VOL. 07 NO. 9,
FEBRUARY, 2025
E-ISSN 3026-958X
P-ISSN 3027-1169

African Scholar Publications & Research International
www.africanscholarpub.com

Journal of Science Innovation & Technology Research (JSITR)

Adversarial Attacks in Cybersecurity: A Machine Learning Perspective

Fatima Rilwan Ododo[1]; & Ridwan Rahmat Sadiq[2]

[1]Department of Computer Science, Montana State University Bozeman, MT 59717, USA. [2]Department of Computer Science, Nasarawa State University Keffi, Nigeria.
Corresponding Author: fatimaododo@montana.edu
DOI: https://doi.org/10.70382/ajsitr.v7i9.031

Abstract

Adversarial machine learning (AML) presents a critical threat to the integrity of machine learning (ML) systems deployed in cybersecurity, where adversarial examples can maintain malicious functionality while evading detection. This literature review synthesizes findings from 35 peer-reviewed sources to investigate the taxonomy, attack strategies, and defense mechanisms associated with AML in cybersecurity domains such as intrusion detection systems (IDS), malware analysis, industrial control systems (ICS), and reinforcement learning in cyber-physical systems. We categorize attacks based on knowledge level, timing, and specificity, and highlight the unique challenges of functionality-preserving adversarial inputs in discrete, protocol-constrained environments. The review further evaluates defensive techniques—including adversarial training, detection frameworks, model hardening, and secure lifecycle integration—and identifies key limitations such as domain-specific overfitting, poor generalizability, and lack of standardized benchmarks. We conclude by advocating for robust, adaptive defenses, attacker-aware datasets, and security-by-design approaches that embed adversarial resilience into the entire ML development lifecycle.

Keywords: Adversarial Machine Learning, Cybersecurity, Evasion Attacks, Intrusion Detection Systems, Malware Detection, Industrial Control Systems,

Model Robustness, Adversarial Training, Functionality-Preserving Attacks, Secure Machine Learning Lifecycle.

Introduction

The growing reliance on machine learning (ML) for cybersecurity applications—ranging from intrusion detection systems (IDS) to malware classification and biometric authentication—has revolutionized threat detection and system resilience (Okoli et al., 2024). These models can learn patterns in large datasets (Kalonde et al., 2024), automate responses to known and unknown threats, and reduce the overhead of manual rule-based detection mechanisms (Xu et al., 2025). However, the very strength of these systems—their ability to generalize from data—has become a critical vulnerability. This weakness has given rise to adversarial machine learning (AML), which leverages deliberately crafted inputs, known as adversarial examples, to fool ML models into making incorrect predictions (Xi, 2020).

Adversarial attacks pose a particularly severe threat in the cybersecurity domain. Unlike attacks in image classification—where altered images may fool a classifier but have minimal real-world implications—adversarial attacks in cybersecurity often aim to preserve malicious functionality while evading detection (Rosenberg et al., 2021). For example, a slightly modified piece of malware must still be executable and dangerous even after the perturbation, which adds complexity to the attack scenario (Rosenberg et al., 2021). Furthermore, attackers in cybersecurity have clear incentives and often real-time feedback loops, making the threat landscape more dynamic and dangerous than in other ML application domains (Xi, 2020).

Recent real-world demonstrations further underscore the threat. In the case of Cylance antivirus software, adversarial perturbations to benign files allowed them to bypass malware detection by altering only non-functional bytecode sections (Rosenberg et al., 2021; Demetrio et al., 2021). Similarly, adversarial network traffic has been crafted to evade even deep learning-based IDS systems, such as those trained on NSL-KDD and CICIDS2017 datasets (Xu et al., 2025; Saini et al., 2024).

There are also significant domain-specific challenges that differentiate cybersecurity-focused AML research from other domains. Many cybersecurity features—like API calls, packet headers, and opcode sequences—are discrete and semantically critical, making it difficult to apply common image-domain techniques like gradient perturbations

without breaking functionality (Ahsan et al., 2022; Rosenberg et al., 2021). Additionally, adversarial examples in cybersecurity must comply with protocol constraints, system policies, and functionality preservation, all while remaining undetected by both human analysts and automated systems (Li, 2024; McCarthy et al., 2022).

