time monitoring, and implementing cryptographic security measures.

V. Adversarial Attacks on Mobile Systems

Mobile networks face a diverse range of security threats, from traditional attack vectors to sophisticated ML-based exploits, see for example, Table I. Traditional attacks include Side Channel Attacks that exploit hardware vulnerabilities, Man-in-the-middle (MitM) attacks intercepting communications, Jamming Attacks disrupting signal transmission, Cross-slice Attacks compromising network isolation, hardware-level vulnerabilities through Untrusted Hardware, Denial of Service (DoS) attacks overwhelming network resources, and Brute-force attacks attempting to break encryption. As mobile networks increasingly incorporate ML systems, they become vulnerable to a new class of AML attacks, including Data Poisoning that corrupts training data, Oracle Attacks exploiting model predictions, Gradient-free Attacks that manipulate model behavior without internal knowledge, Evasion Attacks that fool trained models, and Backdoor Attacks that embed hidden vulnerabilities, making the security landscape significantly more complex.

A. AML Attacks and Impact on Mobile Networks

AML poses a major threat to AI-driven 6G networks, targeting key functions like resource allocation, modulation classification, and network security [9]. The L-BFGS method, combined with Generative Adversarial Networks (GANs), generates adversarial datasets that attack AI-based resource management, disrupting intrusion detection and service orchestration. Similarly, the Fast Gradient Sign Method (FGSM) manipulates machine learning outputs in network slicing and resource allocation, affecting network analytics and access management.

Other attacks exploit AI vulnerabilities in different ways. The Jacobian-based Saliency Map Attack (JSMA) manipulates AI transferability to compromise network resource allocation [10], while DeepFool, despite its high complexity, threatens beamforming and signal quality in AI-powered applications. The Zeroth Order Optimization (ZOO) attack disrupts OFDM-based signal detection, increasing bit errors in radio resource allocation [11]. Similarly, Universal Adversarial Perturbation (UAP) degrades Automatic Speech Recognition (ASR) with high success rates, impacting semantic communication [12].

Advanced techniques like adversarial GANs (advGAN) can mislead modulation classifiers, degrading communication reliability [13]. Adversarial Transformation Networks (ATNs) threaten spectrum management in real-time AI applications, while UPSET and ANGRI attacks compromise communication hardware without prior system knowledge. Transfer

TABLE I: Attack Vectors

Generic Attack Vectors	Attack Vectors from ML Models
• Side Channel Attacks • Man in the Middle (MitM) • Jamming Attacks • Cross-slice Attacks • Untrusted Hardware • Denial of Service (DoS) • Brute-force attacks on encryption	• L-BFGS, FGSM • JSMA, Deepfool • ZOO, UAP • advGAN, ATNs • UPSET, ANGRI • DaST, GAP++ • CG-ES

learning models in 6G networks face vulnerabilities from methods like GAP++ and CG-ES, which manipulate cooperative learning, posing risks to applications such as digital twins and autonomous driving. Additionally, the Decision-based Adversarial Sample Transfer (DaST) method highlights the threat of adversarial transferability, allowing attacks even without access to training data.

As AI becomes central to 6G networks, addressing adversarial threats is crucial to ensuring security and stability in future communication systems.

B. Adversarial Defense

To mitigate adversarial risks, a common approach is to develop robust machine learning models that are resistant to manipulation. Robustness in ML refers to the model's ability to maintain accurate predictions even when subjected to adversarial inputs. However, achieving robustness is not without challenges. First, robust models require significant computational resources and expertise to develop. While training a conventional ML model can cost between $40,000$ and $100,000$ [14], robust models can be 100 to $1,000$ times more expensive due to the need for extensive adversarial training and regularization techniques. Additionally, frequent retraining, such as on a quarterly basis, further amplifies these costs. Second, robust models often exhibit lower accuracy on non-adversarial data compared to their non-robust counterparts. This is because non-robust models tend to rely on correlative features that, while useful for prediction, are vulnerable to exploitation. In contrast, robust models sacrifice some predictive power to reduce susceptibility to adversarial manipulation [15].

Given these trade-offs, the pursuit of robustness may only be justified in specific scenarios. For most organizations, the costs of developing and maintaining robust models may outweigh the benefits, particularly when alternative mitigation strategies are available. This leads to the conclusion that maximizing the accuracy of ML models, rather than their robustness, is often the more practical approach for real-world deployments.

