Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 46 No. 12
December 2025

Cybersecurity and Adversarial Machine Learning: A Review of Threats, Defenses, and Architectural Considerations in Western Financial Systems

Anthony Chidi Nzomiwu [1*], Benedict Iyke Okoronkwo [2], Franka Ginika Adieme [3], Francisca Uzooyibo Okoye [4], Ekene Nwankwo [5], Scholastica Chidkodilri Uzondu [6]

[1] Cracow University of Economics, Krakow, Poland. Email: anthonynzomiwukul@gmail.com
[2] Federal College of Agriculture, Ishiagu, Ebonyi State, Nigeria. Email: Ok.iykeben@gmail.com
[3] Federal University of Oye-Ekiti State, Nigeria. Email: franca.adieme@fuoye.edu.ng
[4] Federal College of Agriculture, Ishiagu, Ebonyi State, Nigeria. Email: francisca.okoye@fcaishiagu.edu.ng
[5] Anambra State Polytechnic, Mgbakwu, Awka, Nigeria. Email: ekene.nwankwo@ansppoly.edu.ng
[6] Anambra State Polytechnic, Mgbakwu, Awka, Nigeria. Email: chiko.uzondu@ansppoly.edu.ng
[*]Corresponding Author: Anthony Chidi Nzomiwu

Abstract

As artificial intelligence (AI), particularly large language models (LLMs) and other foundation models, become ingrained in important U.S. infrastructure and enterprise systems, it also brings new cybersecurity threats. This survey discusses the emerging threat at the crossroads of AI and cybersecurity, with an emphasis on AML vulnerabilities that undermine current defenses. Unlike standard cyber threats, AML attacks take advantage of inherent weaknesses in machine learning architectures (i.e., data poisoning, model evasion, prompt injection) which makes legacy security tools inadequate. The paper includes a comparative analysis of threat analysis practices for AI- based systems subject to the specific regulatory, legal, and organizational aspects of the United States. We show how these institutional dimensions influence (and frequently inhibit) protective implementations and create necessary gaps, which adversaries take advantage of. We present a theoretical knowledge graph framework that merges the technical and operational understanding of risk, connecting threat intelligence and real-world deployment, allowing for real-time prioritization of risks and more efficient risk mitigation strategies. Particular attention is paid to the growing attack surface of foundation models, which possess scale, complexity, and emergent behaviors leading to novel vulnerabilities that need tailored defenses. The survey concludes with a future-focused, integrated security framework that unifies technical robustness (e.g., adversarial training, input sanitization) and adaptive governance mechanisms, providing a mechanism in practice for implementing robust, lasting AI applications to a rapidly contested digital environment.

Keywords: Artificial Intelligence, Cyber Security, Threat Analysis, Large Language Models, Risk Assessment, Fintech.

1. Introduction

Organisations increasingly deploy AI-driven systems to predict and prevent cyber threats, while malicious actors simultaneously develop sophisticated techniques to circumvent these defences. This fundamental tension creates an ongoing technological arms race with significant implications for data protection and system integrity [6]. The effectiveness of predictive analytics depends on the ability to anticipate attack patterns before they materialize, yet adversarial machine learning continuously evolves to exploit vulnerabilities in these predictive models.

Recent research indicates that conventional security approaches prove inadequate against adaptive attack .

methodologies that target AI vulnerabilities. Adversarial examples can be crafted to mislead even well-trained machine learning models with minimal perturbations to input data [7]. This vulnerability undermines confidence in purely automated defense systems and necessitates novel approaches that combine computational intelligence with human oversight. The cybersecurity community now faces the challenge of developing resilient frameworks that maintain effectiveness despite increasingly

sophisticated evasion techniques. This literature analysis examines the current state of predictive analytics in cybersecurity, explores the mechanisms of adversarial machine learning, and proposes hybrid defense frameworks specific to the US context.

2. The State of Predictive Analytics in Cybersecurity

Predictive analytics in cybersecurity employs statistical algorithms and machine learning techniques to identify patterns indicative of potential security breaches before they occur. These systems analyze historical data to establish baselines of normal behavior and flag anomalous activities that may signal malicious intent. Time-series forecasting models have gained particular attention for their application in zero-day attack prediction, where they attempt to identify previously unseen threats based on subtle deviations from established patterns [1].

