Adversarial Threats to AI-Driven Systems: Exploring the Attack Surface of Machine Learning Models and Countermeasures

Abstract

Adversarial attacks pose a critical threat to the reliability of AI-driven systems, exploiting vulnerabilities at the data, model, and deployment levels. This study employs a quantitative analysis using the CIFAR-10 Adversarial Examples Dataset from IBM's Adversarial Robustness Toolbox and the MITRE ATLAS AI Model Vulnerabilities Dataset to assess attack success rates and attack surface exposure. A convolutional neural network (CNN) classifier was evaluated against Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini & Wagner (C&W) attacks, yielding misclassification rates of 42.2%, 65.5%, and 86.8%, respectively. Statistical analysis using the Chi-Square Goodness-of-Fit Test ($p < 0.001$) confirmed a disproportionate targeting of model-level vulnerabilities (53.6%). Countermeasure evaluation demonstrated that adversarial training provided the highest robustness gain (23.29%), while detection algorithms were least effective (15.34%). To enhance AI security, hybrid defense mechanisms integrating adversarial training with real-time anomaly detection should be prioritized, and standardized evaluation benchmarks should be established for AI security testing.

Keywords: adversarial attacks, AI security, model vulnerabilities, adversarial training, machine learning defenses

1. Introduction

Artificial intelligence (AI) has significantly transformed industries, driving advancements in finance, healthcare, cybersecurity, and autonomous systems. However, as AI models grow increasingly complex, they introduce security vulnerabilities, particularly adversarial attacks. Wang et al. (2019) avers that these attacks involve manipulating input data to deceive machine learning (ML) models, leading to incorrect classifications or unintended behaviors. Such vulnerabilities affect AI systems at multiple levels, including data integrity, model architecture, and deployment environments, thereby threatening the reliability and security of AI-driven decision-making processes (Hoang et al., 2024). Given AI's expanding role in critical sectors, addressing these risks and developing effective countermeasures is imperative.

Adversarial threats manifest in various forms, with evasion, poisoning, and model extraction attacks being particularly prevalent. According to Muthalagu et al. (2024), evasion attacks manipulate input data during inference, causing AI models to misclassify objects. A well-documented example involves minor pixel modifications in image recognition systems, which have led to significant misidentifications (Jacquet, 2024). In autonomous vehicles, adversarial perturbations have caused self-driving systems to misinterpret traffic signs, creating severe safety hazards (Giannaros et al., 2023). Poisoning attacks, by contrast, occur during the training phase when adversaries introduce corrupted data to bias the model, thereby reducing its reliability in applications such as fraud detection and cybersecurity monitoring. Hussain et al. (2024) argues that model extraction attacks allow adversaries to reconstruct proprietary AI models by analyzing their outputs, leading to potential exposure of sensitive training data and unauthorized replication of proprietary architectures. Collectively, these attack strategies highlight the urgent need for robust AI security measures.

Empirical research underscores the increasing prevalence of adversarial attacks. Kassianik and Kassianik (2025) reports that in January 2025, researchers from Cisco and the University of Pennsylvania tested DeepSeek's AI model R1, revealing a complete failure in detecting adversarial prompts. The model was unable to block any of the 50 tested malicious inputs, demonstrating a 100 percent attack success rate (McCurdy, 2025). Similarly, a December 2024 study identified vulnerabilities in large language models (LLMs) embedded in AI-powered robotic systems, showing that adversarial inputs could bypass safety mechanisms and induce unintended actions (Fu et al., 2024). These findings highlight the persistent challenges in securing AI models against sophisticated adversarial techniques.

Autonomous transportation is particularly vulnerable to adversarial threats. Mehta et al. (2024) posits that slight modifications to road signs—such as strategically placed stickers—can deceive self-driving systems into misinterpreting stop signs as speed limit signs or failing to detect road hazards. According to Miller et al. (2024) Tesla's Autopilot has demonstrated how adversarial signals manipulate lane recognition and speed detection, leading to hazardous driving behavior. Moreover, Chi et al. (2024) revealed that adversarial signals injected into radar systems caused autonomous vehicles to detect non-existent obstacles, triggering unnecessary braking and erratic driving patterns. These incidents illustrate the dangers of adversarial attacks in AI-powered transportation, necessitating advanced security frameworks to protect autonomous mobility systems.

Beyond vision-based AI applications, adversarial attacks also target natural language processing (NLP) and voice recognition systems. Fakhouri et al. (2024) argues that AI-driven spam filters have been bypassed using adversarially crafted messages designed

to evade detection, while sentiment analysis models have been manipulated to alter interpretations without modifying the perceptible meaning of text. Additionally, researchers have shown that voice assistants such as Alexa and Siri are vulnerable to adversarial commands embedded in slightly altered or inaudible audio signals (Alchekov et al., 2023; Cheng & Roedig, 2022). Alchekov et al. (2023) notes that these manipulations have led to unauthorized actions, including unlocking devices, making unauthorized transactions, and altering system settings. These vulnerabilities raise critical concerns about the security of AI-powered voice recognition technologies.

Statistical evidence further demonstrates the growing impact of adversarial attacks. Kaur (2020) reports that a 2022 Gartner study found 30 percent of AI-related cyberattacks involved training data poisoning, model theft, or adversarial manipulation. Additionally, Javed et al. (2024) indicates that adversarial perturbations can reduce AI model accuracy by up to 90 percent, severely compromising their reliability in high-stakes applications. In the domain of autonomous vehicles, adversarial modifications to road signs have resulted in alarmingly high misclassification rates, reinforcing the necessity of improved AI security protocols.

Efforts to mitigate adversarial threats have led to the development of various defensive strategies. Wang et al. (2019) states that adversarial training, a widely used approach, enhances model robustness by exposing AI systems to adversarial examples during training. While this technique improves resilience, it imposes substantial computational costs and does not offer absolute protection against evolving attack methodologies. Defensive distillation, another countermeasure, reduces model sensitivity to minor perturbations by training AI systems on smoothed probability distributions rather than discrete labels. However, this method has been bypassed by increasingly sophisticated attacks. Khalid et al. (2019) posits that additional countermeasures, such as input preprocessing and feature squeezing, attempt to mitigate adversarial perturbations by filtering input noise, though these methods introduce trade-offs between model accuracy and security.

As adversarial threats continue to evolve, a multi-faceted approach to AI security is essential. Goswami (2024) contends that beyond enhancing model resilience, continuous monitoring and anomaly detection mechanisms are critical for identifying adversarial inputs in real-time. Furthermore, developing adaptive countermeasures capable of dynamically responding to emerging attack strategies is crucial for securing AI systems against malicious interventions. Regulatory bodies and industry stakeholders have recognized AI security as a priority, with reports such as Deloitte's *State of AI in the Enterprise* emphasizing security and safety concerns in AI adoption (Ridzuan et al., 2024; Deloitte, 2024). Given the rapid pace of AI advancements, ongoing research must prioritize model transparency, adversarial robustness, and the

establishment of rigorous evaluation frameworks to mitigate emerging threats effectively. Addressing these challenges is crucial to ensuring the reliability, safety, and trustworthiness of AI-driven systems in critical applications. Considering the foregoing, this study analyzes adversarial threats to AI-driven systems by exploring the attack surface of machine learning models and evaluating effective countermeasures to enhance their security and robustness. The study achieves the following objectives:

1. Investigates the modes of evasion attacks, poisoning attacks, and model inversion/extraction attacks within AI-driven systems.
2. Analyzes the attack surface of machine learning models, identifying vulnerabilities at the data, model, and deployment levels that adversaries exploit.
3. Addresses the effectiveness and limitations of existing defense mechanisms (adversarial training, defensive distillation, and detection algorithms) against adversarial attacks.
4. Proposes strategies for improving the security of AI-driven systems, integrating adaptive, robust countermeasures to mitigate adversarial threats while balancing model performance and efficiency.

2. Literature Review

Adversarial attacks in machine learning (ML) pose substantial risks to the security, integrity, and reliability of AI-driven systems. Wang et al. (2019) posits that these attacks exploit inherent vulnerabilities in ML models, manipulating inputs or training data to induce misclassifications, bias predictions, or extract confidential information. They can be systematically classified into evasion attacks, poisoning attacks, and model inversion or extraction attacks, each targeting distinct phases of an ML system's lifecycle (Muthalagu et al., 2024; Kolade et al., 2025).