Despite growing attention, several challenges persist. Defense mechanisms such as adversarial training, input preprocessing, and anomaly detection often fail to generalize across domains, remain expensive to deploy in real-time systems, or can be bypassed by adaptive adversaries (Alobaid et al., 2025; Xu et al., 2025). Moreover, the lack of standardized benchmarks and publicly available functionality-preserving adversarial datasets hinders consistent evaluation of proposed defenses (Ibitoye et al., 2019).

This paper addresses these gaps through a structured literature review focused on the following objectives:

1. To categorize adversarial attacks in cybersecurity using a unified taxonomy based on threat model, knowledge type, and attack strategy.
2. To examine domain-specific impacts of AML across malware detection, biometric authentication, network traffic analysis, and cyber-physical systems.
3. To synthesize the strengths and limitations of existing defense strategies including adversarial training, detection filters, and secure development integration.
4. To identify open challenges and propose directions for future work, including real-world benchmarks, cross-domain defense strategies, and lifecycle-based robustness evaluation.

In doing so, we aim to provide researchers and practitioners with a comprehensive foundation for developing secure, robust, and future-proof ML models in cybersecurity.

Methodology

This literature review followed a structured and systematic approach to ensure comprehensive coverage of adversarial machine learning (AML) research within the cybersecurity domain. Multiple academic databases were searched, including IEEE Xplore, ACM Digital Library, Elsevier ScienceDirect, SpringerLink, arXiv, and MDPI. These platforms were chosen for their relevance to computer science, information security, and machine learning research.

The search strategy involved specific keywords designed to capture a wide range of studies related to adversarial attacks and defenses in cybersecurity. The primary search

terms included: “Adversarial Machine Learning,” “Adversarial Attacks in Cybersecurity,” “Evasion Attacks ML IDS,” “Adversarial Malware Detection,” and “AML defenses cybersecurity.” These terms were selected to encompass both theoretical foundations and practical implementations of AML in cybersecurity contexts.

To maintain the quality and relevance of the review, strict inclusion criteria were applied. Only peer-reviewed articles published between 2018 and 2024 were considered. The review prioritized scholarly work that addressed AML applications in cybersecurity, including but not limited to survey papers, case studies, experimental evaluations, and conceptual frameworks.

After removing duplicates and screening abstracts for relevance, a final corpus of 35 papers was selected. These included several high-impact publications such as articles from ACM Computing Surveys, IEEE Transactions on Information Forensics and Security, and other reputable journals. Each paper was reviewed in full and analyzed for its contribution to understanding adversarial threats, defense strategies, system vulnerabilities, or evaluation frameworks.

The selected articles were thematically categorized under four primary areas: (1) taxonomy and threat models, (2) domain-specific applications (e.g., malware, IDS, ICS), (3) defense mechanisms and mitigation techniques, and (4) existing limitations and research gaps. This thematic structure formed the basis of the review’s analysis and presentation.

Taxonomy of Adversarial Attacks in Cybersecurity

Adversarial attacks against machine learning models in cybersecurity can be categorized across multiple dimensions that reflect the attacker’s knowledge, timing, goals, and application domain. Understanding these dimensions is crucial for designing effective countermeasures.

Attack Knowledge

Adversarial attacks differ significantly based on what the attacker knows about the target model:

- **White-box attacks** assume full knowledge of the model, including its architecture, parameters, and training data (Ren et al., 2020). This enables attackers to compute gradients and craft adversarial samples using powerful methods like FGSM, JSMA, or Carlini & Wagner (C&W) attacks (Xi, 2020).
- **Black-box attacks** occur when the attacker has no internal access to the model (Cui et al., 2020). Instead, they use input-output queries to build a substitute

model or apply transferability principles (Hodes et al., 2024). Techniques such as OnePixel or Zeroth-Order Optimization (ZOO) exemplify black-box strategies (Alotaibi and Rassam, 2023).