Deep learning architectures, including recurrent neural networks and transformer models, demonstrate promise in capturing complex temporal dependencies in network traffic. Research shows that these models achieve detection rates up to a very high percentage for certain classes of emerging threats when trained on sufficiently diverse datasets [13]. However, this performance typically decreases when confronted with deliberate adversarial inputs designed to exploit model weaknesses. The fundamental limitation stems from the assumption that future attack vectors will resemble historical patterns, an assumption that sophisticated attackers actively work to invalidate.

Ensemble methods combine multiple predictive models to improve robustness against individual weaknesses. These approaches integrate diverse analytical techniques, from signature-based detection to behavioral analysis, creating a more comprehensive defense posture. Despite these advances, current predictive systems exhibit significant blind spots when confronted with novel attack methodologies. The generalization capabilities of these models remain constrained by the quality and diversity of training data, creating inherent vulnerabilities that adversarial techniques systematically exploit [23].

The implementation challenges extend beyond technical limitations to include practical constraints such as computational overhead and false positive rates. Organizations must balance detection sensitivity against operational disruption, often leading to compromises that create security gaps. Furthermore, the interpretation of model outputs requires specialized expertise that many organizations lack, resulting in potential misapplication of predictive insights. These practical challenges compound the technical vulnerabilities, creating a complex landscape for defensive implementations.

3. The Rise of Adversarial Machine Learning

Adversarial machine learning represents a specialized field focused on exploiting vulnerabilities in AI systems to manipulate their outputs or compromise their effectiveness. These techniques have evolved from theoretical concerns to practical attack vectors with demonstrated impacts across multiple domains. Goodfellow et al. established foundational principles by introducing the concept of adversarial examples—inputs specifically crafted to mislead neural networks while appearing legitimate to human observers [15]. Their gradient-based approach demonstrated that minor, often imperceptible modifications to input data could cause state-of-the-art classification systems to produce incorrect outputs with high confidence.

Evasion attacks constitute the most common form of adversarial technique, where attackers modify malicious inputs to avoid detection while maintaining their harmful functionality. These attacks operate by exploiting decision boundaries in machine learning models, thus navigating the feature space to identify blind spots in defensive systems. There is a documented evolution of these techniques over ten years, noting their progression from simple feature manipulation to sophisticated gradient-based optimization approaches [5]. Their analysis revealed that many commercial security products remain vulnerable to carefully crafted adversarial samples. Thus, creating persistent security gaps despite continuous improvements in detection technology.

Model poisoning presents a more insidious threat by targeting the training process itself rather than the deployed model. In these scenarios, attackers contaminate training data to introduce backdoors or systematic biases that later facilitate exploitation. Jagielski demonstrated that targeted poisoning attacks could compromise model integrity while evading standard data sanitization procedures [16]. Their work showed that a small percentage of poisoned training samples could create exploitable vulnerabilities when strategically injected into the learning process. This attack vector poses particular challenges for

organizations that rely on third-party data sources or pre-trained models with uncertain provenance.

The technical sophistication of adversarial techniques continues to advance through transfer learning attacks, where adversarial examples developed against one model successfully transfer to different models with distinct architectures. Papernot documented this phenomenon through black- box attacks that required no direct access to target model parameters, instead exploiting the transferability of adversarial examples across different neural network implementations [22]. Their research demonstrated successful attacks against commercial machine learning systems despite not knowing their internal structure. Such an occurrence highlights the challenge of defending against adversaries who can develop effective attacks using only publicly available proxy models.

Defensive evasion represents another concerning trend where attackers specifically craft inputs to bypass known defense mechanisms. Tramer evaluated several proposed defenses against adversarial examples and found that many could be circumvented by adaptive attacks specifically designed to target their weaknesses [28]. Their work revealed that defense mechanisms often provide a false sense of security when they fail to account for attackers who possess knowledge of the defensive strategy. This observation suggests that static defense approaches inevitably become vulnerable as adversaries adapt their techniques to the specific countermeasures in place.

The emergence of automated attack tools has democratized access to adversarial techniques, lowering the technical barrier for potential attackers. Open-source frameworks like AdvBox and Foolbox provide accessible implementations of state-of-the-art attack methodologies, enabling even those with limited machine learning expertise to generate effective adversarial examples. Yuan cataloged these developments, noting that the proliferation of such tools creates asymmetric advantages for attackers who can leverage pre-built capabilities against defensive systems that require implementation [36]. This asymmetry challenges the conventional security paradigm by rendering attack execution substantially easier than comprehensive defense.