VI. Practical Mitigation Techniques

Assessing the risks of AML in mobile networks requires a structured approach to differentiate between realistic threats and theoretical concerns. The lack of real-world datasets and the limited operational deployment of ML models in mobile networks make it difficult to accurately evaluate the impact of AML attacks. Simulated and experimental data often fail to capture the inherent complexity and randomness of real-world mobile environments, while ML models tested under controlled conditions may not fully reflect their vulnerabilities in live deployments. Despite these challenges, it remains crucial to evaluate the risk and impact of AML attacks, ensuring that security efforts are aligned with practical threats rather than hypothetical scenarios. One approach to navigate these limitations, is through the stylized model proposed by Raff et al. [16], which quantifies risk exposure (RE) as the product of an attack's probability and its potential cost. This analytical framework provides a systematic means of evaluating trade-offs between robust and non-robust ML models, allowing security investments to be prioritized based on realistic attack scenarios. By adopting this structured methodology, AML risks can be assessed more objectively, identifying which threats warrant defensive measures and which pose minimal practical risk in mobile networks.

Through their study they found for example, if a model is 95% accurate and 1% of predictions are adversarial, the robust model must maintain at least 94.05% accuracy to justify its use. Their analysis indicates that if the accuracy loss is large, the cost may become negative, making robustness impractical. The analysis suggests that robust models are not cost-effective for most organizations. Robustness is most beneficial when the base model is highly inaccurate, but this also means the robust model will be less effective. Therefore, the decision to adopt robust models should be based on a careful assessment.

Another key consideration is the cost associated with adversarial errors. In certain applications, adversarial errors are significantly more damaging than normal errors. For instance, in financial fraud detection, a model misclassifying a fraudulent loan application as legitimate could lead to severe financial losses, making robust models a worthwhile investment. Furthermore, the frequency of attacks varies across different ML applications. Models trained on publicly available datasets or used internally are generally less exposed to adversarial manipulation than those directly interacting with malicious actors, such as fraud detection or network intrusion systems. This variation in exposure levels underscores the need for a targeted approach to AML security, where defensive strategies are tailored to specific risks rather than applied universally.

A. Risk Analysis Framework

Adversarial attacks can be categorized into four types based on their feasibility and impact [16]. Realistic attacks have a reasonable chance of success and can be carried out with measurable impact, making a robust model an essential defense. Unrealistic attacks, on the other hand, are unlikely to occur unless due to negligence, and the cost of developing a robust model is not justified. Solvable attacks are practical but can be effectively mitigated using existing techniques, without needing a robust model. Finally, Impractical attacks may be possible but require such significant information or resources that the cost to the attacker would be unreasonable, meaning the probability of occurrence is low and a robust model is unnecessary.

B. 6G: Risk Assessment of AML Attack

The adversarial attacks considered in this study were carefully selected based on their prevalence in recent research on AI-driven security threats in 6G networks. We use the attacks categorization from [9] as presented in Table II. While previous studies have identified numerous AML attack vectors in mobile networks [9], our novel contribution is the development of a practical risk assessment framework that categorizes threats based on real-world feasibility rather than theoretical

impact. Building on Raff et al.'s approach [16], we extend their general adversarial risk categorization (Realistic, Unrealistic, Solvable, Impractical) by specifically mapping these categories to 6G network functions and vulnerabilities. Unlike reference [16], which focuses on broad managerial decision-making across all ML domains, our work provides a technical analysis of how each attack type impacts specific mobile network components such as resource allocation, service orchestration, and network slicing.

Our analysis reveals that not all theoretically powerful attacks pose practical threats to mobile networks. Through extensive evaluation of attack vectors against representative 6G network functions. Table II presents our key findings and synthesis at the intersection of adversarial machine learning (AML) and 6G, building upon the current literature, in particular, recent work such as [16]. Furthermore, this categorization is a slight advancement over existing work [9] by highlighting a research gap in existing approaches and evaluating the practicality of each attack tailored to the field of cellular 6G reality and providing targeted mitigation strategies.

1) Realistic: Realistic attacks pose the most immediate danger to AI-driven 6G networks due to their computational efficiency and ability to disrupt essential functions such as resource allocation, service orchestration, and mobility prediction. FGSM and JSMA are particularly concerning, as they can easily mislead CNN-based security frameworks and IoT detection systems, leading to vulnerabilities in network resource management. UAP is highly adaptable, generating perturbations that generalize across different models, making it a severe threat to radio resource allocation and slicing management. Similarly, ATNs and GAP++ can effectively deceive multi-RAT systems and mobility prediction models, undermining AI-based decision-making in next-generation wireless networks. Given their feasibility and impact, these attacks require immediate security interventions to prevent large-scale adversarial exploitation in 6G infrastructure.

2) Unrealistic: Unrealistic attacks, while theoretically powerful, are limited in real-world scenarios due to their extensive resource demands, strict white-box requirements, or lack of targeted precision. DeepFool, for example, is highly effective in controlled environments but lacks the specificity needed for adversarial manipulation in live mobile networks. Similarly, DaST, despite its potential to compromise data analytics and service orchestration, is computationally expensive, making it impractical for large-scale adversaries. These attacks highlight vulnerabilities in AI-driven models but do not represent immediate threats to mobile networks due to their cost and complexity.