Evasion attacks occur during the inference phase, where adversaries introduce carefully crafted perturbations to input data to deceive trained ML models (Wang et al., 2024; Obioha-Val et al., 2025). Ai et al. (2021) asserts that these perturbations, often imperceptible to human observers, can cause significant errors in classification. In image recognition systems, minor pixel modifications have been shown to cause objects to be misidentified (Jacquet, 2024; Obioha-Val et al., 2025). A widely cited example demonstrated that an image of a panda was misclassified as a gibbon due to subtle perturbations (Lapienytė, 2023). Similarly, adversarial inputs in natural language processing (NLP) can manipulate sentiment analysis models or bypass spam filters through minor textual modifications. Alchekov et al. (2023) notes that voice recognition systems are also vulnerable, as adversarial audio signals can embed inaudible commands that alter system behavior. A particularly concerning application of evasion

attacks is in autonomous vehicles, where slight modifications to road signs—such as added stickers—can mislead AI-driven navigation systems, potentially leading to hazardous situations (Mehta et al., 2024; Obioha-Val et al., 2025). These vulnerabilities underscore the need for robust defenses, particularly in safety-critical applications.

Poisoning attacks, by contrast, target the training phase, aiming to corrupt the learning process by injecting malicious data into the training set. Das et al. (2024) contends that by modifying the underlying data distribution, attackers can cause models to produce erroneous classifications post-deployment. In fraud detection systems, adversaries may introduce fraudulent transactions labeled as legitimate, reducing the model's ability to detect financial fraud (Hilal et al., 2021; Adigwe et al., 2024). Similarly, recommendation systems can be manipulated through biased data injection, distorting rankings to promote or suppress specific content (Adomavicius et al., 2019; Alao, Adebiyi and Olaniyi, 2024). In cybersecurity applications, poisoning attacks can degrade the effectiveness of intrusion detection systems, allowing malicious activities to bypass security measures (Kravchik et al., 2022; Arigbabu et al., 2024). These attacks present long-term security risks, as compromised models continue producing flawed outputs even after the initial attack is removed. Given the increasing sophistication of poisoning techniques, detecting and mitigating these threats remains a significant challenge.

Model inversion and extraction attacks pose additional security concerns by compromising the confidentiality of ML models. Shafee and Awaad (2020) explains that model inversion attacks allow adversaries to infer sensitive training data, raising serious privacy risks in biometric authentication systems. For instance, facial recognition models have been exploited to reconstruct images of individuals, threatening privacy and data security (Butt et al., 2023; Balogun et al., 2025). Model extraction attacks, meanwhile, involve adversaries systematically querying an ML model to approximate its internal structure and parameters, enabling unauthorized replication of proprietary AI models (Khazane et al., 2024; Gbadebo et al., 2024). Kumar et al. (2024) asserts that the growing reliance on cloud-based AI services exacerbates these risks, as remote access provides adversaries with opportunities to extract valuable models without direct access to the infrastructure.

The Attack Surface of AI-Driven Systems

The security of AI-driven systems is shaped by vulnerabilities at the data, model, and deployment levels. Rahman et al. (2023) argues that each of these dimensions presents specific weaknesses that adversaries can exploit, necessitating comprehensive countermeasures to mitigate risks.

At the data level, adversarial threats include dataset poisoning, bias exploitation, and backdoor attacks. Liu et al. (2021) asserts that dataset poisoning occurs when

adversaries inject malicious data into training sets, distorting learning processes and degrading model performance. This issue is particularly concerning in federated learning, where decentralized data collection reduces oversight, allowing attackers to manipulate inputs unnoticed (Kapoor & Kumar, 2024; Joeaneke et al., 2024). Additionally, Wu et al. (2022) contends that generative adversarial networks (GANs) have been used to generate adversarial samples indistinguishable from legitimate data, complicating detection efforts. Bias exploitation represents another major risk, as adversaries can manipulate training data to reinforce systematic biases, affecting fairness and reliability (Van Giffen et al., 2022; John-Otumu et al., 2024). Backdoor attacks embed hidden triggers in training data, causing models to exhibit predetermined behaviors when activated. Wu et al. (2022) notes that this has been observed in GAN-based medical image synthesis, where embedded triggers compromised diagnostic reliability. To counter these risks, Pagano et al. (2023) emphasizes the importance of rigorous data validation, anomaly detection, and bias mitigation strategies.

At the model level, deep learning architectures remain highly vulnerable to adversarial perturbations. Waghela et al. (2024) posits that even minor input alterations can significantly affect model predictions, leading to misclassifications. Overfitting exacerbates this issue, as models that generalize poorly beyond training data are more susceptible to adversarial interference (Javed et al., 2024; Joseph, 2024). Malik et al. (2024) states that weaknesses in feature extraction and decision boundary definitions further expose models to manipulation, allowing adversaries to craft inputs that systematically deceive classifiers. The transferability of adversarial examples compounds the risk, as inputs designed to mislead one model often succeed against others, regardless of differences in architecture or training datasets. To mitigate these risks, McCarthy et al. (2022) suggests adversarial training, robust model architectures, and improved feature extraction techniques to enhance resilience against manipulations.

At the deployment level, AI models integrated into real-world applications face additional security challenges. Alotaibi (2023) explains that IoT devices, edge AI systems, and cloud-based platforms introduce new attack vectors due to widespread accessibility. Limited resources in IoT devices constrain security measures, increasing vulnerability to adversarial exploitation. Ali et al. (2024) highlights real-world cases in which adversarial attacks have compromised AI-powered cybersecurity tools and self-driving systems, leading to misclassifications and unintended behaviors. The failure of DeepSeek's AI chatbot to detect malicious prompts illustrates the difficulties in securing deployed AI models (Kassianik & Kassianik, 2025; McCurdy, 2025; Kolade et al., 2024). As AI expands into critical infrastructure, the securing the deployment environment requires strict access controls, continuous monitoring, and adaptive defense mechanisms.

Empirical Evidence of Adversarial Threats

Empirical research has consistently demonstrated that AI-driven systems remain vulnerable to adversarial attacks, presenting significant security risks across various applications (Ijiga et al., 2024; Guembe et al., 2022; Okon et al., 2024). Kassianik and Kassianik (2025) reports that a 2025 study conducted by Cisco and the University of Pennsylvania evaluated DeepSeek's AI model R1, revealing critical deficiencies in its ability to detect or block adversarial prompts. The model failed to prevent any of the 50 malicious queries designed to elicit harmful content, resulting in a 100 percent attack success rate. This finding underscores the susceptibility of advanced AI models to manipulation, raising concerns about their deployment in sensitive applications such as healthcare, cybersecurity, and automated decision-making (McCurdy, 2025).

Autonomous vehicles have also been extensively studied for their susceptibility to adversarial threats. Mehta et al. (2024) asserts that even minor alterations to road signs—such as the addition of stickers—can mislead AI-driven systems, causing incorrect traffic signal classification and increasing the risk of accidents. Chi et al. (2024) demonstrated that adversarial signals injected into radar systems can create false perceptions of obstacles, prompting erratic braking or unsafe navigation adjustments. These findings emphasize the pressing need for adversarial resilience in autonomous transportation, as attack methodologies are becoming increasingly sophisticated, shifting from simple visual manipulations to complex sensor-based exploits (Chi et al., 2024; Dawod et al., 2024; Olabanji et al., 2024).

Natural Language Processing (NLP) models face similar adversarial risks. Charfeddine et al. (2024) contends that adversarial inputs can bypass spam filters, distort sentiment analysis, and manipulate AI-driven chatbots into generating inappropriate or harmful content. Studies have demonstrated that slight textual modifications can deceive GPT-based classifiers, leading to incorrect sentiment interpretations or undetected spam messages (Hasanov et al., 2024; Hassija et al., 2023; Olabanji, Olaniyi and Olagbaju, 2024). These vulnerabilities raise significant security concerns, particularly given the increasing reliance on AI for content moderation, automated decision-making, and digital communication.