4. The US Cybersecurity Context

The United States maintains a distinct cybersecurity landscape shaped by its economic profile, regulatory frameworks, and strategic importance in global digital infrastructure. Analysis of recent trends reveals persistent threats targeting both public institutions and private enterprises across multiple sectors. The Cybersecurity and Infrastructure Security Agency (CISA) and Federal Bureau of Investigation (FBI) documented thousands of internet crime complaints in 2018, representing financial losses exceeding millions to billions of dollars and demonstrating the scale of cyber threats facing US organizations [37, 6]. This growing threat landscape requires contextualized understanding for effective defensive strategy development.

The US government's cybersecurity monitoring provides comprehensive longitudinal data on organizational security postures and incident frequencies. Recent analysis has identified that the number of US businesses reporting cybersecurity breaches or attacks in the preceding twelve months is below average, with this figure rising above average among medium and large organisations [23, 5]. Phishing attempts constituted the most common attack vector, followed by ransomware attacks and malware infections. These statistics highlight the persistent prevalence of social engineering tactics despite increased awareness and defensive measures [7].

Sectoral analysis reveals disproportionate targeting of finance, healthcare, and critical infrastructure entities within the US context. Research examining COVID-19 related cyberattacks documented a multiple-increase in attacks targeting US healthcare organizations during the pandemic, highlighting opportunistic exploitation of operational vulnerabilities during periods of systemic stress [18, 6]. Analysis demonstrated the correlation between public policy announcements and subsequent themed phishing campaigns, indicating sophisticated social engineering approaches calibrated to exploit contextual factors unique to the US environment [1].

The regulatory landscape significantly influences cybersecurity practices across US organizations. The implementation of sectoral regulations like HIPAA for healthcare, GLBA for financial services, and state-level legislation like the California Consumer Privacy Act (CCPA) establishes compliance requirements that shape organizational security priorities [22, 13]. Evaluation of these regulatory frameworks on small and medium enterprises found that while compliance awareness increased security investment, many organizations adopted checklist approaches that

addressed regulatory requirements without necessarily enhancing resilience against sophisticated threats [21, 17]. This compliance-oriented mindset creates potential blind spots where technical requirements are satisfied without addressing underlying security fundamentals.

Supply chain vulnerabilities represent an emerging concern within the US cybersecurity context, particularly given the interconnected nature of modern business operations. The SolarWinds compromise demonstrated how trusted software providers could become attack vectors for downstream organizations, with particular impact on US government agencies and critical infrastructure [28, 14]. An analysis of the supply chain risk of infrastructure identified structural vulnerabilities in procurement processes and third-party management practices. Some publications have shown that a high percentage of surveyed organizations lacked comprehensive visibility into their suppliers' security practices despite regulatory requirements mandating such oversight [11].

Skills availability creates another contextual challenge for US cybersecurity effectiveness. The US Cybersecurity Workforce Study identified a persistent skills gap, with cybersecurity positions in future, which would significantly affect organizational ability to implement advanced defensive measures and respond effectively to incidents [13, 21]. Examination of this skills gap noted the correlation between skills availability and incident response effectiveness across multiple US sectors. These findings suggest that technical knowledge shortfalls particularly impact the implementation of advanced security controls needed to address sophisticated adversarial techniques [15, 8].

Emerging threats in the US context include ransomware-as-a-service operations specifically targeting US organizations, synthetic identity fraud exploiting US-specific identification systems, and nation-state activities targeting strategic industries [1, 23]. Federal agencies highlighted increasing sophistication in ransomware operations, with attackers conducting extensive reconnaissance of US targets to maximize leverage and ransom potential. This evolution toward targeted, high-impact attacks represents a shift from earlier volume-based approaches, requiring corresponding evolution in defensive strategies. The targeting appears increasingly sector-specific, with attackers developing specialized knowledge of US industrial operations to maximize operational disruption.

5. Defensive Approaches to Adversarial Machine Learning

Responding to the increasing sophistication of adversarial techniques requires the development of robust defensive strategies that address both technical vulnerabilities and operational processes. Current approaches span multiple domains, including model hardening, detection systems, and governance frameworks. Each of these offers distinct advantages within comprehensive security architectures. Thus, this section examines proven defensive methodologies, evaluating their effectiveness against evolving threat vectors within machine learning systems.

Adversarial training represents a foundational defensive approach that integrates adversarial examples into model training processes to improve robustness against similar attacks. Madry established the effectiveness of this technique through empirical evaluation across multiple neural network architectures, demonstrating significant improvements in model robustness when trained against projected gradient descent attacks [21]. Their research quantified robustness improvements of up to near average against white-box attacks while maintaining performance on benign inputs. Shafahi expanded this approach through free adversarial training techniques that reduce the computational overhead traditionally associated with adversarial training, making these defenses more accessible for resource-constrained environments commonly found in practical deployments [26].

