OPEN ACCESS

Volume: 5

Issue: 3

Month: July

Year: 2026

ISSN: 2583-7117

Published: 31.07.2026

Citation:

Ujjwal Deshmukh "Adversarial Machine Learning in Cybersecurity: A Survey of Attacks, Defenses, and Robustness Evaluation" International Journal of Innovations in Science Engineering and Management, vol. 5, no. 3, 2026, pp. 210-217

DOI:

10.69968/ijisem.2026v5i3210-217

Adversarial Machine Learning in Cybersecurity: A Survey of Attacks, Defenses, and Robustness Evaluation

Ujjwal Deshmukh[1]

[1]*B.tech CSE (Data Scienc e), Oriental College Of Science And Technology, Bhopal*

Mail Id: *ujjwaldeshmukh3900@gmail.com*

Abstract

The rise of artificial intelligence (AI) and machine learning (ML) in cybersecurity has made Adversarial Machine Learning (AML) a key field of research. Though ML-based systems are more effective in intrusion detection, malware analysis, spam filtering and authentication, they are also susceptible to adversarial attacks that modify input samples, modify ML models or alter training data. The review explores the key adversarial attack classes: poisoning, evasion, model extraction, model inversion, and membership inference and also white-box, black-box, and grey-box threat models. It also provides an introduction to important defence methods like adversarial training, feature squeezing, defensive distillation, robust optimisation, detection-based methods, and ensemble learning. In addition, robustness evaluation metrics, benchmark datasets and attack assessment techniques to measure model robustness are highlighted. Lastly, the emerging trends are discussed in order to find future research directions in the field of creating trustworthy and resilient AI-based cybersecurity systems, such as Explainable AI, Federated Learning, Large Language Models, Autonomous Cyber defence, and Regulatory structures.

Keywords; Adversarial Machine Learning; Cybersecurity; Adversarial Attacks; Defense Mechanisms; Robustness Evaluation.

INTRODUCTION

Background

Artificial Intelligence (AI) and Machine Learning (ML) are indispensable tools in the modern cybersecurity landscape, as they allow for automated detection, analysis, and response capabilities to increasingly complex cyber threats. In contrast to signature-based security approaches, ML systems can be trained to detect patterns in vast amounts of security data and recognize attacks that have not been previously observed. As a result, AI-powered cybersecurity tools are now commonplace in network monitoring, threat intelligence, and automated incident response, enhancing the efficiency and effectiveness of security operations [1], [2].

Several cybersecurity applications have benefited from machine learning, such as intrusion detection, malware detection, spam and phishing filtering, and user authentication. An ML-based Intrusion Detection System (IDS) will review network traffic to find any malicious activity, while a model for identifying malware will utilise not only the behaviour of the malware, but static features [3]. Likewise, classification algorithms are applied to identify spam and detect phishing, while behavioural biometrics and anomaly detection are employed to enhance authentication. These features help security systems stay ahead of the curve when it comes to fringing the ever-changing cyber threats while minimizing false alarms and enhancing the system's detection capabilities [4], [5].

Although these benefits, ML models are vulnerable to manipulation by adversaries. Adversarial Machine Learning (AML) is the branch of study of attacks aimed at misleading machine learning models and the defense strategies that can enhance their security [6]. An adversarial example is a carefully crafted input that can be used to attack a model during deployment; poisoning attacks can be used to

manipulate training data. These attacks illustrate that high prediction accuracy does not necessarily mean robustness against malicious manipulation and highlight the statistical properties and generalisation behaviour of ML algorithms [7].

This increased use of ML for important applications in cybersecurity has thus spurred the need for reliable and secure models that can sustain good performance in the presence of adversarial actions. The development of effective defence mechanisms and the robustness analysis of models have become basic research needs for the establishment of credible cybersecurity systems with AI technology.