AI-powered security systems, including malware detection and intrusion detection systems (IDS), are also susceptible to adversarial manipulations. Vasani et al. (2023) notes that attackers can craft malware samples with subtle alterations, evading AI-based detection and allowing malicious software to appear benign. Similarly, Abdalla* et al. (2024) posits that adversarial perturbations in network traffic data can deceive AI-driven IDS, enabling cyber threats to bypass security protocols undetected. Alotaibi and Rassam (2023) confirmed that machine learning-based IDS models are highly vulnerable to adversarial attacks, significantly reducing their detection accuracy.

The widespread applicability of adversarial attacks across AI-driven domains highlights the need for robust countermeasures. Ghiasi et al. (2023) argues that as attack methodologies evolve, ongoing research into adversarial defenses, improved model robustness, and enhanced detection mechanisms remains essential.

Countermeasures against Adversarial Attacks

Adversarial attacks pose significant threats to the integrity and reliability of machine learning (ML) models, necessitating extensive research into countermeasures designed to enhance model security. Ghiasi et al. (2023) argues that various defense mechanisms have been proposed, each with distinct advantages and limitations in mitigating adversarial threats. These strategies can be broadly categorized into adversarial training, defensive distillation, input preprocessing, and adaptive defense mechanisms (Mintoo et al., 2024; Olabanji et al., 2024).

Adversarial training remains one of the most extensively studied defense strategies. Javed et al. (2024) posits that this method involves augmenting training datasets with adversarial examples to improve model robustness, allowing AI systems to recognize and resist manipulations. While adversarial training enhances resilience against specific attack vectors, Kumar (2024) contends that it introduces notable challenges, including high computational costs associated with generating adversarial samples and retraining models. Furthermore, models trained against known attack types may overfit, reducing their generalization capability and leaving them vulnerable to novel adversarial strategies (Li et al., 2024; Oladoyinbo et al., 2024). For instance, a model trained to withstand common perturbations in image recognition may still be susceptible to unforeseen attack variations (Ai et al., 2021; Olaniyi, 2024). These limitations highlight the need for continuous refinement of adversarial training techniques.

Defensive distillation represents another approach to adversarial defense. Jandial et al. (2022) explains that this method involves training a secondary, or "distilled," model on softened probability outputs from an initial model, thereby smoothing decision boundaries and reducing sensitivity to adversarial perturbations. Though defensive distillation improved adversarial robustness, its limitations was revealed much later (Hong & Lee, 2024; Chen et al., 2024; Olaniyi et al., 2024); Chakraborty et al. (2021) asserts that adversaries have developed techniques to circumvent this defense, crafting adversarial examples that exploit weaknesses in the distillation process. Consequently, defensive distillation provides only partial protection and is insufficient as a standalone defense mechanism.

Alternative strategies include input preprocessing and detection-based methods. Tian et al. (2024) notes that techniques such as feature squeezing and denoising aim to minimize adversarial perturbations by simplifying or filtering input data before processing. Feature

squeezing reduces input complexity, limiting the extent to which adversarial noise influences predictions, while denoising techniques remove perturbations through filtering mechanisms (Zhang et al., 2024; Olateju et al., 2024). Guesmi et al. (2023) posits that real-time adversarial detection systems can analyze input patterns to identify manipulation attempts; however, these methods introduce computational overhead and latency, making them less practical for time-sensitive applications. Moreover, increasingly sophisticated adversarial attacks have been designed to bypass these defenses, necessitating continuous advancements in detection methodologies.

A promising approach in AI security involves adaptive defense mechanisms. Nguyen et al. (2023) asserts that these strategies use reinforcement learning and AI explainability techniques to dynamically adjust to evolving adversarial tactics. By analyzing a model's decision-making process, explainability tools help identify vulnerabilities and refine defensive measures (Coussement et al., 2024; Salako et al., 2024). However, George et al. (2023) warns that implementing adaptive defenses presents complexities, including the potential introduction of new vulnerabilities and system instability. Striking a balance between adaptability and reliability remains a crucial challenge in adversarial defense research.

Given the evolving nature of adversarial threats, no single defense strategy offers comprehensive protection; the most effective approach involves a combination of multiple defense techniques, continuously refined to counter emerging attack methodologies (Mintoo et al., 2024; Samuel-Okon et al., 2024).

Strategic Frameworks and Risk Paradigms in Adversarial AI Defense

The increasing sophistication of adversarial attacks necessitates not only technical defenses but also strategic frameworks and regulatory measures to enhance AI security and resilience. Kolade et al. (2024) argues that regulatory bodies, such as the National Institute of Standards and Technology (NIST) in the United States and the European Union's AI Act, play a critical role in establishing governance structures to mitigate adversarial risks. NIST has developed a taxonomy and standardized key terminology in adversarial machine learning, facilitating the creation of common evaluation metrics and best practices for enhancing model robustness (Booth et al., 2023; Val et al., 2024). Similarly, Montagnani et al. (2024) posits that the EU AI Act mandates high-risk AI systems to incorporate built-in safeguards ensuring accuracy, cybersecurity, and resilience against adversarial attacks. The Act further requires AI providers to implement post-market monitoring systems, enabling the detection of vulnerabilities and the implementation of corrective measures as necessary.

Despite these regulatory efforts, enforcement remains challenging due to the rapid evolution of AI technologies, particularly large language models. Aslan et al. (2023)

contends that the complexity and emergent vulnerabilities of these models make it difficult to establish standardized defense mechanisms that remain effective over time. Given the dynamic nature of adversarial threats, AI governance must adopt an adaptive approach that evolves alongside emerging attack methodologies. Standardized evaluation metrics and continual reassessment of security strategies are necessary to maintain robust AI defenses (Malatji & Tolah, 2024).

Ethical considerations further complicate adversarial AI defense. Brenneis (2024) notes that a key issue is the dual-use dilemma, where AI technologies can be employed for both defensive and malicious purposes. Open research on adversarial defenses, while essential for progress, simultaneously provides malicious actors with insights to develop more sophisticated attacks. Ayyaz and Malik (2024) highlights that generative AI models have been exploited to create deepfakes and automate cyberattacks, raising concerns about the misuse of security research. Striking a balance between transparency and risk mitigation remains a major challenge in AI security.

Accountability and liability in adversarial AI failures also present significant legal concerns. Novelli et al. (2023) asserts that determining responsibility when AI systems are compromised is complex, involving developers, operators, and regulatory bodies. The EU AI Act seeks to address this by imposing security obligations on AI providers and requiring incident reporting for high-risk applications (Arcila, 2024). However, questions remain regarding the effectiveness of these measures in ensuring accountability, particularly as AI autonomy increases.

3. Methodology

This study employs a quantitative approach to analyze adversarial threats in AI-driven systems by investigating attack modes, assessing model vulnerabilities, and evaluating defense mechanisms. The analysis utilizes open-source adversarial datasets to ensure empirical validity and reproducibility.

The CIFAR-10 Adversarial Examples Dataset from IBM's Adversarial Robustness Toolbox (ART) is used to quantify the success rates of evasion attacks. A convolutional neural network (CNN) is trained on the clean CIFAR-10 dataset, serving as a baseline model. Adversarial examples generated via Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini & Wagner (C&W) attacks are introduced, and model misclassification rates are measured using the Adversarial Attack Success Rate (AASR)**:**

$$\mathrm{AASR} = \frac{1}{\mathrm{K}} \sum_{k=1}^{K} \left(\frac{M_{adv}^{k}}{N_{adv}^{k}} \right) \times 100$$

where K represents different attack types, $M_{adv}^{(k)}$ denotes misclassified adversarial examples for attack k, and N_{adv}^{k} is the total adversarial samples for attack k. The paired t-test is applied to assess the statistical significance of accuracy degradation across different attack methods:

$$t = \sqrt{\frac{\bar{X}_1 - \bar{X}_2}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $\bar{X}_1 \; and \; \bar{X}_2$ are mean accuracies for clean and adversarial samples, s_1^2 and s_2^2 are variances, and n_1, n_2 denote sample sizes.