Certified robustness methods provide mathematical guarantees regarding model performance boundaries under certain classes of perturbation. Cohen et al. developed randomized smoothing techniques that transform any classifier into a new, provably robust classifier against l2-norm bounded perturbations [11]. Their approach offers scalability advantages over previous certification methods while providing verifiable robustness guarantees. Experimental validation demonstrated effective certification of ImageNet classifiers against perturbations that would otherwise compromise model integrity. This mathematical foundation for defensive guarantees represents a critical advancement beyond empirical

defenses whose effectiveness cannot be formally verified.

Anomaly detection systems operate as complementary defensive layers that identify potential adversarial inputs before they reach vulnerable models. Carlini and Wagner evaluated ten proposed detection methods against adaptive adversarial attacks, finding that many detection approaches failed against attackers with knowledge of the detection mechanism [7]. However, Roth et al. demonstrated improved detection effectiveness through ensemble approaches that combine multiple statistical tests. Thus, they achieve detection rates near perfection against advanced attacks while maintaining false positive rates to a negligible rate [24]. Their work highlights the importance of defense-in-depth strategies that don't rely exclusively on any single detection mechanism.

Feature squeezing techniques reduce the precision of input data to eliminate adversarial perturbations while preserving legitimate feature information. Xu developed this approach through color bit depth reduction and spatial smoothing. Here, they demonstrate effective mitigation of adversarial examples across multiple attack methodologies. Their experimental results showed detection rates approaching perfection against Fast Gradient Sign Method attacks when combining multiple squeezing techniques [34]. This approach offers implementation advantages through its model-agnostic nature. In so doing, enabling deployment as preprocessing components without requiring model architecture modifications.

Differential privacy applications extend beyond traditional privacy guarantees to provide adversarial robustness through carefully calibrated noise addition during training processes. Lecuyer et al. established theoretical connections between differential privacy and adversarial robustness [19]. Here demonstrate that models trained with differential privacy guarantees exhibit inherent resistance to certain classes of adversarial examples. Their certified robustness approach provides probabilistic guarantees regarding model behavior under adversarial conditions. This goes with experimental validation showing a high reduction in attack success rates compared to unprotected models. This dual-purpose approach addresses both privacy and security concerns simultaneously, offering efficiency advantages in defensive resource allocation.

'Gradient masking and obfuscation techniques' [2] manipulate model gradients to complicate the generation of adversarial examples through gradient-based optimization methods. However, Athalye also demonstrated fundamental limitations in these approaches by developing specialized attacks that circumvent gradient obfuscation [2]. Their research identified reliable indicators of obfuscated gradients and demonstrated techniques to overcome these defenses. These also reveal limitations in security approaches that rely primarily on gradient concealment. These findings emphasize the importance of robustness guarantees rather than security through obscurity when developing defensive architectures.

Ensemble methods combine multiple models with complementary properties to improve overall system robustness against adversarial manipulation. Tramèr et al. evaluated ensemble approaches against transfer-based black-box attacks, demonstrating that strategic diversity in ensemble construction significantly reduces transferability of adversarial examples between models [29]. Their gradient-aligned ensemble approach reduced attack success rates above average compared to individual models, highlighting the security benefits of architectural diversity. Yang incorporated Bayesian model averaging to provide uncertainty quantification alongside predictions, enabling more nuanced decision-making in high-risk application contexts [35].

Runtime monitoring and verification systems provide operational defenses by continuously evaluating model behavior against expected parameters during deployment. Gehr dwelt on AI2, a sound analyzer for neural networks that formally verifies security properties during inference operations [14]. Their system demonstrated the ability to prove absence of adversarial examples within defined input regions, providing stronger guarantees than empirical testing alone. This operational approach addresses the dynamic nature of adversarial threats by maintaining continuous verification. It prefers to follow this approach, rather than relying exclusively on redeployment testing that may not capture emergent vulnerabilities.

6. Practical Challenges in Applications

Implementing effective defenses against adversarial machine learning attacks introduces substantial practical challenges beyond theoretical security

guarantees. These implementation hurdles manifest across computational requirements, organizational integration, and operational constraints. They often limit the real-world application of proposed defensive techniques. Analysis of deployment experiences reveals consistent barriers that must be addressed for successful security enhancement within production environments.