3) Solvable: Solvable attacks present a manageable security risk as they can be effectively mitigated using existing AI defense techniques. ZOO, a black-box attack on radio source allocation and load prediction, relies on iterative queries but can be neutralized through query detection and rate limiting. UPSET and ANGRI, which target communication protocols and hardware components, can be mitigated by reinforcing network security protocols. Similarly, CG-ES, which exploits federated learning models, can be countered using secure aggregation and differential privacy techniques [17]. While

TABLE II: Categorization of adversarial attacks based on feasibility and impact in 6G networks.

Attack Method	**Category**	**Attack Type**	**Affected Functions in 6G Networks**	**Reasoning**
L-BFGS	Impractical	White-box	Source management, Service orchestration	Requires full model access and high computational cost, making it difficult to execute at scale.
FGSM	Realistic	White-box	Network access management, Data analytics	Fast and computationally cheap, making it a practical attack against CNN-based access control systems.
JSMA	Realistic	White-box	Network resource allocation	Efficient targeted attacks on IoT detection, disrupting network resource allocation.
Deepfool	Unrealistic	White-box	Resource allocation decisions	Despite being efficient, it is non-targeted and requires model access, limiting its real-world impact.
ZOO	Solvable	Black-box	Radio source allocation, Load prediction	Black-box attack that can be mitigated using query monitoring, reducing its overall effectiveness.
UAP	Realistic	White-box	Radio resource block allocation, Slicing management	Generates broad non-targeted attacks that can efficiently disrupt radio resource allocation.
advGAN	Impractical	White-box	Load prediction, Channel estimation	Requires significant computational resources, making large-scale execution impractical.
ATNs	Realistic	White-box	Service orchestration, Mobility prediction	Computationally feasible targeted attacks that pose a risk to ML-based mobility prediction and orchestration services.
UPSET & ANGRI	Solvable	Black-box	Communication protocols, Hardware components	Security protocols can mitigate these attacks, making them less of an urgent threat.
DaST	Unrealistic	Black-box	Service orchestration, Data analytics	Expensive and complex to execute, making it an unlikely large-scale threat despite its impact on service orchestration.
GAP++	Realistic	White-box	Multi-RAT resource allocation, ML multimodal models	Cheap, fast, and effective at disrupting multimodal ML models used for radio resource allocation.
CG-ES	Solvable	Black-box	Network slicing, Federated learning	Targets federated learning but can be addressed with secure aggregation techniques, reducing its impact.

these attacks can disrupt 6G networks, they can be effectively contained with robust security mechanisms, making them less critical than realistic threats.

4) Impractical: Impractical attacks, despite their theoretical effectiveness, are rarely deployed in real-world adversarial settings due to their excessive computational requirements or reliance on complete model access. L-BFGS, for instance, requires full model knowledge and high processing power, making it unrealistic for adversarial exploitation in dynamic 6G environments. Similarly, AdvGAN, though capable of attacking load prediction and channel estimation models, is highly computationally intensive, limiting its practical use outside of controlled research settings. Given their high cost and complexity, these attacks pose minimal immediate risk to mobile networks, as adversaries would likely opt for more efficient attack vectors.

VII. Discussion and Recommendation

The classification of adversarial attacks on AI-driven 6G networks reveals varying degrees of feasibility and impact across different attack methodologies. As adversarial threats grow more sophisticated, AI risk management is becoming increasingly crucial, particularly with the rise of global regulatory frameworks like the EU AI Act, the NIST AI Risk Management Framework (RMF), and ISO/IEC 42001:2023. These frameworks emphasize the need for robust, transparent, and accountable AI systems.

Strategic security efforts should focus on defending against realistic evasion attacks through adversarial training and anomaly detection. Solvable black-box threats can be mitigated with access control, query restrictions, differential privacy, multi-party computation, and federated learning techniques. However, to optimize resource allocation and achieve economic sustainability in AI-driven 6G networks, organizations must avoid excessive investments in robust AI models designed for unrealistic or impractical threats. Effective risk mitigation must strike a balance between security investments and operational efficiency. While implementing AI security mechanisms incurs costs, these must be justified by the potential risks they address. Investments in adversarial training, model robustness, and secure AI governance must align with long-term economic priorities, avoiding unnecessary financial or computational burdens.

As 6G networks become integral to industries such as smart manufacturing, autonomous transport, and immersive digital services, AI security is critical not only for technical robustness but also for maintaining business competitiveness. Enterprises that proactively implement AI security while ensuring compliance with global regulations will gain a competitive edge, reduce liability risks, and strengthen consumer trust.

