World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(RESEARCH ARTICLE)

Check for updates

Adversarial machine learning in cybersecurity: Mitigating evolving threats in AI-powered defense systems

Ebuka Mmaduekwe Paul [1,*], Ugochukwu Mmaduekwe Stanley [2], Joseph Darko Kessie [3] and Mukhtar Dolapo Salawudeen [4]

[1] *Department of Information and communication science, Ball state university, Muncie Indiana, USA.*
[2] *Mechanical Engineering, University of Nigeria Nsukka Nigeria.*
[3] *Department of Cybersecurity, Eastern Illinois University, Charleston, Illinois, United States.*
[4] *IA Technology Risk and Cybersecurity, Goldman Sachs, New York, USA.*

World Journal of Advanced Engineering Technology and Sciences, 2023, 10(02), 309-325

Publication history: Received on 12 October 2023; revised on 26 October 2023; accepted on 29 November 2023

Article DOI: https://doi.org/10.30574/wjaets.2023.10.2.0294

Abstract

The increasing integration of artificial intelligence (AI) in cybersecurity has enhanced the ability to detect and mitigate cyber threats in real-time. However, adversarial machine learning (AML) has emerged as a significant challenge, enabling attackers to manipulate AI models and bypass security measures. This study explores the evolving landscape of AML threats and the vulnerabilities they introduce to AI-powered defense systems. The research identifies key adversarial attack techniques, including evasion, poisoning, model inversion, and model extraction, which threaten the integrity and effectiveness of AI-driven cybersecurity mechanisms. This study evaluates various mitigation strategies to address these threats, such as adversarial Training, model hardening, defensive Distillation, and hybrid AI approaches. Through experimental analysis, we assess the robustness of AI defense systems under adversarial attack and measure their effectiveness using key performance metrics, including model accuracy, false positive rates, and computational efficiency. The findings indicate that while adversarial Training improves model resilience, adaptive attack techniques continue to challenge existing defenses, necessitating continuous advancements in cybersecurity frameworks. This research highlights the need for a multi-layered security approach that integrates AI-based anomaly detection, human-AI hybrid security models, and adaptive learning techniques to counter adversarial threats effectively. Additionally, it discusses the broader implications of AML in cybersecurity, including policy considerations, ethical concerns, and future research directions. The study recommends strategies for enhancing AI-powered cyber defense systems to maintain security, reliability, and resilience against evolving adversarial threats.

Keywords: Adversarial Machine Learning; AI-Powered Cybersecurity; Adversarial Attacks; Intrusion Detection Systems (Ids); Cyber Threat Intelligence

1. Introduction

1.1. Background & Context: Overview of machine learning in cybersecurity

Cybersecurity refers to a set of technologies, processes, and practices to protect and defend networks, devices, software, and data from attack, damage, or unauthorized access. Cybersecurity is becoming complex because of the exponential growth of interconnected devices, systems, and networks. This is exacerbated by advances in the digital economy and infrastructure, leading to significant growth of cyberattacks with serious consequences. In addition, researchers report the continued evolution of nation-state-affiliated and criminal adversaries and the increasing sophistication of cyberattacks, which find new and invasive ways to target even the savviest targets. This evolution is driving an increase

* Corresponding author: Ebuka Mmaduekwe Paul

in the number, scale, and impact of cyberattacks and necessitating the implementation of intelligence-driven cybersecurity to provide a dynamic defense against evolving cyberattacks and to manage big data. Advisory organizations, such as the National Institute of Standards and Technologies (NIST), are also encouraging the use of more proactive and adaptive approaches by shifting towards real-time assessments, continuous monitoring, and data-driven analysis to identify, protect against, detect, and respond to, and catalog cyberattacks to prevent future security incidents.

AI is an intriguing tool that can provide analytics and intelligence to protect against ever-evolving cyberattacks by swiftly analyzing millions of events and tracking various cyber threats to anticipate and act in advance of the problem. For this reason, AI is increasingly being integrated into the cybersecurity fabric and used in multiple use cases to automate security tasks or support the work of human security teams. The flourishing field of cybersecurity and the growing enthusiasm of researchers from both AI and cybersecurity have resulted in numerous studies to solve problems related to the identification, protection, detection, response, and recovery from cyberattacks.

1.2. Problem Statement

The increasing sophistication of adversarial attacks poses a significant challenge to AI-driven cybersecurity solutions. As artificial intelligence becomes integral to threat detection, malware analysis, and network security, adversarial actors are developing more advanced techniques to bypass these AI defenses. Attackers manipulate AI models by injecting adversarial inputs, causing misclassifications or false negatives that can compromise security systems. Traditional cybersecurity methods struggle to counter these dynamic threats, necessitating the development of robust AI-powered defense mechanisms. The growing reliance on AI in critical sectors such as finance, healthcare, and national security further amplifies the risks associated with adversarial attacks, making it imperative to develop adaptive, resilient, and explainable AI security solutions.

1.3. Research Objectives

- Identify Key Adversarial Threats to AI in Cybersecurity: This study aims to examine various adversarial techniques, including evasion, poisoning, and model inversion attacks, that threaten AI-driven security solutions. Understanding these threats will help highlight vulnerabilities in current AI security models and guide the development of more resilient defenses.
- Analyze Existing Mitigation Strategies: The research will evaluate the effectiveness of current adversarial defense mechanisms such as adversarial Training, ensemble learning, and model robustness techniques. This study will determine how well these methods protect AI models from evolving adversarial threats by assessing their strengths and limitations.
- Propose Advanced Techniques to Strengthen AI-Powered Defense Systems: Given the limitations of existing approaches, this study seeks to explore innovative adversarial defense strategies. Potential solutions include hybrid AI security models combining deep Learning with symbolic reasoning, reinforcement learning-based threat detection, and Explainable AI (XAI) techniques to improve model transparency and threat interpretability.

1.4. Significance of the Study

Securing AI-driven cybersecurity systems is crucial in an era where cyber threats are becoming more intelligent and adaptive. The effectiveness of AI in cybersecurity depends on its ability to detect, respond to, and mitigate attacks in real time. However, adversarial techniques can undermine AI's reliability, making security systems vulnerable to exploitation. This study's findings will contribute to developing more resilient AI security frameworks, ensuring robust protection for critical infrastructure and sensitive data. Moreover, by advancing AI-powered defense mechanisms, organizations can enhance their threat response capabilities, minimize cybersecurity risks, and build trust in AI-driven security applications. The research will also provide insights for policymakers, encouraging the establishment of regulatory frameworks that promote AI security standards and responsible AI deployment.

2. Literature Review

2.1. Understanding Adversarial Machine Learning

AML is concerned with identifying vulnerabilities and mitigating them for machine learning algorithms. Execution-time AML attacks against supervised Learning have largely focused on Evasion attacks, which occur when slight human-imperceptible input changes return significantly different outputs from neural networks. AML defenses for evasion attacks in supervised Learning have been broader and include adversarial Training, regularisation, adversarial detection, data preprocessing, and ensembles.

Evasion attacks may be targeted or untargeted. Untargeted evasion attacks, such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), aim to cause the victim to predict a different class. Targeted evasion attacks, such as the Jacobian Saliency Map (JSM) and the Carlini and Wagner method (C&W), aim to cause the victim to output a specific class. Targeted attacks can also be created using variations of untargeted attacks; for example, the One-step target class methods [68] alter the loss function in the FGSM attack.