- **Gray-box attacks** lie between white-box and black-box. The attacker may know the model architecture but not the parameters or may have access to only part of the training data (Lin et al., 2021). Gray-box settings often occur in real-world cybersecurity scenarios like malware analysis, where attackers guess or learn some model traits through probing (Alotaibi and Rassam, 2023).

Attack Timing

This category differentiates whether the attack affects the model during training or inference:

- **Evasion attacks** modify input data at test time to bypass detection (Girhepuje et al., 2024). In cybersecurity, this can involve altering malware binaries, spoofing network traffic, or modifying API call sequences to be misclassified as benign (Xi, 2020).
- **Poisoning attacks** target the training data. Attackers inject carefully crafted samples into the training set to corrupt the model's behavior or implant backdoors (Zhao et al., 2025). In IDS, poisoning can cause the system to misclassify malicious traffic as benign in future sessions (Xi, 2020).
- **Model extraction and inversion attacks** attempt to reconstruct the model or its training data. For example, APIs can be exploited to approximate decision boundaries or infer sensitive attributes from biometric models, posing serious privacy risks (Chakraborty et al., 2021).

Specificity and Targeting

This dimension considers whether the adversary has a specific goal in misclassification:

- **Targeted attacks** aim to force the model to predict a specific incorrect label (Ododo and Addotey, 2025a). For example, a malware file could be designed to appear exactly like a benign program (Rosenberg et al., 2021).
- **Indiscriminate attacks** simply aim to degrade the model's performance overall, increasing misclassification rates without targeting a particular class (Ododo and Addotey, 2025a). This is common in denial-of-service-style AML attacks (Rosenberg et al., 2021).

In binary classification settings like malware detection or spam filtering, targeted and indiscriminate attacks may functionally overlap since forcing misclassification into the only other class always achieves the attack objective.

Domain Context and Application Targets

Adversarial attacks manifest differently across cybersecurity domains, depending on data structure, constraints, and the type of ML system deployed.

- Network Traffic / Intrusion Detection Systems (IDS): Adversaries craft packets with modified headers, payload sizes, or timing patterns to bypass anomaly or signaturebased detection. GAN-based attacks (e.g., IDSGAN) and JSMA have been used to deceive DNN-based IDS trained on datasets like CICIDS2017 (Alotaibi and Rassam, 2023).
- Malware Detection: Attacks often preserve the executable's functionality while altering features used by classifiers (e.g., injected benign strings or added API calls). Gradient-based attacks and byte-level perturbation are common here (Alotaibi and Rassam, 2023).
- IoT and Industrial Control Systems (ICS): In these domains, perturbations must be stealthy and lightweight due to resource constraints. Adversarial examples targeting ICS can modify sensor data in real-time, potentially leading to physical system damage if left undetected (Anthi et al., 2021).
- Biometric Systems and Authentication: In biometric security, adversarial perturbations to facial images, fingerprints, or speech patterns can bypass authentication or impersonate authorized users. These attacks often involve model inversion or data reconstruction (Chakraborty et al., 2021).

Application Domains and Case Studies

Adversarial machine learning has introduced new security risks across various cyber-defense systems. These attacks vary in method and impact depending on the target domain, and have been extensively studied in four critical areas: intrusion detection systems (IDS), malware detection, industrial control systems (ICS), and multi-agent reinforcement learning in cyber-physical systems. This section synthesizes literature findings from each domain.

Intrusion Detection Systems (IDS)

Intrusion Detection Systems (IDS) are vital in detecting unauthorized access and network anomalies. ML-based IDS, such as those using deep neural networks (DNNs), support vector machines (SVMs), or decision trees, have become increasingly popular

for their high accuracy and adaptability to evolving threats (Alotaibi and Rassam, 2023). However, these systems are highly vulnerable to adversarial evasion attacks, especially when adversaries can probe the model using black-box or gray-box techniques (Alotaibi and Rassam, 2023).