Types of Adversarial Attacks in Machine Learning

Adversarial attacks are attacks on vulnerabilities in machine learning (ML) models that involve the manipulation of data, the behavior of the model, or the output of the system in order to decrease the accuracy of the predictions that the model makes or to compromise the confidentiality of the system. Adversarial attacks are not the typical cyber attacks targeting software or network infrastructure, but they specifically target the learning process of AI systems. These attacks can take place during model development or model deployment, and can have an impact on the integrity, availability, or privacy of ML-based cybersecurity applications, depending on the objectives of the attack and the attacker's understanding of the target model [8].

- **Attack Lifecycle (Training Phase vs. Inference Phase):** Adversarial attacks can be generally classified based on the phase of the ML lifecycle being attacked. Attacker(s) intend to disrupt the training process by altering the training data or training labels prior to deployment, leading to incorrect decision boundaries. However, in inference-phase attacks, the attacker can tweak the input samples to change the output of the model without changing the parameters of the model, once it has been trained. This distinction can help determine what aspect of learning has been affected by the attack and if it is model learning or if there are flaws in their use of the model.
- **Poisoning Attacks:** Poisoning attacks try to trick the training data by adding compromised samples or incorrectly labeled samples to the training set. These tainted samples can affect the learning algorithm, leading the trained model to misclassify future samples, or to produce covert backdoors that can be exploited by attackers. The attacks are especially damaging in applications that constantly update models based on crowdsourced or automatically gathered data.
- **Evasion Attacks:** The most common inference-phase attacks are evasion attacks. They produce carefully designed adversarial examples that are sometimes hard to see, but involve adding a small modification to a legitimate input. These changes do not impact a human user, but may cause an ML model to misclassify malware, malicious network traffic or phishing emails so that they slip past security detection systems[9].
- **Model Extraction, Model Inversion, and Membership Inference:** In model extraction attacks, an attacker tries to recover the functionality of a deployed ML model by making multiple queries, thereby mimicking an important proprietary model. Attacks to privacy involve recovering sensitive information from the training process, such as model inversion, to recover information about the training data from the model output, and membership inference, to decide if a particular piece of data was used to train the model or not. These attacks pose threats to both the confidentiality of data and to intellectual property, especially in cloud-based ML services.
- **White-box, Black-box, and Grey-box Attack Settings:** Adversarial attacks can be further broken down based on the knowledge of the target model held by the adversary. In white-box attacks, the adversary knows the entire model architecture, architecture parameters, and how the model was trained, and attacks are very effective. Black-box attacks require no insider information or knowledge, only the knowledge of the model's input and output. Grey-box attacks fall somewhere between these two extremes, where an attacker knows some details of the model, such as the model architecture, or some of the training data. To build strong cybersecurity defenses, it is crucial to test the ML models under these threat scenarios [10].

Table 1: Classification of Major Adversarial Attacks in Machine Learning

Attack Type	Attack Phase	Primary Objective	Impact on Cybersecurity Systems
Poisoning Attack	Training	Manipulate training data to corrupt learning	Reduces IDS and malware detection accuracy

Evasion Attack	Inference	Generate adversarial examples to bypass detection	Evades malware, spam, or intrusion detection
Model Extraction	Inference	Replicate the target ML model	Theft of intellectual property and model functionality
Model Inversion	Inference	Recover sensitive information from the trained model	Privacy leakage of training data
Membership Inference	Inference	Determine whether a record belongs to the training dataset	Disclosure of confidential user information
White-box Attack	Training/ Inference	Exploit complete model knowledge	Produces highly effective adversarial attacks
Black-box Attack	Inference	Attack using only model queries	Practical attack scenario against deployed services
Grey-box Attack	Training/ Inference	Exploit partial model knowledge	Moderate attack capability with realistic assumptions

Defense Mechanisms against Adversarial Attacks

As machine learning (ML) is becoming a key component of cybersecurity, its use has simultaneously made ML models more useful for detecting cyber threats, but also more susceptible to adversarial attacks. The attacks can exploit training data or model predictions to compromise the integrity, confidentiality, or availability of the model. However, none of the defence mechanisms could tackle all of the adversarial threats, leading to complementary strategies to boost the robustness of the models during training and deployment. Good defensive strategies try to reduce misclassification, avoid information leakage, and ensure the reliability and trustworthiness of ML-based cyber security systems and services like intrusion detection, malware classification and user authentication [11].