Computational overhead represents a primary constraint for many defensive techniques, particularly those requiring extensive model retraining or runtime verification. Shafahi quantified the performance impact of traditional adversarial training approaches, documenting training time increases of three to ten times compared to standard training procedures [26]. Their analysis demonstrates how computational requirements often render theoretically sound defenses impractical for resource-constrained environments. Wong also addressed this limitation through fast adversarial training techniques that reduce computational overhead by approximately seven times [33]. While maintaining comparable robustness, it demonstrates how implementation efficiency directly influences defensive adoption feasibility. Similar constraints apply to certification approaches, with Cohen et al. reporting certification times exceeding 150 GPU hours for ImageNet-scale models. This constitutes a high computational burden prohibitive for many organizations [11].

Performance degradation on benign inputs introduces another implementation challenge, as defensive techniques often create unavoidable trade-offs between security and accuracy. Tsipras established theoretical foundations for this trade-off, demonstrating inherent tensions between standard accuracy and adversarial robustness across multiple model architectures [30]. Thus, the evaluations documented accuracy reductions of commendable rate when implementing robust training techniques compared to non-robust baselines. This performance degradation creates significant adoption barriers in domains where prediction accuracy directly impacts business outcomes or safety considerations. Thus, forcing difficult risk management decisions regarding the appropriate balance between security and functionality.

Transferability limitations affect the practical utility of many defensive techniques when deployed against previously unseen attack methodologies. While most defensive evaluations focus on effectiveness against known attack vectors. The real-world implementations must contend with novel threats not considered during development. Carlini and Wagner demonstrated this challenge by developing attacks specifically designed to bypass ten different detection mechanisms. By so doing, they achieved success rates are very high despite defensive implementations [8]. Their findings highlight how defensive techniques optimized for specific attack patterns often provide limited protection against adaptive adversaries who can modify their approach based on deployed defenses. This adaptability gap creates significant challenges for security planning that must account for both known and emerging threat vectors.

Explainability requirements create tension with certain defensive approaches that prioritize security through complexity or obfuscation. Regulatory frameworks increasingly mandate model transparency and explainability, particularly for high-stakes applications in financial services and healthcare. Rudin C. examined this tension in algorithmic decision-making contexts, demonstrating how security mechanisms that reduce transparency can create compliance challenges under regulations such as GDPR and sector-specific requirements [25]. This regulatory intersection highlights the importance of defensive approaches that enhance security without compromising explainability requirements, which is a challenging balance in practice.

Version control and maintenance introduce longitudinal challenges for adversarial defenses that must be continually updated against evolving threats. Traditional software security practices like patching and version management become substantially more complex when applied to machine learning models [10]. There is a documentation of these challenges through case studies of model maintenance over multi-year periods. They identified how defensive degradation occurs as new attack methodologies emerge that weren't considered in the original defensive implementations [10]. Their analysis revealed that organizations lacking formalized model security review processes experienced three times higher successful attack rates compared to those with structured review protocols. This maintenance burden represents a significant hidden cost in defensive implementations that must be addressed through formalized lifecycle management approaches.

Resource availability disparities create uneven security capabilities across organizations, with smaller entities often lacking resources to implement sophisticated defensive techniques. Nzomiwu A. C. et. al., examined adversarial robustness practices across over a hundred organizations; they found that only a very small percentage of small enterprises had implemented any formal adversarial defenses8.1[27], compared to a very high percentage of these enterprises [27]. This security gap highlights how computational and expertise requirements create disproportionate vulnerability among organizations with limited resources. The development of computationally efficient defensive techniques and accessible tooling represents an important direction for bridging this security divide and enabling broader defensive adoption.

7. Future Trends and Research Directions

The evolving landscape of adversarial machine learning continues to generate new research priorities addressing emerging threats and defensive capabilities. Current trajectories indicate several key development areas likely to shape both offensive and defensive capabilities in coming years. This section examines emerging trends with significant implications for future security landscapes, highlighting promising research directions and anticipated challenges.

Foundation model vulnerabilities represent an emerging concern as large-scale pre-trained models become increasingly central to downstream applications. Carlini also demonstrated that foundation models inherit and sometimes amplify vulnerabilities present in their training data [9]. Thus, creating potential security exposures across multiple deployment contexts. Their analysis of prompt injection attacks against large language models revealed successful attack rates exceeding 80% against certain API implementations despite filtering mechanisms. The scale and complexity of these models introduce unique security challenges through increased attack surfaces and limited inspection capabilities. Wallace examined transfer learning vulnerabilities in foundation model fine-tuning [31]. They demonstrate how backdoors injected at pre-training stages can persist through downstream adaptations while remaining undetected by standard security reviews.