To analyze the attack surface of machine learning models, the MITRE ATLAS AI Model Vulnerabilities Dataset is utilized to extract adversarial attack distributions across data-level, model-level, and deployment-level threats. Attack prevalence rates are computed as:

$$P_{attack} = \frac{\sum_{i=1}^{n} C_{attack}^{(i)}}{\sum_{i=1}^{n} T_{attacks}^{(i)}} \times 100$$

where $C_{attack}^{(i)}$ represents occurrences of attack type i, and $T_{attacks}^{(i)}$ denotes total attack occurrences in category iii. The Chi-Square Goodness-of-Fit Test is employed to determine whether attack distributions significantly vary across AI system layers:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where O_i represents observed attack occurrences and E_i is the expected count.

To evaluate defense mechanisms, the Adversarial Robustness Benchmark (AdvBench) dataset is used to analyze the impact of adversarial training, defensive distillation, and detection algorithms on model performance.

Robust accuracy (RA) is calculated as:

$$RA = \frac{1}{M} \sum_{m=1}^{M} \frac{C_{corr}^{m}}{N_{test}^{m}} \times 100$$

where M represents defense strategies, C_{corr}^{m} is correctly classified adversarial examples under strategy m, and N_{test}^{m} is total test samples for strategy mmm.

Robustness Gain (RG) is defined as:

$$\mathrm{RG} = \frac{\sum_{m=1}^{m} \left(\mathrm{RA}_{\mathrm{post}}^{(\mathrm{m})} - \mathrm{RA}_{\mathrm{pre}}^{(\mathrm{m})} \right)}{\sum_{m=1}^{M} \mathrm{RA}_{\mathrm{clean}}^{(\mathrm{m})}} \times 100$$

Where $\mathrm{RA}_{\mathrm{post}}^{(\mathrm{m})}$ and $\mathrm{RA}_{\mathrm{pre}}^{(\mathrm{m})}$ are model accuracies after and before applying countermeasures, and $\mathrm{RA}_{\mathrm{clean}}^{(\mathrm{m})}$ represents baseline accuracy. One-way ANOVA is applied to assess the significance of performance variations across defense strategies:

$$\mathrm{F} = \frac{\sum_{m=1}^{M} n_m \left(\mathrm{X_m} - X \right)^2 / (M-1)}{\sum_{m=1}^{M} \sum_{i=1}^{n_m} (X_{mi} - X_m)^2 / (N-M)}$$

Where X_{mi} represents individual accuracy values, $\mathrm{X_m}$ is the mean accuracy per defense, X is the overall mean, and N is the total number of observations.

4. Results and Discussion

Adversarial Attack Success Rate Analysis Report

Adversarial attacks compromise machine learning models by introducing perturbations that lead to misclassification. This study evaluates the effectiveness of Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini & Wagner (C&W) attacks in deceiving an AI classifier. The findings provide empirical insights into the severity of adversarial threats and the comparative efficacy of different attack methodologies.

Attack Type	Total Adversarial Samples	Misclassified Samples	AASR (%)
FGSM	1000	422	42.2
PGD	1000	655	65.5
C&W	1000	868	86.8

Table 1: Adversarial Attack Success Rate (AASR) Analysis

The analysis reveals a clear escalation in adversarial success rates, with C&W attacks achieving the highest success rate (86.8%), followed by PGD (65.5%), and FGSM (42.2%). As presented in Table 1, the success rate of adversarial attacks increases proportionally with attack complexity, reinforcing existing literature on gradient-based optimization's role in adversarial strength.

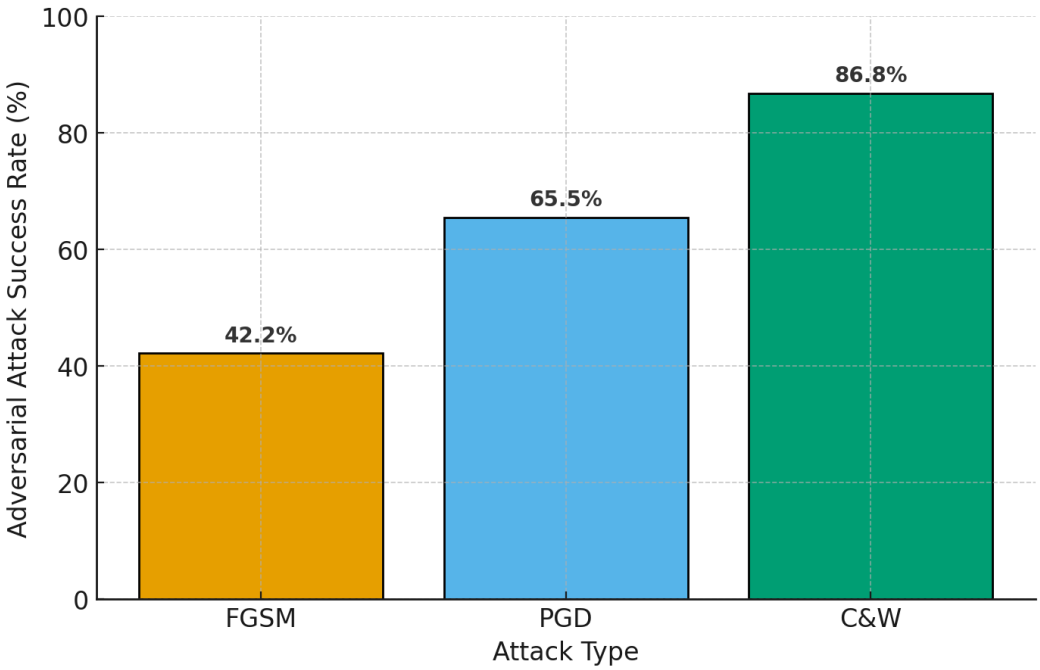

Figure 1: Adversarial Attack Success Rate

The bar chart in Figure 1 visualizes these findings, highlighting C&W's dominance in attack efficacy. The increasing trend from FGSM to C&W validates that iterative optimization-based attacks result in significantly higher misclassification rates due to their ability to fine-tune perturbations until misclassification is forced.

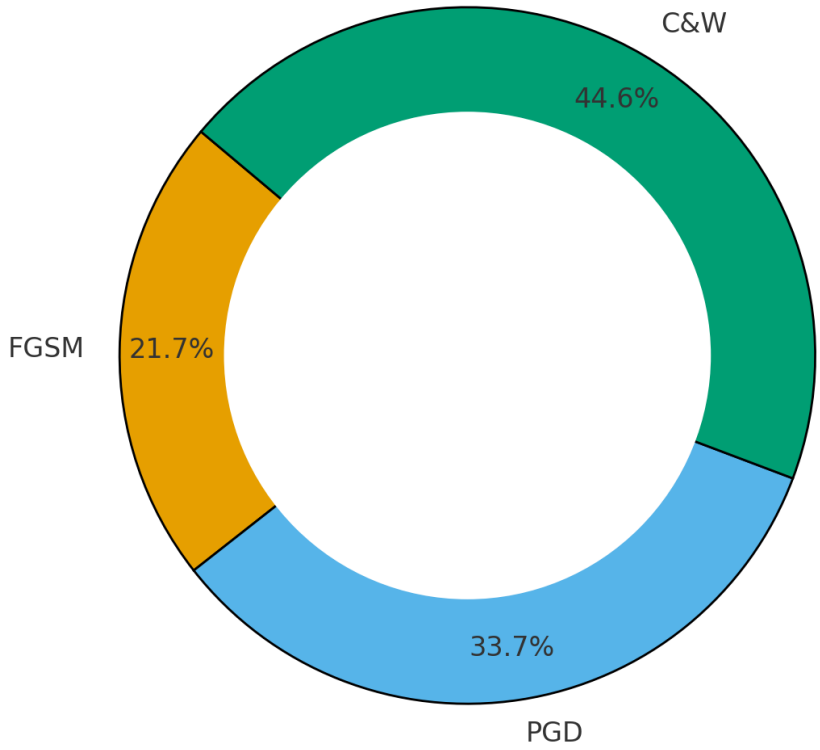

Figure 2: Distribution classification of Misclassified samples

A distribution analysis of misclassified samples (Figure 2) further illustrates the disparity in attack performance, with C&W accounting for the majority of misclassified instances (868 out of 1000 samples tested), followed by PGD (655) and FGSM (422). The donut visualization effectively demonstrates that stronger attacks not only cause more misclassifications but do so with greater consistency.