Attacks such as FGSM, JSMA, DeepFool, and PGD have demonstrated effectiveness in misleading IDS on datasets like NSL-KDD and CICIDS2017 (Alotaibi and Rassam, 2023). For example, adversaries can craft adversarial network traffic by perturbing features like packet timing, header sizes, and byte sequences, causing classifiers to label them as benign (Alotaibi and Rassam, 2023).The use of generative adversarial networks (GANs), such as IDSGAN, further enhances evasion success, producing synthetic traffic that bypasses detection (Yan et al., 2022).

Malware Detection

Malware detection is another critical domain where ML models—especially static and dynamic classifiers—are frequently attacked using functionality-preserving adversarial examples. Static detection relies on file attributes like bytecode, opcode sequences, and imported functions, while dynamic detection analyzes behaviors such as API call patterns and memory use (Ibitoye et al., 2019).

Evasion attacks in this domain include byte padding, code injection, and control-flow obfuscation, all of which maintain malicious behavior while avoiding detection (Yan et al., 2022). Techniques such as MalGAN and GAPGAN utilize GANs to generate adversarial binaries, significantly reducing detection rates by modern classifiers (Ibitoye et al., 2019). Other models like EvadeDroid and AdvAttack manipulate Android malware by iteratively injecting benign features or altering key API calls (Ibitoye et al., 2019).

These attacks are not theoretical. In some cases, detection systems like MaMaDroid, DREBIN, and Sec-SVM experienced evasion rates exceeding 70% (Ibitoye et al., 2019). As adversarial-malware-as-a-service platforms emerge, the ease of generating such evasive malware is becoming an operational concern (Anthi et al., 2021).

Industrial Control Systems (ICS)

Industrial Control Systems (ICS), used in smart grids, manufacturing, and water treatment plants, are increasingly adopting ML-driven IDS solutions. These systems often use supervised learning models like Random Forest, J48, and LSTM for anomaly detection in sensor data and control commands (Anthi et al., 2021).

Adversarial attacks in ICS typically manipulate sensor readings or communication signals to cause misclassification without interrupting operations. Techniques like

JSMA have successfully altered input signals to evade IDS while preserving functionality, even leading to safety hazards like pressure misreporting or actuator delays (Anthi et al., 2021). These attacks are particularly dangerous because many ICS environments rely on legacy hardware and operate in real time, limiting the capacity to update defenses (Anthi et al., 2021).

Experimental work by Erba et al. demonstrated real-time evasion attacks against RNNbased detectors using autoencoders, while other studies reported 6–11% accuracy degradation in classic classifiers under adversarial stress (Anthi et al., 2021).

Multi-Agent Systems and Reinforcement Learning

Reinforcement learning (RL) agents deployed in cyber-physical systems (CPS)—such as autonomous vehicles, smart grids, and industrial robots—have shown vulnerabilities to adversarial policies. These attacks manipulate either the agent's inputs or reward structures to mislead its learned behavior (Standen et al., 2025).

Studies by Lee et al. and Standen et al. introduced adversarial tactics against deep reinforcement learning (DRL) agents using spatiotemporal constraints and action-space poisoning (2025). These methods can cause agents to learn unsafe or suboptimal behaviors, especially in cooperative or multi-agent settings like swarm robotics or distributed control systems (Standen et al., 2025).

Attacks like Functional Adversarial Policies (FAP) and adversarial cheap talk have also proven effective in reducing the trustworthiness of agent-based communication and planning strategies (Standen et al., 2025). Research on STARCRAFT multi-agent environments and CybORG simulations highlights the real-world feasibility of these attacks (Standen et al., 2025).

Defense Mechanism

As adversarial attacks become more sophisticated, researchers have developed various defense mechanisms to secure machine learning (ML) systems in cybersecurity. These defenses span training, detection, architectural enhancements, and lifecycle integration. Despite promising advances, most methods still face trade-offs between robustness, scalability, generalization, and computational efficiency.