Evasion attacks also require a certain amount of knowledge about a victim. Attacks that require a full understanding of a victim are known as White-box attacks. Black-box attacks only require the ability to query the victim. Transfer attacks allow attacks that originally needed information from the white box only to require black-box information. Transfer attacks use black-box information to train a surrogate victim to replicate the behavior of the original victim. Then, white-box techniques are used against the surrogate to discover attacks that are also effective against the original victim. There are also grey-box attacks, which only require partial knowledge of the victim.

To counter evasion attacks, defensive techniques are used to mitigate the vulnerability. Adversarial Training mitigates AML attacks by training victims against original and adversarially perturbed data. However, adversarial Training may fail to defend against other attacks against which it was not taught. Regularisation alters the algorithm's training process to improve its robustness against AML attacks. Regularisation may include using additional terms, such as the Lipschitz constant, which aims to prevent sudden changes in the output caused by slight perturbations to the input. Preprocessing input data is an effective AML defense and includes techniques such as autoencoders. Input data can also be altered to remove adversarial perturbations if they can first be detected; thus, adversarial detection is also a key AML defense. Ensembles are also an AML defense that trains multiple algorithms that collectively decide the output.

2.2. Categories of adversarial attacks: evasion, poisoning, model inversion, and extraction

As AI technologies evolve and integrate into various industries, they become prime targets for sophisticated cyber threats. These threats exploit unique vulnerabilities inherent to AI systems, such as their reliance on data integrity, the transparency of their algorithms, and the security of their supporting infrastructure. Understanding the Nature of these vulnerabilities and the corresponding attack vectors is crucial for developing robust countermeasures that protect AI systems from potential cyber-attacks. Data poisoning represents a critical threat to AI systems, particularly because these systems rely heavily on data integrity for Training and operation. In a data poisoning attack, adversaries deliberately manipulate the training data to compromise the model's learning process, leading to flawed decision-making or predictive abilities. This attack is particularly dangerous because it can be difficult to detect and have far-reaching effects once a model is deployed. Techniques include Injection Attacks and Modification Attacks. Injection Attacks involve inserting malicious data points into the training dataset. Unaware of tampering, the AI system learns from this corrupted data, which can lead to significant deviations in its behavior. For instance, an AI model used for financial forecasting could be taught incorrect associations, leading to erroneous investment recommendations. Modification Attacks, on the other hand, alter existing data within the dataset rather than adding new data points.

Even minor changes to critical data points can retrain the model with false information, resulting in incorrect outputs. Such attacks might be used to manipulate systems like automated surveillance, where altering image data could prevent recognizing specific individuals or objects. Concrete examples of these techniques' real-world impacts include attackers compromising a facial recognition system by introducing subtly altered images into its training set. These alterations, imperceptible to humans, were significant enough to fool the system, failing to identify or misidentify individuals. This vulnerability was exploited to manipulate facial recognition, leading to incorrect tagging or ignoring faces, potentially bypassing security protocols. Another instance involved a traffic control AI in an urban smart city system. Attackers injected faulty data representing fake traffic conditions, such as non-existent traffic jams or accidents. The AI, trained with these false data points, generated incorrect traffic flow predictions, causing chaos in city traffic management and emergency response services.

2.3. Adversarial Threats to AI-Powered Cyber Defense

The emergence and proliferation of artificial intelligence (AI) technologies have revolutionized various aspects of our lives, from healthcare to finance, transportation, and beyond (Sahai and Rath, 2021; Allam and Allam, 2021). However, with this rapid advancement comes an evolving threat landscape with new risks and vulnerabilities. Understanding and mitigating these threats is essential to safeguarding AI systems and the sensitive data they handle. This essay explores the emerging threat landscape in AI, highlighting key risks and vulnerabilities that organizations and cybersecurity professionals must address.

One prominent risk in the emerging threat landscape is the susceptibility of AI systems to adversarial attacks. Adversarial attacks exploit vulnerabilities in AI algorithms by perturbing input data in subtle ways that are

imperceptible to humans but can cause the system to make incorrect predictions or classifications. These attacks seriously affect AI applications in critical domains such as autonomous vehicles, medical diagnosis, and cybersecurity. For example, in autonomous cars, adversarial attacks could manipulate sensor inputs to deceive the vehicle's perception system, leading to potentially catastrophic consequences on the road.

Furthermore, the interconnected nature of AI systems introduces data privacy and confidentiality vulnerabilities. As AI applications increasingly rely on vast amounts of data, ensuring the confidentiality and security of this data becomes paramount. Data breaches can have severe consequences, including financial loss, reputational damage, and regulatory penalties. Moreover, aggregating sensitive data from multiple sources in AI systems raises concerns about unauthorized access and misuse. Adversaries may exploit vulnerabilities in data storage and transmission protocols to gain unauthorized access to sensitive information, posing a significant threat to individuals' privacy and organizational security.

Another emerging threat in the AI landscape is manipulating AI-generated content, often called "deepfakes." Deepfakes use AI algorithms to create realistic but fabricated audio, video, or text content that can be used to spread misinformation, manipulate public opinion, or impersonate individuals. This poses significant challenges for media integrity, political discourse, and cybersecurity. With the proliferation of deepfake technology, distinguishing between authentic and manipulated content becomes increasingly difficult, undermining trust in digital media and exacerbating societal polarization.

Additionally, AI-enabled cyberattacks represent a growing concern in the emerging threat landscape. Adversaries can leverage AI algorithms to automate and enhance various stages of the cyberattack lifecycle, including reconnaissance, exploitation, and evasion. For example, AI-powered malware can autonomously adapt its behavior in response to changes in the target environment, making it more challenging for traditional cybersecurity defenses to detect and mitigate. Furthermore, AI-driven phishing attacks can leverage sophisticated social engineering techniques to deceive users and bypass email security filters, increasing the likelihood of successful compromises.

Moreover, the proliferation of AI-driven IoT devices introduces new attack surfaces and vulnerabilities in interconnected systems. IoT devices often lack robust security mechanisms, making them susceptible to exploitation by adversaries. Compromised IoT devices can be leveraged to launch large-scale distributed denial-of-service (DDoS) attacks, exfiltrate sensitive data, or infiltrate corporate networks. As the number of IoT devices continues to grow exponentially, securing these devices against cyber threats becomes increasingly challenging, necessitating proactive measures to address vulnerabilities at both the device and network levels (Montasari, 2022; Cohen, 2019.).

Furthermore, using AI for offensive cyber operations introduces geopolitical implications and risks. Nation-states and threat actors can leverage AI technologies to develop sophisticated cyber weapons capable of disrupting critical infrastructure, stealing sensitive information, or conducting covert surveillance. The proliferation of AI-driven cyber capabilities exacerbates the threat landscape, raising concerns about the escalation of cyber conflicts and the erosion of international cyberspace norms.

In conclusion, the emerging threat landscape in AI presents a complex and evolving challenge for organizations, governments, and cybersecurity professionals worldwide. The risks and vulnerabilities associated with AI systems are multifaceted and interconnected, from adversarial attacks and data privacy concerns to deepfakes, AI-enabled cyberattacks, and IoT vulnerabilities. Addressing these challenges requires a holistic approach encompassing technological innovation, policy development, and international cooperation (Tremont, 2023; Khatun et al., 2023). By understanding the evolving threat landscape and adopting proactive measures to mitigate risks, stakeholders can enhance the security and resilience of AI systems in an increasingly digital and interconnected world.