VIII. Conclusion and Outlook

This study presents a practical framework for assessing and mitigating AML threats in AI-driven 6G networks. While AI has significantly enhanced operational efficiency and security, it has also introduced vulnerabilities that could compromise network reliability. By categorizing AML attacks based on their feasibility and impact, we identify which threats warrant defensive measures and which pose minimal practical risk. Our findings emphasize that not all theoretically powerful attacks present equal risks in real-world mobile networks. Organizations should prioritize defenses against realistic threats such as FGSM, JSMA, and UAP, which can effectively disrupt resource allocation and service orchestration in 6G systems. For solvable attacks, targeted mitigation strategies like query monitoring and differential privacy offer effective protection without the computational burden of robust models. This risk-based approach to AML defense ensures that security investments are aligned with practical threats rather than hypothetical scenarios, promoting a balanced strategy that maintains both security and performance in next-generation mobile networks.

Acknowledgment

This work has been supported by the Federal Ministry of Education and Research of the Federal Republic of Germany (Förderkennzeichen Open6GHub 16KISK003K). The authors alone are responsible for the content of the paper.

References

[1] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, "The road towards 6G: A comprehensive survey," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 334–366, 2021.

[2] Z. Fang, Y. Li, J. Lu, J. Dong, B. Han, and F. Liu, "Is out-of-distribution detection learnable?" *Advances in Neural Information Processing Systems*, vol. 35, pp. 37 199–37 213, 2022.

[3] M. Aminov, *Adversarial Machine (Deep) Learning-basedRobustification in 5G Networks*, 2023.

[4] H. Rifa-Pous, V. Garcia-Font, C. Nunez-Gomez, and J. Salas, "Security, Trust and Privacy challenges in AI-driven 6G Networks," *arXiv preprint arXiv:2409.10337*, 2024.

[5] Y. Sun, J. Liu, J. Wang, Y. Cao, and N. Kato, "When machine learning meets privacy in 6G: A survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2694–2724, 2020.

[6] B. Flowers, R. M. Buehrer, and W. C. Headley, "Evaluating adversarial evasion attacks in the context of wireless communications," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1102–1113, 2019.

[7] Y. E. Sagduyu, T. Erpek, and Y. Shi, "Adversarial machine learning for 5G communications security," *Game Theory and Machine Learning for Cyber Security*, pp. 270–288, 2021.

[8] O. T. Ajayi, S. O. Onidare, and H. Tajudeen, "AStudy on Adversarial Machine Learning in Wireless Communication Systems," in *International Conference on Computing, Control and Industrial Engineering*, Springer, 2024, pp. 384–392.

[9] V.-T. Hoang, Y. A. Ergu, V.-L. Nguyen, and R.-G. Chang, "Security risks and countermeasures of adversarial attacks on AI-driven applications in 6G networks: A survey," *Journal of Network and Computer Applications*, p. 104 031, 2024.

[10] E. Nowroozi, Y. Mekdad, M. H. Berenjestanaki, M. Conti, and A. El Fergougui, "Demystifying the transferability of adversarial attacks in computer networks," *IEEE Transactions on Network and Service Management*, vol. 19, no. 3, pp. 3387–3400, 2022.

[11] Y. Ye, Y. Chen, and M. Liu, "Multiuser adversarial attack on deep learning for OFDM detection," *IEEE Wireless Communications Letters*, vol. 11, no. 12, pp. 2527–2531, 2022.

[12] Z. Qin, X. Zhang, and S. Li, "Arobust adversarial attack against speech recognition with UAP," *High-Confidence Computing*, vol. 3, no. 1, p. 100 098, 2023.

[13] P. F. de Araujo-Filho, G. Kaddoum, M. Naili, E. T. Fapi, and Z. Zhu, "Multi-objective GAN-based adversarial attack technique for modulation classifiers," *IEEE Communications Letters*, vol. 26, no. 7, pp. 1583–1587, 2022.

[14] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for modern deep learning research," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 13 693–13 696.

[15] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial Examples Are Not Bugs, They Are Features," 2019. arXiv: 1905.02175.

[16] E. Raff, M. Benaroch, and A. L. Farris, "You Don't Need Robust Machine Learning to Manage Adversarial Attack Risks," *(In Submission/Technical Report)*, 2023, 304 Sentinel Dr, Annapolis Junction, MD 20701 USA.

[17] S. P. Sanon, R. Reddy, C. Lipps, and H. D. Schotten, "Secure Federated Learning: An Evaluation of Homomorphic Encrypted Network Traffic Prediction," in *2023 IEEE 20th Consumer Communications & Networking Conference (CCNC)*, 2023, pp. 1–6.