Adversarial Training

Adversarial training is one of the most commonly employed methods to improve model robustness (ODODO and ADDOTEY, 2025b). It involves incorporating adversarial samples—crafted using methods like FGSM or PGD—into the training process to help the model learn to resist perturbations (Zhou et al., 2022). Notable

strategies include standard min-max optimization, ensemble adversarial training, and curriculum adversarial training. These approaches have shown strong defense capabilities, especially in image and malware detection tasks (Xi, 2020) (Ibitoye et al., 2019).

However, adversarial training suffers from significant limitations. It is computationally expensive due to the iterative generation of adversarial samples, especially for large-scale datasets like CIFAR-100 or NSL-KDD (Zhou et al., 2022). Additionally, models trained with adversarial examples may overfit to specific perturbation types and generalize poorly to unseen attacks (Ibitoye et al., 2019). Recent improvements, such as Triplet Loss regularization and Latent Adversarial Training, aim to address these concerns by enhancing latent space robustness (Zhou et al., 2022).

Detection and Filtering

Detection-based defenses aim to identify adversarial inputs before they reach the model. Notable strategies include feature squeezing and autoencoder-based detection. Xu et al. introduced feature squeezing by reducing input variability through bit-depth reduction and spatial smoothing, then comparing outputs between squeezed and original samples to flag adversarial inputs (2025). Meng and Chen proposed the MagNet framework, which employs multiple autoencoders and divergence detectors to reform or reject suspicious inputs (2022).

These methods are generally lightweight and easy to implement, but they often suffer from high false positives or can be bypassed by adaptive attacks that anticipate the detector's behavior. Furthermore, detection defenses are more effective in controlled settings than in complex, real-world cybersecurity environments such as malware traffic or ICS logs.

Model Hardening

Model hardening involves architectural or algorithmic changes that make models more resilient to adversarial manipulation. One popular method is gradient masking (Apruzzese et al., 2020; Garba et al., 2019), where gradients are hidden or obfuscated to prevent attackers from computing effective perturbations (Wang et al., 2023). However, this technique often fails against black-box attacks or transfer-based attacks and may degrade model performance.

Other strategies include defensive distillation, where soft labels from a teacher network are used to train a smaller student network (Wang et al., 2023), which smooths decision boundaries and reduces vulnerability to attacks. Ensemble learning

approaches—training multiple diverse models—can also increase robustness by reducing the chance that a single attack is universally effective across models (Wang et al., 2023).

Despite their promise, model hardening techniques often lead to reduced clean-data accuracy or increased model complexity. As such, their deployment must balance robustness with performance, especially in real-time systems.

Secure ML Lifecycle Integration

Most current defenses are post hoc solutions rather than proactive security measures embedded in the ML development process. To address this, researchers and industry leaders advocate for integrating adversarial resilience into the Secure Development Lifecycle (SDL) (Olutimehin et al., 2025). This approach involves testing for adversarial vulnerabilities at every stage of ML system development—design, training, deployment, and monitoring (Kumar et al., 2020).

Industry case studies reveal that many organizations lack structured processes to secure ML systems from adversarial threats. A survey of 28 companies across sectors found that while SDL is widely known in software engineering, its adoption for ML systems is still limited. Only a few organizations conducted adversarial testing before deployment (Kumar et al., 2020).

Lifecycle-oriented frameworks propose systematic adversarial threat modeling, curated repositories of known AML attacks (akin to the MITRE ATT&CK framework), and regular security assessments of deployed models. These practices mirror traditional secure coding standards and represent a promising direction for more resilient ML cybersecurity systems (Kumar et al., 2020).

Challenges and Open Problems

Despite the progress in defending against adversarial attacks, several fundamental challenges continue to hinder effective and scalable adversarial machine learning (AML) defenses in cybersecurity contexts. These open problems span dataset realism, evaluation frameworks, domain transferability, and attack realism.