2.4. Existing Defense Mechanisms

We consider several categories in the classification of AML defenses for MARL, DRL, and MAL: the type of defense, when the defense occurs, and what attacks the defense counters. We have identified several AML defenses used to defend MARL and DRL algorithms from AML attacks. These are Adversarial Training, Competitive Training, Robust Learning, Adversarial Detection, Input Alteration, Memory, Regularisation, and Ensembles. When considering when a defense is applied, we identify four general times: during training execution, before and during an attack. Figure 2 shows our classification.

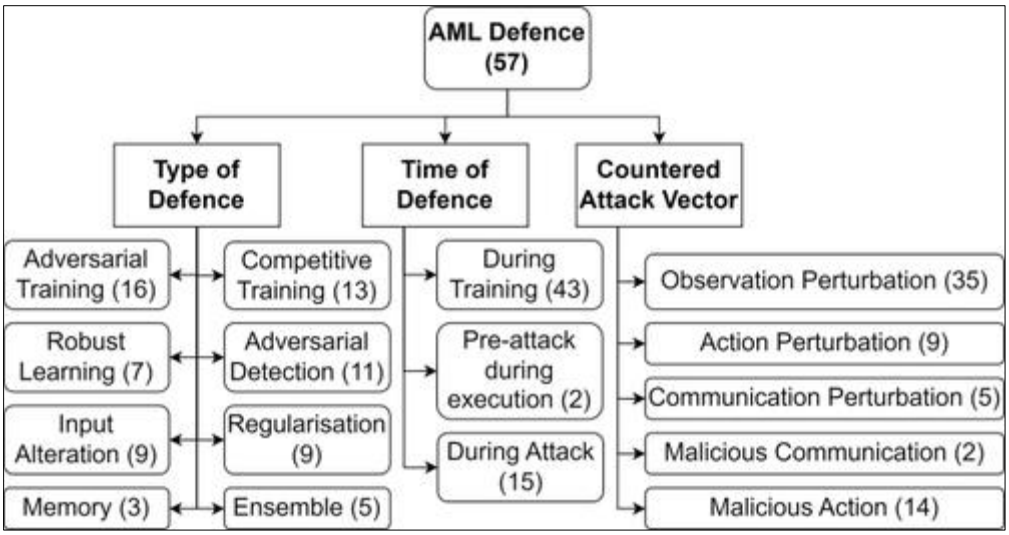

Figure 1 Classification of AML defenses for DRL and the number of papers considering those categories.

Adversarial Training retrains the original agent against the AML attack. Adversarial training occurs during the training phase of the machine learning pipeline, and its main purpose is to counter Observation Perturbations. It has also been shown to be effective in countering Communication Perturbations. Table 1 presents the adversarial training techniques, the attack vector that countered the defense's impact on performance, and the framework the paper used to present the defense. A defense's impact on a model's clean performance shows potential benefits or drawbacks. We use five classifications, namely, positive and negative, for defenses that improve or degrade clean performance, respectively; no change, for defenses that do not significantly change the performance; mixed, for defenses that show a mixture of positive and negative changes for different evaluation conditions, and not evaluated, for papers that do not include enough information to allow a comparison of the clean performance of their defense.

Table 1 Adversarial Training Defences

Name of Defence	**Countered Attack Vectors**	**Impact on**	
Performance	Framework		
Adversarial training	Observation Perturbations	Negative	MDP
Robustifying models	Communication Perturbations	Positive	N/A
DQWAE	Observation Perturbations	Positive	MDP
Adversarial training	Observation Perturbations	Not evaluated	MDP
SA DRL	Observation Perturbations	Positive	SA-MDP
RADIAL-RL	Observation Perturbations	Mixed	MDP
Robust training	Observation Perturbations	Positive	MDP
PA-ATLA	Observation Perturbations	Mixed	MDP
CIQ	Observation Perturbations	No Change	POMDP-IO
BCL	Observation Perturbations	Negative	MDP
RMA3C	Observation Perturbations	Not evaluated	SAM
Semi-Contrastive Adversarial Augmentation	Observation Perturbations	Not evaluated	Goal-Conditioned MDP
SAFER	Observation Perturbations	Mixed	CMDP

We exclude online training from this category despite its potential for performing adversarial training during execution. Online training against an adversary poses a risk as the adversary may influence the data collected from the

environment. The algorithm will use this adversary-influenced data during the online training process. An adversary may intentionally poison this training data to cause the algorithm to learn a poor policy. Thus, online adversarial training is needed to handle data poisoning attacks before they can be deployed as an effective defense.

Robust Learning increases the robustness of an agent in an environment. These methods do not consider specific attacks but instead consider robustness against specific adversary capabilities, such as the perturbation magnitude. Some of these approaches can also certify the robustness of agents' decisions against a certain level of adversarial capability. Table 9 shows robust learning techniques and their impact on performance.

Table 2 Robust Learning Defences

Name of Defences	Impact on Performance
A2PD	Positive
CARL	Not evaluated
CROP	Not evaluated
CPPO	Positive
TRC	Positive
PATROL	Positive
ReCePS	Positive

A limitation of our work is that our data collection only found papers that considered improved robustness in defending against an AML attack. The robust learning category could be extended to consider approaches that do not consider specific AML adversary capabilities but instead look at improving an agent's robustness to other conditions, such as against non-stationary environments.

3. Methodology

3.1. Research Design

This study employs an experimental approach to investigate the effectiveness of AI-driven cybersecurity models against adversarial attacks. The research simulates adversarial attacks on AI-based security systems to evaluate their resilience, identify vulnerabilities, and test advanced defense mechanisms. The experimental setup includes designing and executing adversarial attacks such as evasion attacks, poisoning attacks, and model inversion attacks on machine learning models commonly used for cybersecurity, such as deep neural networks (DNNs), random forests, and support vector machines (SVMs). The study will assess the models' Accuracy, false positive/negative rates, and robustness under adversarial conditions. By leveraging simulation environments, this research aims to provide empirical insights into the strengths and weaknesses of existing AI-based cybersecurity defenses and propose improvements.

3.2. Data Collection & Sources

The study will rely on publicly available cybersecurity datasets that contain real-world and simulated network traffic data, intrusion attempts, and malicious activities. The primary datasets include:

- NSL-KDD Dataset – A widely used dataset for network intrusion detection, containing labeled instances of normal and attack traffic. It improves upon the older KDD'99 dataset by removing redundant and duplicate records, making it more suitable for evaluating AI-driven intrusion detection models.
- CICIDS Dataset (Canadian Institute for Cybersecurity Intrusion Detection System Datasets) – A set of modern intrusion detection datasets that provide realistic network traffic, including normal behavior and a wide range of cyberattacks. These datasets are valuable for Training and testing AI models to detect evolving cyber threats.

Additional data sources may include open-source repositories, security research databases, and real-time threat intelligence feeds to ensure the diversity and representativeness of adversarial attack scenarios. Data preprocessing techniques such as feature selection, normalization, and augmentation will be applied to enhance model training and

evaluation. The experimental results derived from these datasets will inform recommendations for improving AI-driven cybersecurity defenses against adversarial attacks.

3.3. Real-world adversarial attack logs

3.3.1. Adversarial attacks on pervasive applications of industrial interest

In the last section of this article, we will mainly introduce the typical attack algorithms and methods. Most were initially designed for image classification tasks. However, these methods can also be applied to other domains, such as image/video segmentation, 3D recognition, audio recognition, and reinforcement learning, attracting growing attention from academia and industry. Besides, specific data and applications could lead to unique adversarial attacks. Hence, in this section, we sketch these unique adversarial attacks on the other pervasive applications.