- **Adversarial Training:** One of the most popular defence techniques is the adversarial training technique. It increases the robustness of the model by adding adversarial examples to the training set so that the model learns to make more robust decision boundaries that are less sensitive to malicious perturbations. Adversarial training is very effective at training against known attacks, but it adds computational expense and may offer some protection against previously unseen methods of attack.
- **Input Preprocessing and Feature Squeezing:** The goal of the input preprocessing is to eliminate adversarial perturbations prior to the model analysing the data. Image denoising, image normalisation, input transformation, or feature squeezing are common techniques, such as reducing the color depth of the input image or smoothing the feature to be extracted from the input image. These techniques can help you withstand basic attacks, but they are not strong enough against adaptive attackers.
- **Defensive Distillation and Robust Optimisation:** Defensive distillation involves training a second model with softened probability outputs from an initial model, which decreases sensitivity to small changes in the inputs. To produce models that are more accurate in the worst-case attack scenario, robust optimisation adds an extension to this concept by including adversarial perturbations directly into the optimisation objective during model training.
- **Detection-Based Defences, Ensemble Learning, and Deployment Challenges:** Detection-based methods detect adversarial inputs prior to classification, through examining variations in statistics or inconsistencies in predictions. Another way to make successful attacks more difficult is by using ensemble learning which combines predictions made by multiple models that are diverse. However, despite these progressions, it is still difficult to use them practically as defence algorithms may introduce significant computational costs, slow down inference, and may not be generalizable to unknown or adaptive attacks. Thus, it is believed that multiple complementary defence mechanisms is the best solution to ensure the security of cyber systems based on ML.

Table 2: Major Defense Mechanisms against Adversarial Attacks

Defense Mechanism	Principle	Strengths	Limitations
Adversarial Training	Train using adversarial examples	High robustness against known attacks	Computationally expensive; limited against novel attacks
Input Preprocessing & Feature Squeezing	Remove or reduce adversarial perturbations	Simple to implement; improves baseline security	Vulnerable to adaptive attacks

Defensive Distillation	Train with softened probability outputs	Reduces sensitivity to perturbations	Less effective against advanced attack methods
Robust Optimisation	Optimise for worst-case perturbations	Strong theoretical robustness	High training complexity
Detection-Based Methods	Identify adversarial inputs before prediction	Prevents malicious inputs from reaching the model	Detection accuracy varies across attacks
Ensemble Learning	Combine predictions from multiple models	Improves robustness and reliability	Increased computational and memory requirements

Robustness Evaluation and Security Assessment of ML Models

With the growing adoption of machine learning (ML) models in cybersecurity applications, it has become crucial to assess their robustness against adversarial attacks to guarantee dependable and secure performance. Adversarial machine learning considers robustness as the ability of a model to give a correct prediction even when maliciously designed perturbations are added to fool the model. A robust model should not only generate a high prediction accuracy in normal circumstances, but it should also be capable of being robust when faced with adversarial inputs. As a result, the robustness evaluation has been so far established as a key part of security evaluation, which assists to find vulnerabilities in the model and to compare the level of security of defence mechanisms [12].

Evaluation metrics like accuracy under attack (accuracy of model on adversarial samples), robust accuracy (percentage of correctly classified adversarial inputs) and perturbation limits (maximum possible input modification without changing the prediction) are commonly used measures of robustness. To evaluate and compare defence methods, a variety of standard benchmark datasets such as MNIST and CIFAR-10 for image classification or NSL-KDD and CICIDS2017 for intrusion detection are used and applied with the same experimental setup [13].

Well known attack evaluation approaches are also used for security assessment. Adversarial examples are created using a single perturbation with the Fast Gradient Sign Method (FGSM) and a multi-step iterative perturbation with Projected Gradient Descent (PGD), which is considered one of the strongest first-order attacks. In the Carlini and Wagner (CW) attack, the authors use optimisation techniques to create very effective and hard-to-detect adversarial examples. While robust procedures offer greater security, they can also result in increased computational difficulty, and may even lead to a decrease in performance when working with clean data, leading to a compromise between robustness, efficiency and model accuracy. Standardized assessment methods and repeatable benchmarks are crucial for objective cross-comparison of defense strategies and for the development of reliable and credible ML-based cybersecurity systems [14], [15].