Lack of Realistic Datasets

A major challenge in evaluating AML defenses is the scarcity of realistic, diverse, and up-todate datasets (Eghaghe et al., 2024). Most AML research in cybersecurity uses outdated or synthetic datasets such as NSL-KDD and CICIDS2017, which fail to reflect current attack patterns, protocols, and network behaviors (He et al., 2023). These datasets often lack the richness of real-world traffic, contain imbalanced class

distributions, or are gathered in controlled environments that do not simulate operational variability (He et al., 2023).

Moreover, privacy concerns and legal restrictions make it difficult to publicly release datasets containing real user traffic or sensitive operational logs (Khaleel et al., 2024). As a result, most defense strategies are not evaluated in production-grade scenarios, making their real-world effectiveness uncertain. Future research should explore the application of federated learning and transfer learning techniques to build realistic, privacy-preserving datasets while maintaining generalizability (He et al., 2023).

Functionality-Preserving Evaluation

A core limitation of current AML defenses is their reliance on threat models adapted from computer vision, where adversarial perturbations are designed to be imperceptible to humans (Mintoo et al., 2024). This paradigm is inappropriate for cybersecurity applications, where attackers must preserve the functionality of the malware, payload, or attack traffic.

Studies by Demetrio et al. and Labaca-Castro et al. emphasize that modifications to portable executable (PE) files or API call sequences must not corrupt the binary or disable malicious behavior (2021). In IDS contexts, perturbations must maintain protocol compliance, timing structure, and semantic behavior to avoid detection while still executing the intended attack.

This requirement adds constraints not present in image or text domains. Yet, many defenses continue to evaluate against adversarial samples generated without these constraints. This mismatch leads to inflated defense scores and undermines progress toward deployable solutions.

Generalizability Across Domains

Defenses that perform well in one cybersecurity application (e.g., NIDS) often fail when transferred to others (e.g., ICS, malware detection). This lack of generalization stems from the domain-specific nature of data formats, constraints, and operational contexts.

For example, a defense designed for packet-level feature perturbation in an enterprise IDS may be irrelevant for binary feature modification in Android malware detection. Likewise, defenses tailored for static malware detection may not translate to ICS, where sensor readings and actuation control loops follow temporal and physical laws. This challenge necessitates the development of adaptive, modular, and domain-aware AML defense architectures that can accommodate variability in feature types and

threat models. A meta-learning approach to adversarial robustness—where models adapt across task boundaries—remains a promising direction.

Lack of Standardized Benchmarks and Metrics

There is currently no unified framework for evaluating adversarial robustness in cybersecurity ML models. This has led to fragmented research outputs, inconsistent performance reporting, and difficulty in comparing defenses under standardized conditions.

Benchmarks used in CV and NLP (e.g., ImageNet, GLUE) offer robust baselines, but cybersecurity lacks such reference datasets, evaluation protocols, and scoring rubrics. Furthermore, widely used metrics like accuracy and F1-score are ill-suited to adversarial contexts, especially when datasets are imbalanced or perturbations are constrained by functionality.

Emerging metrics such as CLEVER, empirical robustness, and adversarial risk surfaces should be explored and adapted to the cybersecurity setting. There is also a need for benchmarking platforms and shared leaderboards to drive reproducibility and comparative analysis.

Conclusion and Future Work

This literature review has explored the landscape of adversarial machine learning (AML) in cybersecurity, highlighting the growing threat of adversarial attacks on ML-based systems such as intrusion detection, malware analysis, and industrial control networks. While progress has been made in categorizing attacks and developing defensive strategies—such as adversarial training, input filtering, and model hardening—most existing approaches remain domain-specific, computationally expensive, and difficult to generalize. Key challenges identified include the lack of realistic, attacker-aware datasets, the failure to evaluate functionality-preserving adversarial inputs, and the absence of standardized benchmarks for fair comparison. Furthermore, AML defense remains largely reactive, with minimal integration into the broader machine learning development lifecycle. Future research should focus on developing realistic datasets and adversarial testbeds, designing generalizable and cross-domain defense frameworks, embedding AML resilience into secure ML pipelines, and establishing common evaluation protocols. Addressing these challenges will be crucial in building trustworthy and robust AI-driven cybersecurity systems capable of resisting evolving, intelligent threats.