Adversarial attacks on semantic Segmentation models

Xie et al. were the first to propose a systematic algorithm— dense adversarial generation (DAG)—to generate adversarial plans for object-detection and segmentation tasks, as shown in The basic idea of DAG is to consider all the targets in the detection/segmentation task simultaneously and optimize the overall loss. Moreover, to tackle the larger number of proposals in the pixel-level object-detection task (i.e., scaling in O K2, where K is the number of pixels), DAG preserves an increased but reasonable number of proposals by changing the intersection-over-union rate in the optimization process. In Ref., the authors observe that for the segmentation task, the relationship between the widely used adversarial losses and the accuracy is not as well-established as in the classification task. Therefore, they propose a new surrogate loss called Houdini to approximate the real adversarial loss, which is the product of a stochastic margin and a task loss. The stochastic margin characterizes the difference between the predicted probability of the ground truth and that of the expected target. The task loss is independent of the model, corresponding to the maximization objective. Also, it further derives an approximation for the gradient of the new surrogate loss concerning the input to enable the gradient-based optimization over the input. Experiments show that Houdini achieves state-of-the-art attack performance on semantic segmentation, making adversarial perturbations more imperceptible to human vision.

Adversarial attacks on 3D recognition

Point-cloud is an important 3D data representation for 3D object recognition. PointNet, PointNet++, and dynamic graph CNN (DGCNN) are the three most popular DL models for point-cloud-based classification/segmentation. However, these three models were also recently found vulnerable to adversarial attacks. In Ref., the authors first extend the C&W attack to the 3D point-cloud DL models. The point locations correspond to the pixel values, and the C&W loss is optimized by shifting the points (i.e., perturbing the point locations). Similarly, the work proposed in Ref. applies BIM/PGD to point-cloud classification and achieves high attack success rates. In Ref., the authors propose a new attack by dropping the existing points in the point clouds. They approximate the contribution of each point to the classification result by point-shifting to the center of the point cloud and dropping the points with large positive contributions. With a certain number of points dropped, the classification accuracy of PointNet, PointNet++, and DGCNN is significantly reduced. Besides, works in Ref. propose to add adversarial perturbations on 3D meshes such that the 2D projections of the 3D meshes can mislead 2D-image classification models. This 3D attack is implemented by optimizing a hybrid loss with the adversarial loss to attack the target 2D model and a penalty loss to keep the 3D adversarial meshes perceptually realistic.

Adversarial attacks on audio and text recognition

Carlini and Wagner successfully constructed high-quality audio adversarial samples by optimizing the C&W loss. Up to 50 words in the text translation can be modified for an audio signal by only adversarial perturbing 1% of the audio signal on DeepSpeech. They also found that the constructed adversarial audio signals are robust to pointwise noise and MP3 compression. However, due to the nonlinear effects of microphones and recorders, the perturbed audio signals do not remain adversarial after being played over the air. The authors propose simulating the nonlinear effects and the noise while considering them in the attack process. Specifically, the authors model the received signal as a function of the transmitted signal, which consists of transformations for modeling the effects of the band-pass filter, impulse response, and white Gaussian noise. The adversarial loss is defined in the received signals instead of the transmitted signals. The proposed attack successfully generates adversarial audio samples in the physical world, which can attack the audio-recognition models even after being played in the air. Liang et al. propose three word-level perturbation strategies on text data for text recognition: insertion, modification, and removal. The attack first identifies the important text items for classification and then exploits one of the perturbation approaches on those text items. Experiments show that this attack can successfully fool some state-of-the-art DNN-based text classifiers. Moreover, TextBugger adopts five types of perturbation operations on text data, including insertion, deletion, swap, character substitution, and word substitution, as shown in Fig. 7. In the white-box setting, those five operations are also conducted on the important

words identified by the Jacobian matrix. However, in the black-box threat model, the Jacobian loss J(h, x, y) with the parameters h, the input of the policy x, and a weighted score over all possible actions y. FGSM is used to attack feed-forward policies trained with three algorithms: Deep Q-networks, asynchronous advantage actor-critic, and trust region policy optimization.

In most cases, the proposed attack can reduce the agent's accuracy by 50% under the white-box setting. In the black-box setting, this attack is also effective. The adversarial effects can transfer across those three algorithms, although the attack performance may degrade. Ref. Proposes perturbing the input states st in the Q-function Q (st+1, a, ht), such that the learning process will produce an adversarial action a'. FGSM and JSMA are nominated as the adversarial-perturbation-crafting method. Lin et al. propose two attack tactics for deep reinforcement learning: the strategically timed attack and the enchanting attack. In the strategically timed attack, the reward is minimized by only perturbing the image inputs for a few specific time steps. This attack is conducted by optimizing the perturbations over the reward. The enchanting attack adversarially perturbs the image frames to lure the agent to the target state. This attack requires a generative model to predict future states and actions to formulate a misleading sequence of actions as guidance for generating perturbations in the image frames.

3.3.2. Adversarial defenses

This section summarizes the representative defenses developed in recent years, mainly including adversarial Training, randomization-based schemes, denoising methods, provable defenses, and some other new defenses. We also briefly discuss their effectiveness against different attacks under different settings.

Adversarial Training

Adversarial Training is an intuitive defense method against adversarial samples, which attempts to improve the robustness of a neural network by training it with adversarial samples. For- mally, it is a min-max game that can be formulated as follows:

Matrix is unavailable on sentences and documents. The adversary is assumed to have access to the confidence values of the prediction. In this context, the importance of each sentence is defined as its confidence value regarding the predicted class. The importance of each word in the most salient sentence is determined by the difference between the confidence values of the sentence with and without the word.

1. Adversarial attacks on deep reinforcement learning

Huang et al. show that existing attack methods can also be used to degrade the performance of the trained policy in deep reinforcement learning by adding adversarial perturbations to the policy's raw inputs. In particular, the authors construct a surrogate where J(h, x', y) is the adversarial loss, with network weights h, adversarial input x,' and ground-truth label y. D(x, x') represents a certain distance metric between x and x'. The inner maximization problem is to find the most effective adversarial samples, which are solved by a well-designed adversarial attack, such as FGSM and PGD. Outer minimization is the standard training process for minimizing loss. The resulting network is supposed to resist the adversarial attack used for the adversarial sample generation in the training stage. Recent studies in Refs. Show that adversarial training is one of the most effective defenses against adversarial attacks. In particular, it achieves state-of-the-art accuracy on several benchmarks.

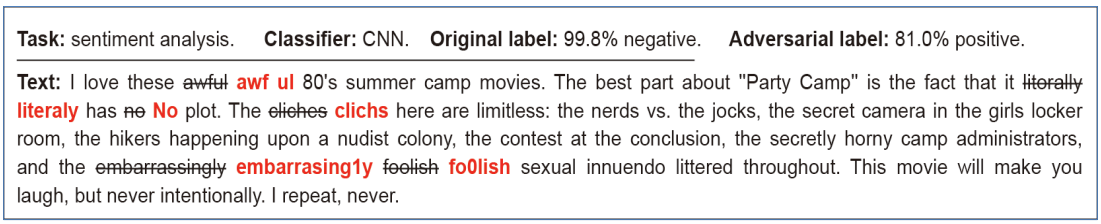

Figure 2 Adversarial text generated by TextBugger: A negative comment is misclassified as a positive commentTherefore, this section elaborates on the best-performing adversarial training techniques in the past few years

FGSM adversarial training

Goodfellow et al. first propose enhancing the robustness of a neural network by training it with benign and FGSM-generated adversarial samples. Formally, the proposed adversarial objective can be formulated as follows:

~J(h, x, y) = cJ(h, x, y) + (1 — c)J(h, x + s • sign[∇xJ(h, x, y)], y) Where x + s • sign[∇xJ(h, x, y)] is the FGSM-generated adversarial sample for the benign sample x, and c is used to balance the Accuracy on benign and adversarial samples as a hyperparameter. Experiments in Ref. show that the network becomes somewhat robust to FGSM-generated adversarial samples. Specifically, with adversarial Training, the error rate on adversarial samples dramatically fell from 89.4% to 17.9%. However, the trained model is still vulnerable to iterative/optimization-based adversarial attacks despite its effectiveness when defending FGSM-generated adversarial samples. Therefore, several studies further dig into adversarial Training with stronger adversarial attacks, such as BIM/PGD attacks.