Emerging Trends

The landscape of adversarial machine learning is continually evolving, and researchers are actively working on sophisticated methods to make AI-based cybersecurity systems more secure, transparent, and resilient. Emerging trends are not just about making models more robust to adversarial attacks, but also about trusted decision-making, privacy protection, and adherence to new regulatory requirements. It is anticipated these developments will have a significant impact on the next generation of intelligent cyber defence systems [16], [17].

- **Explainable AI (XAI) for Trustworthy Cybersecurity:** Explainable AI (XAI) makes AI models more transparent by offering explanations that human users can understand for security decisions. In cybersecurity, XAI assists analysts in comprehending the reasons why a phishing email, malware sample, or intrusion is malicious, helping to boost user confidence, investigate threats, and pinpoint adversarial manipulation [18].
- **Federated Learning and Adversarial Robustness:** This collaborative training of ML models without the sharing of raw data is called federated learning, which enhances privacy. Research efforts are currently aimed at creating strong federated learning methods that are more resilient to poisoning attacks, malicious client updates, and can ensure that the model performs well in distributed settings.
- **Large Language Models (LLMs) and New Adversarial Risks:** The widespread use of Large Language Models (LLMs) has brought with it a new set of security problems, such as prompt injection, jailbreak attacks, data leakage, and the creation of malicious content. As a result, ensuring secure alignment, adversarial testing, and trustworthy safety mechanisms are still significant research challenges for ensuring the secure deployment of LLMs [19].
- **AI-Driven Autonomous Cyber Defence:** Autonomous cyber defence is a combination of

machine learning and automated threat detection, incident response, and adaptive security management, powered by AI. These intelligent systems have the ability to constantly monitor the network, detect new attacks and trigger quick countermeasures with minimal human input, enhancing cyber-resilience to complex threats.

- **Standardisation and Regulatory Considerations:** As AI continues its growing presence in critical infrastructure, the importance of standardized robustness evaluation frameworks and regulatory guidelines has come into focus. The NIST AI Risk Management Framework and the European Union AI Act, among other international efforts, highlight the importance of trustworthy AI, risk assessment, transparency, and security considerations for safe and responsible use of AI-powered cybersecurity solutions.

LITERATURE REVIEW

(TEKESTE et al., 2026) [20] The topic, which was initially centred on algorithmic robustness, has developed into a complicated nexus of assurance, security, and policy. We use a lifecycle-oriented taxonomy to map attack vectors and defence mechanisms to specific phases of the AI pipeline, from data collection to deployment. This expands "the traditional Confidentiality, Integrity, and Availability (CIA)" triad to include Governance and Regulation. We pinpoint important research needs, such as rising risks in Generative AI and verified robustness for "Natural Language Processing (NLP)". We examine five domain-specific case studies—"autonomous vehicles, medical artificial intelligence (AI), financial systems, natural language processing (NLP), and the Internet of Things (IoT)"—to put these theoretical insights into reality. This survey is unique in that it maps technical AML results to developing standards, such as "ISO/IEC 42001, MITRE ATLAS, and the NIST AI Risk Management Framework (RMF)", bridging the gap between academic research and industry practice. Finally, we offer a path for practitioners, scholars, and regulators to develop AI systems that are verifiable, reliable, and compliant.

(Singh & Gupta, 2026) [21] The discipline of adversarial machine learning (AML) is experiencing rapid growth, particularly in settings where security is a critical concern, as machine learning models are increasingly implemented. With a focus on the relationship between AML and network security, this review delves deeply into 746 papers from the Scopus database. We examined publishing patterns, study domains, productive authors, cooperation networks, and topic concentrations using Biblioshiny and Scopus tools. Thirteen charts that highlight significant contributions, prominent publications, popular keywords, and citation patterns illustrate the evolution of AML research over time. Our findings show that the subject is diverse and has research facilities worldwide. Additionally, it indicates that researchers are no longer collaborating as much and that the emphasis is shifting from attack modelling to robustness and interpretability. Some significant gaps are identified at the study's conclusion, including the clichés of ethical governance and real correspondent execution. In order to strengthen and increase the accessibility of the AML research ecosystem, the paper also suggests future research directions.