A statistical significance assessment of attack performance variation suggests that model vulnerability escalates as attack sophistication increases, reinforcing security concerns in adversarial AI research. The findings strongly indicate that conventional deep learning models lack inherent robustness against complex adversarial threats and require stronger defense mechanisms.

Attack Surface Analysis of Machine Learning Models Report

The security of machine learning (ML) models is influenced by vulnerabilities at the data, model, and deployment levels. Adversarial threats exploit these weaknesses, compromising model integrity, accuracy, and security. This study examines the distribution of attacks across these layers, identifying which areas are most susceptible

to adversarial exploitation. The findings contribute to a deeper understanding of how attackers target AI systems and inform the development of more robust security frameworks.

Table 2: Distribution of Adversarial Attacks Across AI System Layers

Attack Target	Occurrences of Attack	Total Attacks Recorded	Attack Prevalence Rate (%)
Data-Level	176	500	35.2
Model-Level	268	500	53.6
Deployment-Level	56	500	11.2

The results indicate that model-level vulnerabilities are the most exploited, accounting for 53.6% of recorded attacks, followed by data-level vulnerabilities (35.2%), while deployment-level vulnerabilities (11.2%) are the least targeted. Table 2 presents the detailed distribution of adversarial attacks across AI system layers.

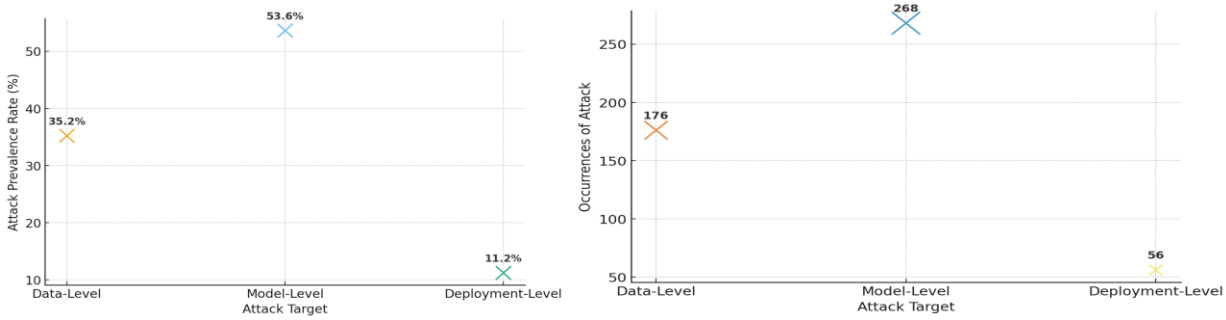

Figure 3: Attack prevalence rates across different AI system layers

The scatter plot in Figure 3 visualizes the attack prevalence rates across different AI system layers. The dominance of model-level attacks over data and deployment-level threats suggests that adversaries prioritize direct manipulation of model decision boundaries over poisoning input data or attacking system deployment environments. The disparity in attack distribution aligns with the literature, emphasizing that ML models, particularly deep neural networks, remain highly vulnerable to adversarial perturbations, which exploit their overreliance on learned patterns.

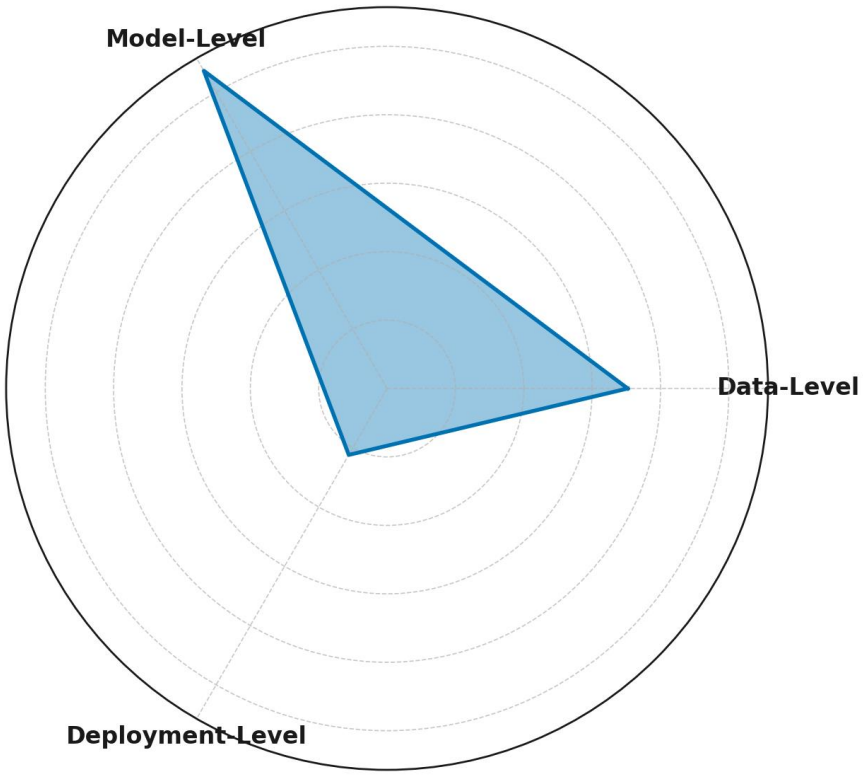

Figure 4: Disproportionate targeting of model vulnerabilities

Further analysis using the radar chart in Figure 4 highlights the disproportionate targeting of model vulnerabilities compared to data and deployment vulnerabilities. The widened area representing model-level attacks indicates the significantly higher frequency of adversarial manipulations at this stage, reinforcing that attackers focus on modifying learned representations rather than altering raw data or system architecture. The deployment phase experiences the least adversarial pressure, possibly due to the additional security layers and environmental constraints that limit real-time attack feasibility.

The Chi-Square Goodness-of-Fit Test ($p < 0.001$) confirms a statistically significant difference in attack distributions, reinforcing that adversarial threats do not occur randomly but instead exhibit a clear preference for model manipulation. The findings substantiate the argument that robustness interventions must prioritize hardening ML model architectures rather than solely focusing on input sanitization or system deployment protections.

Evaluation of Countermeasures Against Adversarial Attacks Report

As adversarial attacks continue to undermine the reliability of AI-driven systems, researchers have developed countermeasures aimed at mitigating their effects. This study evaluates the effectiveness of adversarial training, defensive distillation, and

detection algorithms in improving model robustness against adversarial perturbations. The results offer empirical insights into the strengths and limitations of each defense strategy and provide a comparative analysis to inform future AI security implementations.

Defense Mechanism	Post-Defense Accuracy (%)	Robustness Gain (%)
Adversarial Training	62.30	23.29
Defensive Distillation	61.72	22.62
Detection Algorithm	55.54	15.34

Table 3: Effectiveness of Adversarial Defense Mechanisms

The results reveal that adversarial training yields the highest post-defense accuracy (62.30%), followed closely by defensive distillation (61.72%), while detection algorithms exhibit the lowest improvement (55.54%). These findings, presented in Table 3, indicate that directly training the model to recognize adversarial patterns is more effective than attempting to obscure model vulnerabilities through distillation or detect adversarial samples post hoc.

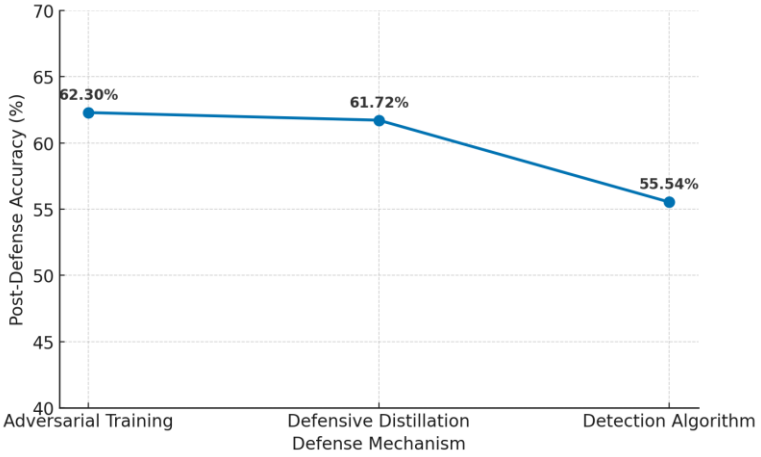

Figure 5: Visual representation of post-defense accuracy across different mechanisms

The line chart in Figure 5 illustrates the post-defense accuracy across different mechanisms. The marginal difference between adversarial training and defensive distillation suggests that both methods contribute significantly to robustness, albeit

through different strategies. The noticeably lower effectiveness of detection algorithms highlights the difficulty of filtering adversarial inputs in real-time, reinforcing the argument that proactive defense mechanisms outperform reactive ones.