PGD adversarial training

Extensive evaluations demonstrate that a PGD attack is probably a universal first-order L∞ attack. If so, model robustness against PGD implies resistance against a wide range of first-order L∞ attacks. Based on this conjecture, Madry et al. propose using PGD to train a robust network adversarially. Surprisingly, PGD adversarial training improves the robustness of CNNs and ResNets against typical first-order L∞ attacks, such as FGSM, PGD, and CW∞ attacks under black-box and white-box settings. Even the currently strongest L∞ attack, that is, DAA, can only reduce the Accuracy of the PGD adversarially trained MNIST model to 88.56% and the Accuracy of the CIFAR-10 model to 44.71%. In the recent Competition on Adversarial Attacks and Defenses (CAADs), the first-ranking defense against ImageNet adversarial samples relied on PGD adversarial training. With PGD adversarial training, the baseline ResNet achieves over 50% accuracy under 20-step PGD, while the denoising architecture proposed in Ref. Only increases the Accuracy by 3%. All the above studies and results indicate that PGD adversarial training is the most effective countermeasure against L∞ attacks. However, due to the large computational cost required for PGD adversarial sample generation, PGD adversarial training is inefficient. For example, PGD adversarial training on a simplified ResNet for CIFAR-10 requires approximately three days on a TITAN V graphics processing unit (GPU), and the first-ranking model in CAAD costs 52 hours on 128 Nvidia V100 GPUs. Besides, a PGD adversarially trained model is only robust to L∞ attacks and is vulnerable to other Lp-norm adversaries, such as EAD and CW2

4. Results

4.1. Performance Comparison of AI Defense Models Under Adversarial Attack

Several AI models are employed in cybersecurity for intrusion detection, malware classification, and network anomaly detection. However, their vulnerability to adversarial attacks varies depending on their architecture, training methodology, and defense mechanisms.

Table 3 Traditional AI Models vs. Robust AI Models

Model Type	Advantages	Limitations	Vulnerability to Adversarial Attacks
Deep Neural Networks (DNNs)	High accuracy in normal conditions, learns complex patterns	Highly vulnerable to adversarial perturbations	High
Random Forest (RF)	Less affected by adversarial noise, interpretable	Struggles with high-dimensional data	Moderate
Support Vector Machines (SVMs)	Strong generalization, effective in small datasets	Inefficient in large-scale real-time threat detection	Moderate
Graph Neural Networks (GNNs)	Effective in network-based threat detection	Computationally intensive	Moderate to High
Ensemble Learning Models	Combines multiple classifiers for higher robustness	Increased computational complexity	Lower than individual models
Adversarially Trained Models	Specifically trained to resist adversarial examples	The trade-off with Accuracy on clean data	Low

Table 4 Performance Metrics for Different Models Under Adversarial Attacks

Model	Standard Accuracy (%)	Adversarial Accuracy (%)	Robustness Score	False Positive Rate (FPR)	Computational Cost
DNN	95	42	Low	12%	High
RF	89	60	Moderate	8%	Moderate
SVM	86	55	Moderate	7%	Moderate
GNN	92	50	Moderate-High	10%	High
Ensemble Learning	94	72	High	5%	High
Adversarially Trained DNN	92	80	Very High	6%	Very High

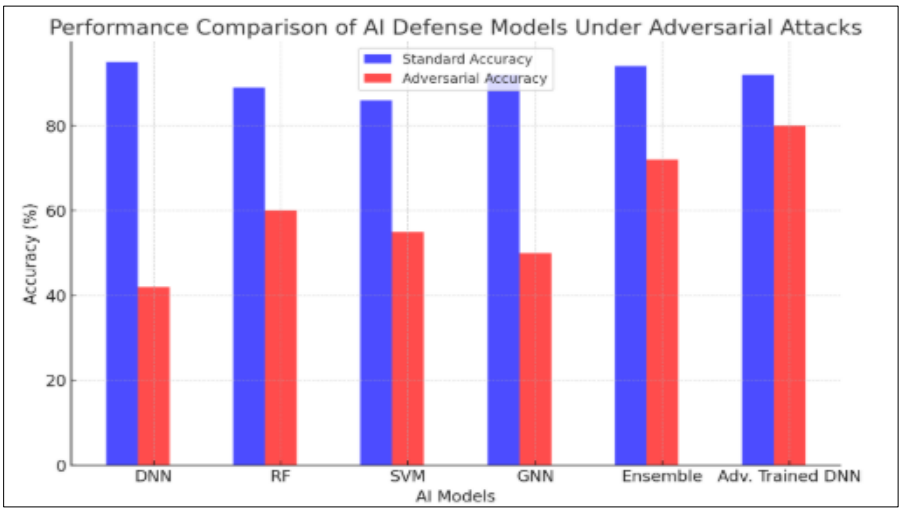

Figure 3 performance comparison of AI defense models under adversarial Attacks

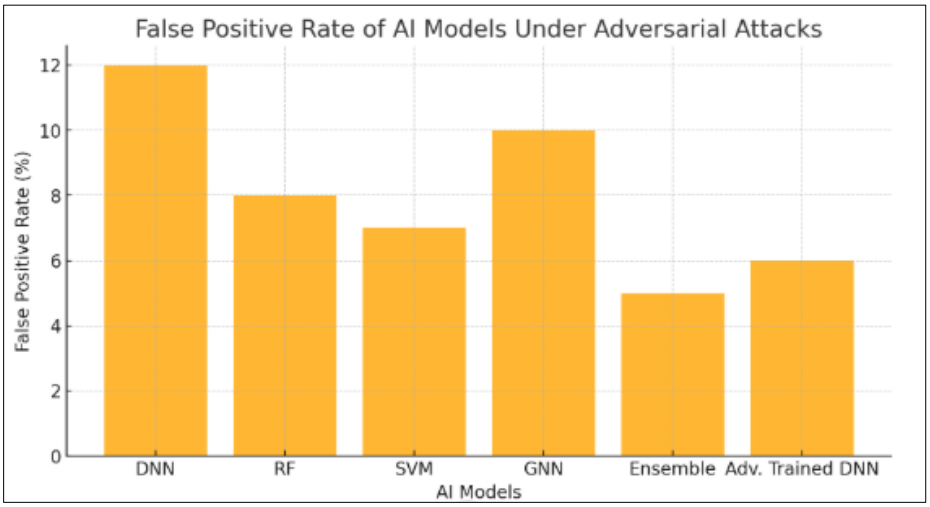

Figure 4 False Positive Rate of Ai models under Adversarial Attacks

4.1.1. Key Insights

- Standard DNNs suffer the most under adversarial attacks, with adversarial Accuracy dropping significantly.
- Ensemble learning and adversarial Training improve robustness but come at a higher computational cost.