References

Abayomi Titilola Olutimehin, Adekunbi Justina Ajayi, Olufunke Cynthia Metibemu, Adebayo Yusuf Balogun, Tunboson Oyewale Oladoyinbo, and Oluwaseun Oladeji Olaniyi. Adversarial threats to ai-driven systems: Exploring the attack surface of machine learning models and countermeasures. *Available at SSRN 5137026*, 2025.

Abdul Awal Mintoo, Ashrafur Rahman Nabil, Md Ashraful Alam, and Imran Ahmad. Adversarial machine learning in network security: A systematic review of threat vectors and defense mechanisms. *Innovatech Engineering Journal*, 1(01):80–98, 2024.

Afnan Alotaibi and Murad A Rassam. Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense. *Future Internet*, 15(2):62, 2023.

Ahmad Alobaid, Talal Bonny, and Maher Alrahhal. Disruptive attacks on artificial neural networks: A systematic review of attack techniques, detection methods, and protection strategies. *Intelligent Systems with Applications*, page 200529, 2025.

Aliyu Garba, Sandip Rakshit, and Fatima Rilwan. Detection and sentiment analysis of hate speech on twitter in nigerian politics. In *Proceedings of: 2nd International Conference of the IEEE Nigeria*, page 285, 2019.

Andrew McCarthy, Essam Ghadafi, Panagiotis Andriotis, and Phil Legg. Functionalitypreserving adversarial machine learning for robust classification in cybersecurity and intrusion detection domains: A survey. *Journal of Cybersecurity and Privacy*, 2(1):154–190, 2022.

Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1):25–45, 2021.

Bowei Xi. Adversarial machine learning for cybersecurity and computer vision: Current developments and challenges. *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(5):e1511, 2020.

Eirini Anthi, Lowri Williams, Matilda Rhode, Pete Burnap, and Adam Wedgbury. Adversarial attacks on machine learning cybersecurity defences in industrial control systems. *Journal of Information Security and Applications*, 58:102717, 2021.

FATIMA ODODO and NICHOLAS ADDOTEY. Advancements and challenges in deep learning for cyber threat detection. *International Journal of Science Research and Technology*, 2025a.

FATIMA ODODO and NICHOLAS ADDOTEY. Understanding the influence of outliers on machine learning model interpretability. *International Journal of African Sustainable Development Research*, 2025b.

Gilbert Kalonde, Samuel Boateng, Lateefat Sanni, Silas Chotwe, and Fatima Ododo. Artificial intelligence and special education: The use and the integration. In *Society for Information Technology & Teacher Education International Conference*, pages 1926–1932. Association for the Advancement of Computing in Education (AACE), 2024.

Giovanni Apruzzese, Mauro Andreolini, Michele Colajanni, and Mirco Marchetti. Hardening random forest cyber detectors against adversarial attacks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 4(4):427–439, 2020.

Ishai Rosenberg, Asaf Shabtai, Yuval Elovici, and Lior Rokach. Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)*, 54(5):1–36, 2021.

Jian Xu, Yong Wang, Heyao Chen, Zepeng Shen, et al. Adversarial machine learning in cybersecurity: Attacks and defenses. *International Journal of Management Science Research*, 8(2):26–33, 2025.

Jing Lin, Long Dang, Mohamed Rahouti, and Kaiqi Xiong. Ml attack models: Adversarial attacks and data poisoning attacks. *arXiv preprint arXiv:2112.02797*, 2021.

Ke He, Dan Dongseong Kim, and Muhammad Rizwan Asghar. Adversarial machine learning for network intrusion detection systems: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 25(1):538–566, 2023.

Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. Adversarial attacks and defenses in deep learning. *Engineering*, 6(3):346–360, 2020.

Li Li. Comprehensive survey on adversarial examples in cybersecurity: Impacts, challenges, and mitigation strategies. *arXiv preprint arXiv:2412.12217*, 2024.