(Kakkad et al., 2026) [22] carried out two thorough studies to investigate students' and industry professionals' perceptions on various AML vulnerabilities and their instructional approaches. In our initial study, we polled professionals online and found a significant relationship between cybersecurity education and awareness of AML risks. In order to show a poisoning assault on the training data set, we created two CTF challenges for our second research that incorporate Natural Language Processing and Generative AI ideas. Carnegie Mellon University undergraduate and graduate students were surveyed to assess the efficacy of these challenges, and the results showed that a CTF-based strategy successfully piques interest in AML concerns. We offer comprehensive recommendations highlighting the vital necessity of integrated security education within the ML curriculum based on the replies of the research participants.

(Penmetsa et al., 2025) [10] The focus of adversarial machine learning (AML) is on how attackers alter data inputs to take advantage of weaknesses in ML systems, which can result in "misclassification, data breaches, and system failures". This article provides a thorough analysis of adversarial attacks against machine learning models in cybersecurity, classifying them according to their goals and techniques into evasion, poisoning, and inference assaults. Threat models, adversary knowledge levels (white-box, black-box, and gray-box), and popular attack methods like FGSM, Deep Fool, and model extraction are also examined. In order to demonstrate how academics and practitioners are attempting to strengthen the resilience of AI-powered security systems, defensive distillation, adversarial training, and input preprocessing are also covered. This paper intends to provide a thorough description of the adversarial

environment in cybersecurity, serving as a reference for further research and real-world applications of robust and secure machine learning systems.

(Jedrzejewski et al., 2024) [23] The act of attacking and defending machine learning (ML) models, a crucial component of artificial intelligence (AI), is covered in adversarial machine learning (AML). In the future, the industry will pay even more attention to AI and ML, yet the risks posed by previously identified assaults that explicitly target ML models are either overlooked, disregarded, or managed improperly. Current AML research examines ML attack and defence scenarios in many industrial contexts with differing levels of academic rigour and practical applicability. To the best of our knowledge, however, there isn't a synthesis of the current level of academic rigour with practical usefulness. The rigour and relevance ratings of each study are measured and analysed in this evaluation of the literature on AML in the context of industry. All of the studies had a low relevance score and a high rigour score overall, meaning that while they are well-designed and documented, they fail to contain touch points that practitioners can relate to.

(Rosenberg et al., 2021) [24] provides a thorough overview of the most recent research on adversarial attacks against machine learning-based security systems and highlights the dangers they provide. The adversarial attack techniques are first described according to the attacker's objectives and capabilities, as well as the stage at which they occur. Next, we classify how adversarial attack and defence techniques are used in the field of cyber security. Lastly, we examine the influence of recent developments in other adversarial learning domains on future research paths in the cyber security area and highlight certain features found in recent research. To the best of our knowledge, this study is the first to address the particular difficulties associated with putting end-to-end adversarial assaults into practice in the field of cyber security, map them into a single taxonomy, and utilise the taxonomy to identify areas for future research.

(Babatunde et al., 2020) [25] offers a thorough analysis of AML in the context of cybersecurity, emphasising the types of adversarial attacks, how they affect various ML systems, and cutting-edge defences. In areas like "malware detection, spam filtering, and network intrusion detection", case studies demonstrate how vulnerable deep neural networks, support vector machines, and ensemble techniques are to minute but deliberate perturbations. We also look at ethical and regulatory issues, such as the necessity of transparent and explicable AI to promote confidence in adversarially resistant systems. In its conclusion, the report identifies outstanding research issues and suggests future options, including the integration of AML defences into the larger cybersecurity ecosystem, hybrid human-AI defence frameworks, and standardised assessment criteria for adversarial resistance. Our results highlight the critical need for a multidisciplinary strategy that integrates policy formulation, technical innovation, and ongoing monitoring to protect machine learning applications in cybersecurity against changing adversary threats.