- Random Forest and SVM provide moderate robustness but may lack scalability in large-scale cybersecurity applications.

4.2. Effectiveness of Mitigation Strategies

Various defense strategies have been developed to counter adversarial attacks, each with varying levels of effectiveness, resource requirements, and trade-offs.

Table 5 Comparison of Different Mitigation Strategies

Mitigation Strategy	Effectiveness Against Attacks	Computational Overhead	Impact on Standard Accuracy	Applicability in Real-Time Systems
Adversarial Training	High (mitigates evasion attacks)	High	There is a slight drop in clean Accuracy	Moderate (real-time infeasible for large-scale models)
Defensive Distillation	Moderate (obfuscates gradients)	Moderate	Minimal impact	Good
Feature Squeezing	Moderate (reduces adversarial noise)	Low	Minimal impact	High
Gradient Masking	Low (adaptive attacks can bypass)	Low	Minimal impact	High
Ensemble Defense Models	High (resilient to multiple attack types)	High	There is a slight drop in clean Accuracy	Moderate
AI-Augmented Human Analysis	Very High (combines AI with expert review)	High	Minimal impact	Low (slow due to manual intervention)

Table 6 Effectiveness of Defense Mechanisms Against Specific Attacks

Defense Mechanism	Evasion Attack Resilience	Poisoning Attack Resilience	Model Extraction Prevention	Computational Cost
Adversarial Training	High	Moderate	Low	High
Defensive Distillation	Moderate	Low	Low	Moderate
Feature Squeezing	Moderate	Low	Low	Low
Ensemble Models	High	High	Moderate	High
Robust Federated Learning	High	High	High	Very High

4.2.1. Key Observations

- Adversarial Training remains one of the most effective strategies but is computationally expensive.
- Defensive Distillation is a moderate solution but can be bypassed by sophisticated attacks.
- Ensemble models provide high resilience at the cost of increased processing time.
- Feature squeezing and gradient masking offer lightweight solutions but are ineffective against complex attacks.

4.3. Key Observations and Trends in Adversarial Behavior

4.3.1. Trends in Adversarial Attacks

Increased Sophistication of Adversarial Attacks

- Attackers use adaptive adversarial attacks that dynamically modify perturbations based on real-time defense mechanisms.

- Black-box attacks are becoming more prevalent, where attackers exploit models without direct access to their parameters.

4.3.2. Rise in Transferability of Adversarial Examples

- Adversarial samples crafted against one model often deceive other AI models, making model-specific defenses less effective.

4.3.3. Growing Threat of Data Poisoning

- Attackers manipulate training datasets to introduce hidden backdoors, compromising model reliability from the learning phase.

4.3.4. Emergence of Model Extraction and Privacy Attacks

- Attackers use model inversion techniques to reconstruct training data and extract sensitive patterns from AI models.

4.3.5. Defensive Trends and Future Directions

Hybrid AI-Based Security Models

- Combining symbolic reasoning with deep learning is being explored to improve AI interpretability and robustness.

4.3.6. Adversarially Robust AI Frameworks

- Research focuses on AI models that can self-adjust to evolving adversarial threats through reinforcement learning.

4.3.7. Explainable AI (XAI) for Cybersecurity

- Efforts are being made to make AI decisions more transparent, helping cybersecurity experts analyze adversarial threats effectively.

4.3.8. AI-Augmented Human-in-the-Loop Security Systems

- AI models are increasingly combined with human security analysts to improve decision-making and threat mitigation.

4.3.9. Quantum Machine Learning for Cybersecurity

- Quantum AI techniques are being studied for their potential to resist adversarial perturbations due to the randomness of quantum states.

5. Discussion

5.1. Interpretation of Findings

Analyzing AI-based cybersecurity defenses against adversarial machine learning (AML) attacks reveals significant insights into their strengths and weaknesses. The findings indicate that while certain models exhibit high accuracy in normal conditions, their resilience to adversarial attacks varies widely.

5.1.1. Strengths and Weaknesses of AI Defenses

Strengths

- Adversarially Trained Models: These models, particularly adversarially trained deep neural networks (DNNs), demonstrate strong resilience against adversarial attacks by learning to recognize adversarial perturbations.
- Ensemble Learning Models: Combining multiple classifiers improves robustness and reduces false positive rates, making them more resilient to attacks.

- Graph Neural Networks (GNNs): GNNs perform well in network-based threat detection and exhibit moderate resistance to adversarial modifications.

Weaknesses

- Standard Deep Neural Networks (DNNs): Highly vulnerable to adversarial perturbations, leading to significant drops in adversarial accuracy.
- Computational Overhead: Models like adversarially trained DNNs and ensemble learning approaches require significant computational resources, limiting real-time deployment in large-scale cybersecurity applications.
- False Positive Rates: Some models, such as GNNs and traditional DNNs, suffer from relatively high false positive rates, leading to unnecessary security alerts and operational inefficiencies.

5.2. Implications for Cybersecurity

The evolving nature of adversarial attacks presents a significant challenge for AI-powered defense mechanisms, requiring continuous advancements in cybersecurity strategies. As attackers develop increasingly sophisticated adversarial techniques, AI defenses must adapt accordingly to remain effective. Static AI models are insufficient for long-term protection, making dynamic learning approaches, such as reinforcement learning-based adversarial training, more viable for enhancing resilience. Additionally, industries that rely on AI-driven cybersecurity, including finance, healthcare, and defense, face heightened security risks, necessitating the integration of more robust adversarial defense mechanisms to prevent catastrophic breaches. However, implementing such defenses often comes at the cost of increased computational complexity, creating a critical trade-off between security and performance. Organizations must carefully balance these factors to ensure strong protection without compromising system efficiency.

5.3. Industry and Policy-Level Recommendations

5.3.1. For Industry

To strengthen AI-powered cybersecurity against adversarial threats, organizations should adopt hybrid AI security models that combine symbolic reasoning with deep learning, enhancing interpretability and robustness. Implementing multi-layered defense mechanisms is also crucial, as AI models alone may not be sufficient; integrating traditional cybersecurity approaches such as intrusion detection systems (IDS), behavioral analysis, and human-in-the-loop monitoring can provide an added layer of protection. Enhancing model explainability through Explainable AI (XAI) techniques can help cybersecurity professionals better understand and mitigate adversarial threats. Regular adversarial testing, including red-team exercises, should also be conducted to simulate attacks and assess the resilience of AI security systems, ensuring continuous improvements in defense strategies.

5.3.2. For Policy-Makers

Governments play a crucial role in strengthening AI-powered cybersecurity by establishing regulatory frameworks that enforce robust security standards and ensure organizations implement effective adversarial defenses. Mandating secure AI model training is essential, requiring AI systems to undergo rigorous testing against adversarial attacks before deployment in critical sectors such as finance, healthcare, and defense. Fostering public-private collaboration can accelerate advancements in AI cybersecurity by facilitating information sharing between government agencies, research institutions, and private organizations. This collaborative approach enhances threat intelligence, promotes the development of resilient AI models, and helps create a unified defense strategy against evolving cyber threats.

5.4. Future Challenges in Adversarial Machine Learning

5.4.1. Evolving Attack Techniques

- Attackers increasingly use adaptive adversarial attacks, where perturbations evolve in real time to bypass AI defenses.
- The transferability of adversarial examples means that an attack designed for one model may also work on others, making targeted defenses less effective.