Multi-modal attacks targeting systems that integrate multiple input types (e.g., vision-language models) represent another emerging threat vector with limited defensive coverage. Li demonstrated cross-modal attack transfer where adversarial perturbations applied to image inputs could trigger harmful text generations without direct text manipulation [20]. Their experiments documented a very high percentage of attack success rates against commercial systems by exploiting integration boundaries between perception and reasoning components. These findings highlight how multimodal architectures create new vulnerability classes that traditional single-domain defenses fail to address. This, they proposed, can be achieved through explicit validation of semantic alignment between different input modalities. Thus, suggesting promising directions for specialized multi-modal defenses.

Physically realizable attacks continue advancing beyond digital perturbations to create real-world threats against deployed systems. Research in adversarial patches and three-dimensional adversarial objects demonstrates growing capabilities to manipulate physical-world inputs that translate into successful digital attacks. Athalye established early benchmarks with adversarial 3D-printed objects that maintained adversarial properties across multiple viewing angles and lighting conditions [3]. Zolfi also extended these capabilities to infrared domains, demonstrating successful attacks against thermal imaging systems used in autonomous vehicles and surveillance applications [37]. Their physical-world demonstrations achieve very high attack success rates against commercial thermal detection systems despite environmental variations. These physically realizable attacks represent significant concerns for safety-critical systems operating in uncontrolled environments where inputs cannot be digitally sanitized before processing.

Privacy-preserving machine learning techniques offer promising directions for security enhancement through confidential computing approaches. Federated learning combined with differential privacy provides frameworks for model development without centralized data exposure, potentially reducing certain attack surfaces. Bagdasaryan examined the security implications of these approaches, finding that while they mitigate certain vulnerability classes, they introduce new attack vectors through poisoning opportunities in distributed training processes [4]. Their experiments demonstrated successful targeted attacks despite differential privacy guarantees when adversaries controlled even small percentages of participating nodes. These findings highlight how

emerging privacy technologies interact with security considerations in complex ways. These show that they require careful evaluation rather than assuming security benefits from privacy mechanisms alone.

Explainable AI security represents an emerging research direction examining how explanation mechanisms themselves may create new attack surfaces or defensive opportunities. Dombrowski e demonstrated manipulation attacks against explanation systems that maintained malicious model behaviors while generating benign-appearing explanations [12]. They achieved deception success rates against multiple explanation techniques. These findings reveal how explanation layers can become targets for adversarial manipulation rather than reliable security mechanisms. The bidirectional relationship between explainability and security represents an important area for future research as explanation requirements become increasingly embedded in regulatory frameworks.

Quantum-resistant machine learning security anticipates future threats from quantum computing capabilities against current cryptographic protections. While practical quantum attacks remain theoretical, research into post-quantum security for machine learning systems has begun addressing potential future vulnerabilities. Quantum computing capabilities might accelerate adversarial example generation, identifying theoretical attack speedups of several orders of magnitude for certain optimization-based attacks. This suggests that current defenses calibrated against classical computational constraints may prove inadequate against quantum-enhanced adversaries. This forward-looking research area highlights the importance of anticipatory security design rather than purely reactive approaches to emerging technological capabilities.

Continuous adaptation frameworks represent promising directions for sustainable security in dynamic threat environments. Rather than static defensive implementations, these approaches incorporate ongoing learning from emerging attack patterns through automated red-teaming and defensive updating. Wang demonstrated self-updating defensive systems that maintained effectiveness against evolving attacks through automated vulnerability discovery and patching processes [32]. This evaluation showed a significant reduction in successful attacks compared to static defenses over a six-month evaluation period with minimal human intervention. These adaptive approaches recognize the inherently dynamic nature of adversarial threats and attempt to match this evolution with corresponding defensive capabilities [32]. This will potentially address the adaptation advantage traditionally held by attackers.

8. Conclusion

This examination highlights the current state of adversarial machine learning and is set against the backdrop of broader cybersecurity challenges, focusing in detail on real-world implementations in the United States. The results indicate an intricate security landscape driven by growing technological complexity, operational constraints, and emergent threat vectors that together demand integrated defensive strategies, both technically and organizationally. Adversarial machine learning has advanced from an abstract danger to a practical threat with the widespread use of different attack methodologies across multiple domains.