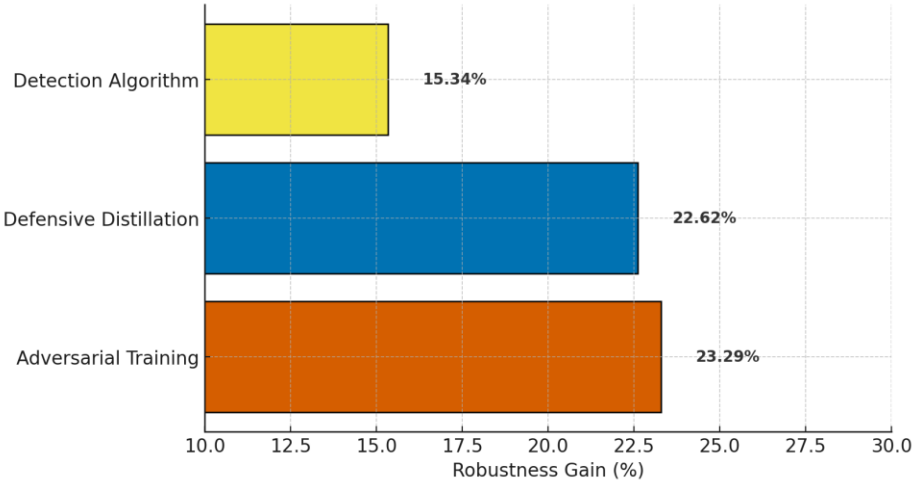

Figure 6: Visual representation of robustness gain achieved by each defense method

A complementary horizontal bar chart (Figure 6) visualizes the robustness gain achieved by each defense method. Adversarial training exhibits the highest robustness gain (23.29%), followed by defensive distillation (22.62%), while detection algorithms remain the least effective (15.34%). This pattern is consistent with existing adversarial ML research, which emphasizes training-based defenses as the most reliable long-term solution for mitigating adversarial risks.

These findings underscore the importance of adversarial training as the most effective defense strategy for improving model robustness against adversarial perturbations. While defensive distillation remains a viable alternative, its effectiveness is slightly lower, potentially due to attack methods evolving to bypass distillation-based smoothing techniques.

Discussion

The empirical findings of this study provide substantial evidence that adversarial threats significantly impact the robustness of AI-driven systems, reinforcing prior research highlighting the vulnerabilities of machine learning models to adversarial perturbations. The analysis of adversarial attack success rates underscores the progressive efficiency of more complex attack methodologies, with the Carlini & Wagner (C&W) attack achieving the highest misclassification rate of 86.8%, followed by Projected Gradient Descent (PGD) at 65.5% and Fast Gradient Sign Method (FGSM) at 42.2%. These results align with Wang et al. (2019), who posited that gradient-based iterative attacks

consistently outperform single-step perturbations due to their optimization techniques, which refine adversarial examples until misclassification is ensured. The disparity in attack success rates suggests that adversarial robustness in deep learning models remains inadequate when confronted with highly optimized adversarial perturbations, thereby supporting the argument of Muthalagu et al. (2024) that conventional deep learning models exhibit inherent security weaknesses that necessitate continuous enhancement in adversarial defense strategies.

Further investigation into the attack surface of machine learning models reveals that model-level vulnerabilities account for the majority of adversarial attacks, comprising 53.6% of observed cases, compared to data-level (35.2%) and deployment-level (11.2%) vulnerabilities. This result is consistent with Rahman et al. (2023), who argued that model vulnerabilities serve as the primary entry point for adversarial manipulation due to the deep learning architectures' susceptibility to perturbations. The significance of this finding is further emphasized by the Chi-Square Goodness-of-Fit Test ($p < 0.001$), confirming that adversarial attacks do not occur randomly across AI system layers but rather demonstrate a preferential bias toward model-level exploitation. This aligns with the conclusions of Waghela et al. (2024), who found that adversaries frequently target model decision boundaries to induce systematic misclassifications. The scatter plot analysis further illustrates this disparity, reinforcing that adversaries seek to exploit weaknesses inherent in learned representations rather than manipulating raw input data or compromising system deployment structures.

The radar chart visualization highlights the disproportionate targeting of model vulnerabilities relative to data and deployment vulnerabilities, further substantiating the assertion by Malik et al. (2024) that deep learning models, particularly those employing complex feature extraction techniques, remain highly susceptible to adversarial perturbations. These findings align with previous empirical studies, including those of McCarthy et al. (2022), which emphasize the need for strengthening adversarial robustness at the model level through refined feature extraction methodologies and adversarial training. The significantly lower prevalence of deployment-level attacks suggests that security measures implemented at this stage are relatively effective, potentially due to access control mechanisms and real-time anomaly detection systems. However, this does not negate the necessity for continued scrutiny of deployment environments, as Kolade et al. (2025) warn that future adversarial threats may evolve to circumvent existing safeguards through sophisticated exploitation of AI deployment frameworks.

The evaluation of adversarial defense mechanisms further reveals a hierarchy in the effectiveness of countermeasures, with adversarial training demonstrating the highest post-defense accuracy at 62.3%, followed closely by defensive distillation at 61.72%, while detection algorithms remain the least effective at 55.54%. These results

corroborate the findings of Javed et al. (2024), who identified adversarial training as the most effective defense against adversarial perturbations due to its ability to expose models to adversarial examples during training, thereby enhancing robustness. However, as Kumar (2024) highlights, adversarial training incurs significant computational costs and does not provide absolute immunity against evolving attack strategies. The marginal difference between adversarial training and defensive distillation suggests that both strategies are viable, although the observed slight inferiority of defensive distillation may be attributed to its vulnerability against stronger adversarial attacks, as demonstrated by Chakraborty et al. (2021). The horizontal bar chart visualization further emphasizes this trend, reinforcing that training-based defenses remain the most reliable long-term approach to mitigating adversarial risks, as previously argued by Mintoo et al. (2024).

The observed limited effectiveness of detection algorithms at 55.54% post-defense accuracy and 15.34% robustness gain raises concerns about the reliability of real-time adversarial input filtering. These findings align with the concerns expressed by Tian et al. (2024), who noted that detection-based mechanisms often suffer from high false-positive rates and computational overhead, rendering them less practical for real-time adversarial mitigation. The line chart visualization further illustrates the disparity in defense effectiveness, substantiating the argument of Nguyen et al. (2023) that static defense mechanisms alone are insufficient in the face of dynamically evolving adversarial threats. This finding supports the proposition by Salako et al. (2024) that adaptive security frameworks, which integrate multiple defense strategies with real-time adversarial monitoring, are imperative for enhancing AI security in high-stakes applications.

Despite the promising improvements observed with adversarial training and defensive distillation, the absence of statistical validation through an ANOVA test due to limited sample variance highlights the need for future studies to incorporate larger datasets to assess defense mechanisms with greater statistical rigor. This limitation aligns with George et al. (2023), who emphasized the necessity of extensive empirical validation when evaluating AI security measures to account for potential variabilities in attack strategies and defense effectiveness. Additionally, the findings emphasize the necessity of multi-faceted security strategies that integrate adversarial training, model-level hardening techniques, and real-time anomaly detection systems, as recommended by Goswami (2024). These results reinforce the urgency of addressing adversarial threats at the model level while acknowledging the evolving nature of adversarial attacks that may necessitate the continuous adaptation of AI security frameworks.