Luca Demetrio, Battista Biggio, Giovanni Lagorio, Fabio Roli, and Alessandro Armando. Functionality-preserving black-box optimization of adversarial windows malware. *IEEE Transactions on Information Forensics and Security*, 16:3469–3478, 2021.

Maxwell Standen, Junae Kim, and Claudia Szabo. Adversarial machine learning attacks and defences in multi-agent reinforcement learning. *ACM Computing Surveys*, 57(5): 1–35, 2025.

Mostofa Ahsan, Kendall E Nygard, Rahul Gomes, Md Minhaz Chowdhury, Nafiz Rifat, and Jayden F Connolly. Cybersecurity threats and their mitigation approaches using machine learning—a review. *Journal of Cybersecurity and Privacy*, 2(3):527–555, 2022.

Olakunle Ibitoye, Rana Abou-Khamis, Mohamed el Shehaby, Ashraf Matrawy, and M Omair Shafiq. The threat of adversarial attacks on machine learning in network security–a survey. *arXiv preprint arXiv:1911.02621*, 2019.

Pinlong Zhao, Weiyao Zhu, Pengfei Jiao, Di Gao, and Ou Wu. Data poisoning in deep learning: A survey. *arXiv preprint arXiv:2503.22759*, 2025.

Ram Shankar Siva Kumar, Magnus Nystr¨om, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xia. Adversarial machine learningindustry perspectives. In *2020 IEEE security and privacy workshops (SPW)*, pages 69–75. IEEE, 2020.

Sahil Girhepuje, Aviral Verma, and Gaurav Raina. A survey on offensive ai within cybersecurity. *arXiv preprint arXiv:2410.03566*, 2024.

Scott G Hodes, Kory J Blose, and Timothy J Kane. Black box phase-based adversarial attacks on image classifiers. In *Automatic Target Recognition XXXIV*, volume 13039, pages 12–31. SPIE, 2024.

Senming Yan, Jing Ren, Wei Wang, Limin Sun, Wei Zhang, and Quan Yu. A survey of adversarial attack and defense methods for malware classification in cyber security. *IEEE Communications Surveys & Tutorials*, 25(1):467–496, 2022.

Shalini Saini, Anitha Chennamaneni, and Babatunde Sawyerr. A review of the duality of adversarial learning in network intrusion: Attacks and countermeasures. *arXiv preprint arXiv:2412.13880*, 2024.

Shuai Zhou, Chi Liu, Dayong Ye, Tianqing Zhu, Wanlei Zhou, and Philip S Yu. Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity. *ACM Computing Surveys*, 55(8):1–39, 2022.

Ugochukwu Ikechukwu Okoli, Ogugua Chimezie Obi, Adebunmi Okechukwu Adewusi, and Temitayo Oluwaseun Abrahams. Machine learning in cybersecurity: A review of threat detection and defense mechanisms. *World Journal of Advanced Research and Reviews*, 21(1):2286–2295, 2024.

VO Eghaghe, OS Osundare, CP Ewim, and IC Okeke. Advancing aml tactical approaches with data analytics: Transformative strategies for improving regulatory compliance in banks. *Finance & Accounting Research Journal*, 6(10):1893–1925, 2024.

Weiyu Cui, Xiaorui Li, Jiawei Huang, Wenyi Wang, Shuai Wang, and Jianwen Chen. Substitute model generation for black-box adversarial attack based on knowledge distillation. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 648–652. IEEE, 2020.

Yahya Layth Khaleel, Mustafa Abdulfattah Habeeb, AS Albahri, Tahsien Al-Quraishi, OS Albahri, and AH Alamoodi. Network and cybersecurity applications of defense in adversarial attacks: A state-of-the-art using machine learning and deep learning methods. *Journal of Intelligent Systems*, 33(1):20240153, 2024.

Yulong Wang, Tong Sun, Shenghong Li, Xin Yuan, Wei Ni, Ekram Hossain, and H Vincent Poor. Adversarial attacks and defenses in machine learning-empowered communication systems and networks: A contemporary survey. *IEEE Communications Surveys & Tutorials*, 25(4):2245–2298, 2023.