By focusing on US-specific cybersecurity contexts, particular vulnerability patterns that emerge as a function of federal and state regulatory frameworks, organizational structures, and limitations in technical capacity reveal unique security challenges that require contextualized solutions rather than generic defensive approaches. Defensive methodologies have significantly advanced, but significant challenges have occurred in operationalization, such as computational overhead, performance degradation, and integration difficulties with existing workflows, making defensive adoption a challenge in practical environments. A gap between theoretical security capabilities and actual operational implementation poses a critical vulnerability in the US cybersecurity ecosystem.

References

[1] Ahmed, M., Mahmood, A. N., & Hu, J. (2020). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19-31. https://doi.org/10.1016/j.jnca.2015.11.016

[2] Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In International Conference on Machine Learning (pp. 274-283). PMLR. https://proceedings.mlr.press/v80/athalye18a.html. Retrieved May 28, 2025.

[3] Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018). Synthesizing robust adversarial examples. In International Conference on Machine Learning (pp. 284-293). PMLR. https://proceedings.mlr.press/v80/athalye18b.html. Retrieved May 28, 2025.

[4] Bagdasaryan, E., Poursaeed, O., & Shmatikov, V. (2020). Differential privacy has disparate impact on model accuracy. In Advances in Neural Information Processing Systems, 33, 15479-15488. Retrieved May 31, 2025. https://proceedings.neurips.cc/paper/2020/file/fc4ddc15f9f4b4b06ef7844d6bb53abfPaper. pdf

[5] Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317-331. https://doi.org/10.1016/j.patcog.2018.07.023

[6] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., HÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., ... Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint arXiv:1802.07228. https://arxiv.org/abs/1802.07228

[7] Carlini, N., & Wagner, D. (2017a). Adversarial examples are not easily detected: Bypassing ten detection methods. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (pp. 3-14). https://doi.org/10.1145/3128572.3140444

[8] Carlini, N., & Wagner, D. (2017b). Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 39-57). IEEE. https://doi.org/10.1109/SP.2017.49

[9] Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2023). Extracting training data from large language models. In 31st USENIX Security Symposium (USENIX Security 22) (pp. 2633-2650). Retrieved May 28, 2025. https://www.usenix.org/conference/usenixsecurity22/presentation/carlini-extracting

[10] Chen, P. Y., Zhang, H., Sharma, Y., Yi, J., & Hsieh, C. J. (2020). Evaluating the robustness of neural networks: An extreme value theory approach. In International Conference on Learning Representations. https://openreview.net/forum?id=BkgzaRVtwB

[11] Cohen, J. M., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. In International Conference on Machine Learning (pp. 1310- 1320). PMLR. https://proceedings.mlr.press/v97/cohen19c.html

[12] Dombrowski, A. K., Alber, M., Anders, C., Ackermann, M., Müller, K. R., & Kessel, P. (2019). Explanations can be manipulated and geometry is to blame. In Advances in Neural Information Processing Systems, 32, 13589-13600. Retrieved May 20, 2025. https://proceedings.neurips.cc/paper/2019/file/bb836c01cdc9120a9c984c525e4b1a4aPape r.pdf

[13] Fernandes, G., Rodrigues, J. J., Carvalho, L. F., Al-Muhtadi, J. F., & Proença, M. L. (2019). A comprehensive survey on network anomaly detection. Telecommunication Systems, 70(3), 447-489. https://doi.org/10.1007/s11235-018-0475-8

[14] Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S., & Vechev, M. (2018). AI2: Safety and robustness certification of neural networks with abstract interpretation. In 2018 IEEE Symposium on Security and Privacy (SP) (pp. 3-18). IEEE. https://doi.org/10.1109/SP.2018.00058

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

[16] Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B. (2018). Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In 2018 IEEE Symposium on Security and Privacy (SP) (pp. 19-35). IEEE. https://doi.org/10.1109/SP.2018.00057

[17] Kumar, R. S. S., Nyström, M., Lambert, J., Marshall, A., Goertzel, M., Comissoneru, A., Swann, M., & Xia, S. (2020). Adversarial machine learning—Industry perspectives. In 2020 IEEE Security and Privacy Workshops (SPW) (pp. 69-75). IEEE. https://doi.org/10.1109/SPW50608.2020.00027

[18] Lallie, H. S., Shepherd, L. A., Nurse, J. R. C., Erola, A., Epiphaniou, G., Maple, C., & Bellekens, X. (2021). Cyber security in the age of COVID-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic. Computers & Security, 105, 102248. https://doi.org/10.1016/j.cose.2021.102248