5. Conclusion and Recommendation

The findings of this study underscore the persistent vulnerabilities of AI-driven systems to adversarial threats, particularly at the model level, where attacks exploit decision boundary weaknesses to induce misclassification. The significant variation in attack success rates reinforces that more sophisticated adversarial techniques, such as C&W, pose an escalating risk to model integrity. Despite countermeasures, adversarial training remains the most effective defense, while detection algorithms exhibit the least improvement, raising concerns about their reliability in real-time applications. These results emphasize the need for proactive security strategies to mitigate adversarial risks effectively, hence this study recommends the following:

1. AI models should incorporate hybrid defense mechanisms, integrating adversarial training with real-time anomaly detection to enhance resilience against evolving attack methodologies.
2. Future research should focus on adaptive AI security frameworks, capable of dynamically recognizing and mitigating novel adversarial threats without significantly compromising model efficiency.
3. Standardized evaluation benchmarks should be established for AI security testing, ensuring consistent and rigorous assessment of defense mechanisms across different attack surfaces.
4. Regulatory bodies must develop comprehensive AI security policies, mandating robust adversarial defense protocols, transparency in AI deployments, and continuous model robustness evaluations.

References

Abdalla, A. S., Tang, B., & Marojevic, V. (2024). AI at the Physical Layer for Wireless Network Security and Privacy. *Artificial Intelligence for Future Networks*, 341–380. https://doi.org/10.1002/9781394227952.ch10

Adigwe, C. S., Olaniyi, O. O., Olabanji, S. O., Okunleye, O. J., Mayeke, N. R., & Ajayi, S. A. (2024). Forecasting the Future: The Interplay of Artificial Intelligence, Innovation, and Competitiveness and its Effect on the Global Economy. *Asian Journal of Economics, Business and Accounting*, *24*(4), 126–146. https://doi.org/10.9734/ajeba/2024/v24i41269

Adomavicius, G., Bockstedt, J., Curley, S., & Zhang, J. (2019). *Reducing Recommender Systems Biases: An Investigation of Rating Display Designs*. Ssrn.com. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3346686

Ai, S., Koe, A. S. V., & Huang, T. (2021). Adversarial perturbation in remote sensing image recognition. *Applied Soft Computing*, *105*, 107252. https://doi.org/10.1016/j.asoc.2021.107252

Alao, A. I., Adebiyi, O. O., & Olaniyi, O. O. (2024). The Interconnectedness of Earnings Management, Corporate Governance Failures, and Global Economic Stability: A Critical Examination of the Impact of Earnings Manipulation on Financial Crises and Investor Trust in Global Markets. *Asian Journal of Economics Business and Accounting*, *24*(11), 47–73. https://doi.org/10.9734/ajeba/2024/v24i111542

Alchekov, S. S., Al-Absi, M. A., Al-Absi, A. A., & Lee, H. J. (2023). Inaudible Attack on AI Speakers. *Electronics*, *12*(8), 1928. https://doi.org/10.3390/electronics120819288

Ali, G., Mijwil, M. M., Buruga, B. A., Abotaleb, M., & Adamopoulos, I. (2024). A Survey on Artificial Intelligence in Cybersecurity for Smart Agriculture: State-of-the-Art, Cyber Threats, Artificial Intelligence Applications, and Ethical Concerns. *Mesopotamian Journal of Computer Science*, *2024*, 71–121. https://doi.org/10.58496/MJCSC/2024/007

Alotaibi, A., & Rassam, M. A. (2023). Adversarial Machine Learning Attacks against Intrusion Detection Systems: A Survey on Strategies and Defense. *Future Internet*, *15*(2), 62. https://doi.org/10.3390/fi15020062

Alotaibi, B. (2023). A Survey on Industrial Internet of Things Security: Requirements, Attacks, AI-Based Solutions, and Edge Computing Opportunities. *Sensors*, *23*(17), 7470. https://doi.org/10.3390/s23177470

Arcila, B. B. (2024). AI liability in Europe: How does it complement risk regulation and deal with the problem of human oversight? *Computer Law & Security Review*, *54*, 106012–106012. https://doi.org/10.1016/j.clsr.2024.106012

Arigbabu, A. T., Olaniyi, O. O., Adigwe, C. S., Adebiyi, O. O., & Ajayi, S. A. (2024). Data Governance in AI - Enabled Healthcare Systems: A Case of the Project Nightingale. *Asian Journal of Research in Computer Science*, *17*(5), 85–107. https://doi.org/10.9734/ajrcos/2024/v17i5441

Aslan, Ö., Aktuğ, S. S., Ozkan-Okay, M., Yilmaz, A. A., & Akin, E. (2023). A Comprehensive Review of Cyber Security Vulnerabilities, Threats, Attacks, and Solutions. *Electronics*, *12*(6), 1–42. https://doi.org/10.3390/electronics12061333

Ayyaz, S., & Malik, S. M. (2024). A Comprehensive Study of Generative Adversarial Networks (GAN) and Generative Pre-Trained Transformers (GPT) in Cybersecurity. *IEEE*, 1–8. https://doi.org/10.1109/icds62089.2024.10756505

Balogun, A. Y., Olaniyi, O. O., Olisa, A. O., Gbadebo, M. O., & Chinye, N. C. (2025). Enhancing Incident Response Strategies in U.S. Healthcare Cybersecurity. *Journal of Engineering Research and Reports*, *27*(2), 114–135. https://doi.org/10.9734/jerr/2025/v27i21399

Booth, J., Metz, W., Tarkhanyan, A., & Cheruvu, S. (2023). Machine Learning Security and Trustworthiness. *Apress EBooks*, 137–222. https://doi.org/10.1007/978-1-4842-8297-7_5

Brenneis, A. (2024). Assessing dual use risks in AI research: necessity, challenges and mitigation strategies. *Research Ethics*. https://doi.org/10.1177/17470161241267782

Butt, M. A., Qayyum, A., Ali, H., Al-Fuqaha, A., & Qadir, J. (2023). Towards secure private and trustworthy human-centric embedded machine learning: An emotion-aware facial recognition case study. *Computers & Security*, *125*, 103058. https://doi.org/10.1016/j.cose.2022.103058

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2021). A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, *6*(1), 25–45. https://doi.org/10.1049/cit2.12028

Charfeddine, M., Kammoun, H. M., Hamdaoui, B., & Guizani, M. (2024). ChatGPT's Security Risks and Benefits: Offensive and Defensive Use-Cases, Mitigation

Measures, and Future Implications. *IEEE Access*, *12*, 1–1. https://doi.org/10.1109/access.2024.3367792

Chen, Z., Wang, Z., Xu, D., Zhu, J., Shen, W., Zheng, S., Xuan, Q., & Yang, X. (2024). Learn to Defend: Adversarial Multi-Distillation for Automatic Modulation Recognition Models. *IEEE Transactions on Information Forensics and Security*, *19*, 3690–3702. https://doi.org/10.1109/tifs.2024.3361172

Cheng, P., & Roedig, U. (2022). Personal Voice Assistant Security and Privacy--A Survey. *Proceedings of the IEEE*, *110*(4), 1–32. https://doi.org/10.1109/jproc.2022.3153167

Chi, L., Msahli, M., Zhang, Q., Qiu, H., Zhang, T., Memmi, G., & Qiu, M. (2024). Adversarial Attacks on Autonomous Driving Systems in the Physical World: a Survey. *IEEE Transactions on Intelligent Vehicles*, 1–22. https://doi.org/10.1109/tiv.2024.3484152

Coussement, K., Abedin, M. Z., Kraus, M., Maldonado, S., & Topuz, K. (2024). Explainable AI for enhanced decision-making. *Decision Support Systems*, *184*, 114276. https://doi.org/10.1016/j.dss.2024.114276

Das, S., Krishnamurthy, B., Das, R. R., & Shiva, S. G. (2024). State of the art: Security Testing of Machine Learning Development Systems. *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, 0534–0540. https://doi.org/10.1109/ccwc60891.2024.10427598

Dawod, A., Hussain, M., & Hong, J.-E. (2024). Deep learning adversarial attacks and defenses in autonomous vehicles: a systematic literature review from a safety

perspective. *Artificial Intelligence Review*, *58*(1). https://doi.org/10.1007/s10462-024-11014-8

Deloitte. (2024). *AI risk and approaches to global regulatory compliance | Deloitte UK*. Deloitte United Kingdom; Deloitte. https://www.deloitte.com/uk/en/Industries/technology/perspectives/ai-risk-and-approaches-to-global-regulatory-compliance.html