CONCLUSION

As artificial intelligence and machine learning systems become more ubiquitous in security-critical systems, adversarial machine learning has emerged as one of the core fields in cybersecurity research. The review demonstrated various types of adversarial attacks on machine learning models throughout their lifecycle, such as poisoning, evasion, model extraction, and privacy-oriented attacks. It also explored important defence strategies like adversarial training, defensive distillation, robust optimisation, input preprocessing, and ensemble learning, all of which enhance model robustness to adversarial manipulation. To ensure the objective assessment of the security and reliability of ML models, it is still crucial to evaluate their robustness, relying on standard metrics, benchmark datasets, and attack methods. Additionally, breakthroughs in Explainable AI, federated learning, Large Language Models, autonomous cyber defence and international regulation highlight the dynamic nature of adversarial AI research. Going forward, it is crucial to focus on establishing standardized evaluation protocols, adaptive defence strategies, and on ensuring adherence to established principles of trustworthiness in AI to create secure, transparent, and resilient machine learning systems that can tackle more complex cyber threats.

REFERENCES

[1] N. Mohamed, “Artificial intelligence and machine learning in cybersecurity: a deep dive into state-of-the-art techniques and future paradigms Nachaat,” *Knowl. Inf. Syst.*, vol. 67, 2025, doi: 10.1007/s10115-025-02429-y.

[2] E. Kesavan, “Internet of Things (IoT): A Review of Security Challenges and Solutions,” *Int. J. Innov. Sci. Eng. Manag.*, vol. 2, no. 4, pp. 65–71, 2023, doi: 10.69968/ijisem.2023v2i465-71.

[3] L. A. M. Hernández, S. P. Arteaga, A. L. S. Orozco, and L. J. G. Villalba, “Adversarial Attacks on Machine Learning Models for Network Traffic Filtering,” *Eng. Proc.*, 2026.

[4] S. Rahman, S. Pal, A. Habib, L. Pan, and C. Karmakar, "Attack-data independent defence mechanism against adversarial attacks on ECG signal," *Comput. Networks*, vol. 258, p. 111027, 2025, doi: 10.1016/j.comnet.2024.111027.

[5] P. Dixit and P. B. Singh, "The Impact of Artificial Intelligence on Various Sectors: Benefits, Challenges, and Future Prospects," *Interantional J. Sci. Res. Eng. Manag.*, vol. 3, no. 2, pp. 140–144, 2024, doi: 10.69968/ijisem.2024v3si2140-144.

[6] M. U. Tariq, "Defense Mechanisms Against Adversarial Attacks Strengthening AI Security in Cy ersecurity Applications," in *Challenges and Solutions for Cybersecurity and Adversarial Machine Learning*, 2025, pp. 199–201. doi: 10.4018/979-8-3373-2200-1.ch007.

[7] A. D. Lahe, G. G. Parrkhed, and B. L. Rathi, "A Study on Security Challenges in Machine Learning," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, pp. 166–176, 2023.

[8] E. Alshahrani, D. Alghazzawi, R. Alotaibi, and O. Rabie, "Adversarial attacks against supervised machine learning based network intrusion detection systems," *PLoS One*, pp. 1–14, 2022, doi: 10.1371/journal.pone.0275971.

[9] P. Xiong, S. Buffett, S. Iqbal, P. Lamontagne, M. Mamun, and H. Molyneaux, "Towards a Robust and Trustworthy Machine Learning System Development: An Engineering Perspective," *arXiv*, pp. 1–58, 2022.

[10] M. Penmetsa, J. R. Bhumireddy, R. Chalasani, S. R. Vangala, R. M. Polam, and B. Kamarthapu, "Adversarial Machine Learning in Cybersecurity: A Review on Defending Against AI-Driven Attacks," *Eur. J. Appl. Sci. Eng. Technol.*, vol. 3, no. 4, pp. 4–14, 2025, doi: 10.59324/ejaset.2025.3(4).01.