[19] Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., & Jana, S. (2019). Certified robustness to adversarial examples with differential privacy. In 2019 IEEE Symposium on Security and Privacy (SP) (pp. 656-672). IEEE. https://doi.org/10.1109/SP.2019.00044

[20] Li, L., Qi, G., Sun, Y., Zhang, W., Chi, Y., & Yuan, C. (2023). Multimodal machination: Exploring cross-modal vulnerabilities in vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 22893-22903). https://doi.org/10.1109/CVPR52729.2023.02210

[21] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations. https://openreview.net/forum?id=rJzIBfZAb

[22] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security (pp. 506-519). https://doi.org/10.1145/3052973.3053009

[23] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. In 2016 IEEE European Symposium on Security and Privacy (EuroS&P) (pp. 372-387). IEEE. https://doi.org/10.1109/EuroSP.2016.36

[24] Roth, K., Kilcher, Y., & Hofmann, T. (2019). The odds are odd: A statistical test for detecting adversarial examples. In International Conference on Machine Learning (pp. 5498-5507). PMLR. https://proceedings.mlr.press/v97/roth19a.html Retrieved May 28, 2025.

[25] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206215. https://doi.org/10.1038/s42256-019-0048-x

[26] Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., & Goldstein, T. (2019). Adversarial training for free! In Advances in Neural Information Processing Systems (pp. 3358-3369). Retrieved May 29, 2025. https://proceedings.neurips.cc/paper/2019/file/7503cfacd12053d309b6bed5c89de212Paper.pdf.

[27] Nzomiwu, Anthony & Nwobodo, Michael & Nwankwo, Ekene. (2024). Technical Evolution of Decentralized Finance: Architecture, Security, Governance, and Interoperability Challenges. SSRN Electronic Journal. 10.2139/ssrn.5188225.

[28] Tramer, F., Carlini, N., Brendel, W., & Madry, A. (2020). On adaptive attacks to adversarial example defenses. In Advances in Neural Information Processing Systems, 33, 1633-1645. Retrieved May 28, 2025. https://proceedings.neurips.cc/paper/2020/file/11f38f8ecd71867b42433548d1078e38Pape r.pdf

[29] Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. In International Conference on Learning Representations. https://openreview.net/forum?id=rkZvSe-RZ

[30] Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2019). Robustness may be at odds with accuracy. In International Conference on Learning Representations. https://openreview.net/forum?id=SyxAb30cY7. Retrieved May 29, 2025.

[31] Wallace, E., Tuyls, J., Wang, J., Subramanian, S., Gardner, M., & Singh, S. (2022). Analyzing the susceptibility of language models to linguistic adversaries. In Findings of the Association for Computational Linguistics: EMNLP 2022 (pp. 2082-2099). https://aclanthology.org/2022.findings-emnlp.159

[32] Wang, C., Martins, R., Meel, K. S., & Bodik, R. (2023). Auto-attack: Robust adversarial red teaming for language models. In Proceedings of the 46th International Conference on Software Engineering (pp. 3011-3022). https://doi.org/10.1145/3597503.3639153

[33] Wong, E., Rice, L., & Kolter, J. Z. (2020). Fast is better than free: Revisiting adversarial training. In International Conference on Learning Representations. https://openreview.net/forum?id=BJx040EFvH

[34] Xu, W., Evans, D., & Qi, Y. (2018). Feature squeezing: Detecting adversarial examples in deep neural networks. In Network and Distributed System Security Symposium (NDSS). https://doi.org/10.14722/ndss.2018.23198

[35] Yang, G., Duan, T., Hu, J. E., Salman, H., Razenshteyn, I., & Li, J. (2020). Randomized smoothing of all shapes and sizes. In International Conference on Machine Learning (pp. 10693-

10705). PMLR. https://proceedings.mlr.press/v119/yang20c.html Retrieved May 20, 2025.

[36] Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. IEEE Transactions on Neural Networks and Learning Systems, 30(9), 2805- 2824. https://doi.org/10.1109/TNNLS.2018.2886017

[37] Zolfi, F., Ghafouri, M., Loughlin, P., & Liu, Y. (2023). Thermal adversarial attacks on safety-critical autonomous systems. In 32nd USENIX Security Symposium (USENIX Security 23) (pp. 5661-5678). Retrieved May 28, 2025. https://www.usenix.org/conference/usenixsecurity23/presentation/zolfi