5.4.2. Scalability and Efficiency Issues

- Many adversarial defense mechanisms require significant computational resources, limiting their real-world applicability.
- Real-time adversarial detection remains challenging, as sophisticated attacks often evade detection without disrupting normal operations.

5.4.3. Ethical and Legal Concerns

- The development of adversarial AI techniques raises ethical concerns, particularly regarding their misuse in cyber warfare, misinformation campaigns, and AI-driven fraud.
- The lack of global regulatory standards on adversarial AI defense frameworks leads to inconsistent cybersecurity protections across industries.

5.5. Limitations of the Study

While this study provides valuable insights into adversarial machine learning and AI defenses, it is subject to several limitations:

AI-driven cybersecurity faces several limitations that impact its effectiveness against adversarial threats. One major constraint is dataset diversity, as training and testing datasets may not fully capture the complexity of real-world cyber threats. Many publicly available datasets primarily focus on known attack patterns, while adversarial machine learning threats continuously evolve, making it difficult to prepare AI models for emerging attack strategies. Additionally, model generalization issues pose a significant challenge, as AI models trained on specific datasets may struggle to detect new and unseen attack techniques. While transfer learning and meta-learning approaches offer potential solutions for improving adaptability, further research is required to enhance their effectiveness. Another limitation is computational constraints, as some high-performing adversarial defense techniques, such as ensemble models and adversarial Training, demand substantial computational resources. This makes them less feasible for organizations with limited infrastructure, highlighting the need for optimized, resource-efficient defense strategies.

6. Conclusion

This study has explored the evolving landscape of AI-powered cybersecurity, emphasizing the vulnerabilities of AI-driven security systems to adversarial threats and the need for continuous adaptation. Despite their efficiency in threat detection, key findings reveal that AI models face challenges such as dataset diversity constraints, model generalization issues, and high computational demands for advanced defenses. As adversarial attacks grow more sophisticated, AI security strategies must incorporate hybrid models, explainable AI (XAI), and multi-layered defense mechanisms. Continuous model adaptation is crucial to enhance AI-powered cybersecurity, requiring reinforcement learning-based adversarial training, meta-learning approaches, and self-learning AI systems that can dynamically adjust to emerging threats. Integrating human-AI hybrid security models, where human analysts collaborate with AI-driven systems, can improve decision-making accuracy and threat mitigation. Future research should focus on quantum-resistant adversarial defenses, as quantum computing advancements threaten traditional cryptographic security, necessitating post-quantum cryptography and quantum-enhanced AI models. Ethical considerations in adversarial AI must also be addressed to ensure transparency, fairness, and accountability, preventing misuse while fostering responsible AI deployment. Strengthening AI security requires a holistic approach that combines technological advancements with strategic policy interventions, ensuring resilient and trustworthy AI-driven cybersecurity systems capable of countering evolving adversarial threats.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Bhardwaj, A., Alshehri, M. D., Kaushik, K., Alyamani, H. J., & Kumar, M. (2022). (Retracted) Secure framework against cyber-attacks on cyber-physical robotic systems. Journal of Electronic Imaging, 31(6), 061802-061802.

[2] Chithaluru, P., Al-Turjman, F., Kumar, M., & Stephan, T. (2023). Computational-intelligence-inspired adaptive opportunistic clustering approach for industrial IoT networks. IEEE Internet of Things Journal, 10(9), 7884-7892.

[3] Rouquerol, J., Avnir, D., Fairbridge, C. W., Everett, D. H., Haynes, J. M., Pernicone, N. ... & Unger, K. K. (1994). Recommendations for the characterization of porous solids (Technical Report). Pure and applied chemistry, 66(8), 1739-1758.

[4] Tabassi, E., Burns, K. J., Hadjimichael, M., Molina-Markham, A. D., & Sexton, J. T. (2019). A taxonomy and terminology of adversarial machine learning. NIST IR, 2019, 1-29.

[5] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.

[6] Wang, Y., Sun, T., Li, S., Yuan, X., Ni, W., Hossain, E., & Poor, H. V. (2023). Adversarial attacks and defenses in machine learning-empowered communication systems and networks: A contemporary survey. IEEE Communications Surveys & Tutorials, 25(4), 2245-2298.

[7] Shaham, U., Stanton, K. P., Zhao, J., Li, H., Raddassi, K., Montgomery, R., & Kluger, Y. (2017). Removal of batch effects using distribution-matching residual networks. Bioinformatics, 33(16), 2539-2546.

[8] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). ªExplaining and harnessing adversarial examples, º. Proc. ICLR'15.

[9] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. 6th Int. In Conf. Learn. Represent. ICLR.

[10] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016, March). The limitations of deep learning in adversarial settings. In 2016 IEEE European symposium on security and privacy (EuroS&P) (pp. 372-387). IEEE.

[11] Carlini, N., & Wagner, D. (2017, May). Towards evaluating the robustness of neural networks. 2017, the symposium on security and privacy (sp) was held (pp. 39-57). Ieee.

[12] Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial machine learning at scale (2016). arXiv preprint arXiv:1611.01236.

[13] Gouk, H., Frank, E., Pfahringer, B., & Cree, M. J. (2021). Regularisation of neural networks by enforcing Lipschitz continuity. Machine Learning, 110, 393-416.

[14] Folz, J., Palacio, S., Hees, J., & Dengel, A. (2020, March). Adversarial defense based on structure-to-signal autoencoders. In 2020 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 3568-3577). IEEE.

[15] Guembe, B., Azeta, A., Misra, S., Osamor, V.C., Fernandez-Sanz, L. and Pospelova, V. (2022) The Emerging Threat of AI-Driven Cyber Attacks: A Review. Applied Artificial Intelligence, 36, Article 2037254.

[16] https://doi.org/10.1080/08839514.2022.2037254

[17] Sun, G., Cong, Y., Dong, J., Wang, Q., Liu, L. and Liu, J. (2022) Data Poisoning Attacks on Federated Machine Learning. IEEE Internet of Things Journal, 9, 11365-11375.

[18] https://doi.org/10.1109/jiot.2021.3128646

[19] Tufail, S., Batool, S. and Sarwat, A.I. (2021) False Data Injection Impact Analysis in AI-Based Smart Grid. SoutheastCon 2021, Atlanta, 10-13 March 2021, 1-7.

[20] https://doi.org/10.1109/southeastcon45413.2021.9401940

[21] De Mello, F.L. (2020) A Survey on Machine Learning Adversarial Attacks. Journal of Information Security and Cryptography (Enigma), 7, 1-7.

[22] https://doi.org/10.17648/jisc.v7i1.76

[23] Sadeghi, K., Banerjee, A. and Gupta, S.K.S. (2020) A System-Driven Taxonomy of Attacks and Defenses in Adversarial Machine Learning. IEEE Transactions on Emerging Topics in Computational Intelligence, 4, 450-467.

[24] https://doi.org/10.1109/tetci.2020.2968933

[25] Ramirez, M.A., Kim, S.K., Hamadi, H.A., Damiani, E., Byon, Y.J., Kim, T.Y., Yeun, C.Y., et al. (2022) Poisoning Attacks and Defenses on Artificial Intelligence: A Survey. arXiv: 2202.10276.

[26] https://doi.org/10.48550/arXiv.2202.10276

[27] Sahai, A.K., & Rath, N. (2021). Artificial intelligence and the 4th industrial revolution. In Artificial intelligence and machine learning in business management (pp. 127-143). CRC Press.

[28] Allam, Z., & Allam, Z. (2021). Big data, artificial intelligence, and the rise of autonomous smart cities. The rise of autonomous smart cities: technology, economic performance and climate resilience, 7-30.