[11] S. S. R. Banait, A. C. M, M. Sen, D. Mondal, D. Dhabliya, and J. R. R. Kumar, *Advancements in Defense Mechanisms against Adversarial Attacks in Computer Vision*, vol. 1, no. 1. Association for Computing Machinery, 2024. doi: 10.1145/3745812.3745886.

[12] J. C. Palomares-salas, S. Aguado-gonzález, and J. M. Sierra-Fernández, "Robustness of Machine Learning and Deep Learning Models for Power Quality Disturbance Classification: A Cross-Platform Analysis," *Appl. Sci.*, pp. 1–16, 2025.

[13] J. Zhang, J. Li, and Z. Yang, "Dynamic robustness evaluation for automated model selection in operation," *Inf. Softw. Technol.*, vol. 178, p. 107603, 2025, doi: 10.1016/j.infsof.2024.107603.

[14] S. Pal *et al.*, "Vulnerabilities in Machine Learning for cybersecurity: Current trends and future research directions Shantanu," *J. Inf. Secur. Appl.*, vol. 96, p. 104269, 2026, doi: 10.1016/j.jisa.2025.104269.

[15] J. YU, A. V. SHVETSOV, and S. H. ALSAMHI, "Leveraging Machine Learning for Cybersecurity Resilience in Industry 4.0: Challenges and Future Directions," *IEEE Access*, vol. 12, pp. 159579–159596, 2024, doi: 10.1109/ACCESS.2024.3482987.

[16] K. Razzaq and M. Shah, "Emerging Trends in Cybersecurity: Machine Learning as a Game-Changer in Next Generation Cybersecurity Applications," *F1000Research*, vol. 15, no. 276, 2026.

[17] J. Sharma, "Cyber Crime Causes and prevention: An Analysis," *Int. J. Innov. Sci. Eng. Manag.*, vol. 4, no. 3, pp. 66–72, 2025, doi: 10.69968/ijisem.2025v4i366-72.

[18] A. Singh and N. Shanker, "Redefining Cybercrimes in light of Artificial Intelligence: Emerging threats and Challenges," *Int. J. Innov. Sci. Eng. Manag.*, vol. 3, no. 2, pp. 192–201, 2024, doi: 10.69968/ijisem.2024v3si2192-201.

[19] S. Joshi, "Review of Gen AI Models for Financial Risk Management: Architectural Frameworks and Implementation Strategies," *Int. J. Innov. Sci. Eng. Manag.*, vol. 4, no. 3, pp. 207–222, 2025, doi: 10.69968/ijisem.2025v4i2207-222.

[20] B. TEKESTE, K. AL-HUSSAENI, B. C. M. FUNG, I. ALAWADHI, and C. FACHKHA, "Adversarial Machine Learning: A 20-Year Survey of Attacks, Defenses, and Standards," *IEEE Access*, vol. 14, pp. 69778–69812, 2026, doi: 10.1109/ACCESS.2026.3690620.

[21] S. Singh and A. Gupta, "Adversarial Attacks on Machine Learning Models in Cybersecurity: A Systematic Literature Review," *J. Artif. Intell. Res. Adv.*, vol. 13, no. 1, pp. 23–38, 2026.

[22] V. Kakkad, P. Chung, H. Hibshi, and M. Woo, "Comparative Insights on Adversarial Machine Learning from Industry and Academia: A User-Study Approach," *arXiv*, 2026.

[23] F. V. Jedrzejewski, L. Thode, J. Fischbach, T. Gorschek, D. Mendez, and N. Lavesson,

"Adversarial Machine Learning in Industry: A Systematic Literature Review," *Comput. Secur.*, vol. 145, p. 103988, 2024, doi: 10.1016/j.cose.2024.103988.

[24] I. Rosenberg, A. Shabtai, Y. Elovici, and L. Rokach, "Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain," *ACM Comput. Surv.*, vol. 54, no. 5, 2021.

[25] L. A. Babatunde *et al.*, "Adversarial Machine Learning in Cybersecurity: Vulnerabilities and Defense Strategies," *J. Front. Multidiscip. Res.*, vol. 01, no. 02, pp. 31–45, 2020.