[29] Montasari, R. (2022). Cyber threats and national security: the use and abuse of artificial intelligence. In Handbook of Security Science (pp. 679-700). Cham: Springer International Publishing.

[30] Cohen, S.A. (2019). Cybersecurity for critical infrastructure: addressing threats and vulnerabilities in Canada.

[31] Tremont, T.M. (2023). Human-AI: Using Threat Intelligence to Expose Deepfakes and the Exploitation of Psychology (Doctoral dissertation, Capitol Technology University).

[32] Khatun, M.A., Memon, S.F., Eising, C., & Dhirani, L.L. (2023). Machine Learning for Healthcare-IoT Security: A Review and Risk Mitigation. IEEE Access.

[33] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). ªExplaining and harnessing adversarial examples, º. Proc. ICLR'15.

[34] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. 6th Int. In Conf. Learn. Represent. ICLR.

[35] James Tu, Tsunhsuan Wang, Jingkang Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. 2021. Adversarial attacks on multi-agent communication. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 7748–7757

[36] Yi Han, David Hubczenko, Paul Montague, Olivier De Vel, Tamas Abraham, Benjamin I. P. Rubinstein, Christopher Leckie, Tansu Alpcan, and Sarah Erfani. 2020. Adversarial reinforcement learning under partial observability in autonomous computer network defense. In Proceedings of the International Joint Conference on Neural Networks. 1–8.

[37] Panagiota Kiourti, Kacper Wardega, Susmit Jha, and Wenchao Li. 2020. TrojDRL: Evaluation of backdoor attacks on deep reinforcement learning. In Proceedings of the ACM/IEEE Design Automation Conference. 1–6.

[38] Michael Everett, Björn Lütjens, and Jonathan P. How. 2021. Certifiable robustness to adversarial state uncertainty in deep reinforcement learning. IEEE Transactions on Neural Networks and Learning Systems 33, 9 (2021), 4184–4198.

[39] Xinghua Qu, Abhishek Gupta, Yew-Soon Ong, and Zhu Sun. 2023. Adversary agnostic robust deep reinforcement learning. IEEE Transactions on Neural Networks and Learning Systems 34, 9 (2023), 6146–6157.

[40] Fan Wu, Linyi Li, Zijian Huang, Yevgeniy Vorobeychik, Ding Zhao, and Bo Li. 2022. CROP: Certifying robust policies for reinforcement learning through functional smoothing. In Proceedings of the International Conference on Learning Representations.

[41] Erwan Lecarpentier and Emmanuel Rachelson. 2019. Non-stationary Markov decision processes a worst-case approach using model-based reinforcement learning. In Proceedings of the Advances in Neural Information Processing Systems, Vol. 32.

[42] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention; 2015 Oct 5–9; Munich, Germany; 2015. p. 234–41.

[43] Grundmann M, Kwatra V, Han M, Essa I. Efficient hierarchical graph-based video segmentation. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2010 Jun 13–18; San Francisco, CA, USA; 2010. p. 2141–8.

[44] Su H, Maji S, Kalogerakis E, Learned-Miller E. Multi-view convolutional neural networks for 3D shape recognition. In: Proceedings of the IEEE International Conference on Computer Vision; 2015 Dec 7–13; Santiago, Chile; 2015. p. 945–53.

[45] Qi CR, Su H, Mo K, Guibas LJ. PointNet: deep learning on point sets for 3D classification and segmentation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA; 2017. p. 652–60.

[46] Lee H, Pham P, Largman Y, Ng AY. Unsupervised feature learning for audio classification using convolutional deep belief networks. In: Proceedings of the 23rd Conference on Neural Information Processing Systems; 2009 Dec 7–10; Vancouver, BC, Canada; 2009. p. 1096–104.

[47] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, et al.

[48] Human-level control through deep reinforcement learning Nature, 518 (7540) (2015), pp. 529-533

[49] Xie C, Wang J, Zhang Z, Zhou Y, Xie L, Yuille A. Adversarial examples for semantic segmentation and object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision; 2017 Oct 22–29; Venice, Italy; 2017. p. 1369–78.

[50] Cisse M, Adi Y, Neverova N, Keshet J. Houdini: fooling deep structured prediction models. 2017. arXiv:1707.05373.

[51] Qi CR, Yi L, Su H, Guibas LJ. PointNet+: deep hierarchical feature learning on point sets in a metric space. In: Proceedings of the 31st Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA; 2017. p. 5099–108.

[52] Wang Y, Sun Y, Liu Z, Sarma SE, Bronstein MM, Solomon JM. Dynamic graph CNN for learning on point clouds. 2018. arXiv:1801.07829.

[53] Xiang C, Qi CR, Li B. Generating 3D adversarial point clouds. 2018. arXiv:1809.07016.

[54] Liu D, Yu R, Su H. Extending adversarial attacks and defenses to deep 3D point cloud classifiers. 2019. arXiv:1901.03006.

[55] Xiao C, Yang D, Li B, Deng J, Liu M. MeshAdv: adversarial meshes for visual recognition. 2018. arXiv:1810.05206v2.

[56] Carlini N, Wagner D. Audio adversarial examples: targeted attacks on speech-to-text. In: Proceedings of 2018 IEEE Security and Privacy Workshops; 2018 May 24; San Francisco, CA, USA; 2018. p. 1–7.

[57] Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, et al. Deep speech: scaling up end-to-end speech recognition. 2014. arXiv:1412.5567.

[58] Yakura H, Sakuma J. Robust audio adversarial example for a physical attack. 2018. arXiv:1810.11793.

[59] Liang B, Li H, Su M, Bian P, Li X, Shi W. Deep text classification can be fooled. 2017. arXiv:1704.08006.

[60] Huang S, Papernot N, Goodfellow I, Duan Y, Abbeel P. Adversarial attacks on neural network policies. 2017. arXiv:1702.02284.

[61] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing Atari with deep reinforcement learning. 2013. arXiv:1312.5602.

[62] Mnih V, Badia AP, Mirza M, Graves A, Harley T, Lillicrap TP, et al. Asynchronous methods for deep reinforcement learning. In: Proceedings of the 33rd International Conference on Machine Learning; 2016 Jun 19–24; New York, NY, USA; 2016. p. 1928–37.

[63] Schulman J, Levine S, Moritz P, Jordan M, Abbeel P. Trust region policy optimization. In: Proceedings of the 32nd International Conference on Machine Learning; 2015 Jul 6–11; Lille, France; 2015. p. 1889–97.

[64] Behzadan V, Munir A. Vulnerability of deep reinforcement learning to policy induction attacks. In: Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition; 2017 Jul 15–20; New York, NY, USA; 2017. p. 262–75

[65] Lin YC, Hong ZW, Liao YH, Shih ML, Liu MY, Sun M. Tactics of adversarial attack on deep reinforcement learning agents. 2017. arXiv:1703.06748.

[66] Carlini N, Katz G, Barrett C, Dill DL. Ground-truth adversarial examples. In: ICLR 2018 Conference; 2018 Apr 30; Vancouver, BC, Canada; 2018.

[67] Papernot N, Faghri F, Carlini N, Goodfellow I, Feinman R, Kurakin A, et al. Technical report on the CleverHans v2.1.0 adversarial examples library. 2016. arXiv:1610.00768v6.

[68] Sharma Y, Chen PY. Attacking the Madry defense model with L1-based adversarial examples. 2017. arXiv:1710.10733v4.

[69] Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale. 2016. arXiv: 1611.01236.