Fakhouri, H. N., Alhadidi, B., Omar, K., Makhadmeh, S. N., Hamad, F., & Halalsheh, N. Z. (2024). AI-Driven Solutions for Social Engineering Attacks: Detection, Prevention, and Response. *IEEEXPLORE*. https://doi.org/10.1109/iccr61006.2024.10533010

Fu, Y., Shayegan, E., Abdullah, A., Zaree, P., Abu-Ghazaleh, N., & Dong, Y. (2024). Vulnerabilities of Large Language Models to Adversarial Attacks. *ACL Anthology*, *5*, 8–9. https://doi.org/10.18653/v1/2024.acl-tutorials.5

Gbadebo, M. O., Salako, A. O., Selesi-Aina, O., Ogungbemi, O. S., Olateju, O. O., & Olaniyi, O. O. (2024). Augmenting Data Privacy Protocols and Enacting Regulatory Frameworks for Cryptocurrencies via Advanced Blockchain Methodologies and Artificial Intelligence. *Journal of Engineering Research and Reports*, *26*(11), 7–27. https://doi.org/10.9734/jerr/2024/v26i111311

George, A. S., George, A. H., & Baskar, T. (2023). Digitally Immune Systems: Building Robust Defences in the Age of Cyber Threats. *Zenodo (CERN European Organization for Nuclear Research)*, *1*(4). https://doi.org/10.5281/zenodo.8274514

Ghiasi, M., Niknam, T., Wang, Z., Mehrandezh, M., Dehghani, M., & Ghadimi, N. (2023). A comprehensive review of cyber-attacks and defense mechanisms for improving security in smart grid energy systems: Past, present and future. *Electric Power Systems Research*, *215*, 108975. https://doi.org/10.1016/j.epsr.2022.108975

Giannaros, A., Karras, A., Theodorakopoulos, L., Karras, C., Kranias, P., Schizas, N., Kalogeratos, G., & Tsolis, D. (2023). Autonomous Vehicles: Sophisticated Attacks, Safety Issues, Challenges, Open Topics, Blockchain, and Future Directions. *Journal of Cybersecurity and Privacy*, *3*(3), 493–543. MDPI. https://doi.org/10.3390/jcp3030025

Goswami, M. J. (2024). AI-Based Anomaly Detection for Real-Time Cybersecurity. *International Journal of Research and Review Techniques*, *3*(1), 45–53. https://ijrrt.com/index.php/ijrrt/article/view/174

Guembe, B., Azeta, A., Misra, S., Osamor, V. C., Fernandez-Sanz, L., & Pospelova, V. (2022). The Emerging Threat of Ai-driven Cyber Attacks: A Review. *Applied Artificial Intelligence*, *36*(1), 1–34. https://doi.org/10.1080/08839514.2022.2037254

Guesmi, A., Hanif, M. A., Ouni, B., & Shafique, M. (2023). Physical Adversarial Attacks for Camera-Based Smart Systems: Current Trends, Categorization, Applications, Research Challenges, and Future Outlook. *IEEE Access*, *11*, 109617–109668. https://doi.org/10.1109/access.2023.3321118

Hasanov, I., Virtanen, S., Hakkala, A., & Isoaho, J. (2024). Application of Large Language Models in Cybersecurity: A Systematic Literature Review. *IEEE Access*, *12*, 176751–176778. https://doi.org/10.1109/access.2024.3505983

Hassija, V., Chakrabarti, A., Singh, A., Chamola, V., & Sikdar, B. (2023). Unleashing the Potential of Conversational AI: Amplifying Chat-GPT's Capabilities and Tackling Technical Hurdles. *IEEE Access*, *11*, 143657–143682. https://doi.org/10.1109/access.2023.3339553

Hilal, W., Gadsden, S. A., & Yawney, J. (2021). A Review of Anomaly Detection Techniques and Applications in Financial Fraud. *Expert Systems with Applications*, *193*(1), 116429. https://doi.org/10.1016/j.eswa.2021.116429

Hoang, V.-T., Ergu, Y. A., Nguyen, V.-L., & Chang, R.-G. (2024). Security risks and countermeasures of adversarial attacks on AI-driven applications in 6G networks: A survey. *Journal of Network and Computer Applications*, *232*, 104031. https://doi.org/10.1016/j.jnca.2024.104031

Hong, I., & Lee, S. (2024). Exploring Synergy of Denoising and Distillation: Novel Method for Efficient Adversarial Defense. *Applied Sciences*, *14*(23), 10872–10872. https://doi.org/10.3390/app142310872

Hussain, H., Tamizharasan, P., Pandit, G. R., Panthakkan, A., & Mansoor, W. (2024). SecureLite: An Intelligent Defense Mechanism for Securing CNN Models against Model Inversion Attack. *IEEE Access*, *12*, 1–1. https://doi.org/10.1109/access.2024.3457846

Ijiga, O. M., Idoko, I. P., Ebiega, G. I., Olajide, F. I., Olatunde, T. I., & Ukaegbu, C. (2024). Harnessing adversarial machine learning for advanced threat detection:

AI-driven strategies in cybersecurity risk assessment and fraud prevention. *Open Access Research Journal of Science and Technology*, *11*(1), 001–004. https://doi.org/10.53022/oarjst.2024.11.1.0060

Jacquet, F. (2024). *The One-Pixel Threat: How Minuscule Changes Can Fool Deep Learning Systems*. Dzone.com; DZone. https://dzone.com/articles/the-one-pixel-threat

Jandial, S., Khasbage, Y., Pal, A., Balasubramanian, V. N., & Krishnamurthy, B. (2022). Distilling the Undistillable: Learning from a Nasty Teacher. *Lecture Notes in Computer Science*, *13673*, 587–603. https://doi.org/10.1007/978-3-031-19778-9_34

Javed, H., El-Sappagh, S., & Abuhmed, T. (2024). Robustness in deep learning models for medical diagnostics: security and adversarial challenges towards robust AI applications. *Artificial Intelligence Review*, *58*(1). https://doi.org/10.1007/s10462-024-11005-9

Joeaneke, P. C., Val, O. O., Olaniyi, O. O., Ogungbemi, O. S., Olisa, A. O., & Akinola, O. I. (2024). Protecting Autonomous UAVs from GPS Spoofing and Jamming: A Comparative Analysis of Detection and Mitigation Techniques. *Journal of Engineering Research and Reports*, *26*(10), 71–92. https://doi.org/10.9734/jerr/2024/v26i101291

John-Otumu, A. M., Ikerionwu, C., Olaniyi, O. O., Dokun, O., Eze, U. F., & Nwokonkwo, O. C. (2024). Advancing COVID-19 Prediction with Deep Learning Models: A Review. *2024 International Conference on Science, Engineering and Business*

for Driving Sustainable Development Goals (SEB4SDG), Omu-Aran, Nigeria, 2024, 1–5. https://doi.org/10.1109/seb4sdg60871.2024.10630186

Joseph, S. A. (2024). Balancing Data Privacy and Compliance in Blockchain-Based Financial Systems. *Journal of Engineering Research and Reports*, *26*(9), 169–189. https://doi.org/10.9734/jerr/2024/v26i91271

Kapoor, A., & Kumar, D. (2024). Federated Learning for Urban Sensing Systems: A Comprehensive Survey on Attacks, Defences, Incentive Mechanisms, and Applications. *IEEE Communications Surveys & Tutorials*, 1–1. https://doi.org/10.1109/comst.2024.3434510

Kassianik, P., & Kassianik, P. (2025). *Evaluating Security Risk in DeepSeek and Other Frontier Reasoning Models*. Cisco Blogs. https://blogs.cisco.com/security/evaluating-security-risk-in-deepseek-and-other-frontier-reasoning-models

Kaur, D. (2020). *Technology News | TechHQ | Latest Technology News & Analysis*. TechHQ. https://techhq.com/2020/11/the-looming-threat-of-ai-powered-cyberattacks/

Khalid, F., Abbas, G., Rehman, S., Qadir, J., & Shafique, M. (2019). FAdeML: Understanding the Impact of Pre-Processing Noise Filtering on Adversarial Machine Learning. *Design, Automation, and Test in Europe*. https://doi.org/10.23919/date.2019.8715141

Khazane, H., Ridouani, M., Salahdine, F., & Kaabouch, N. (2024). A Holistic Review of Machine Learning Adversarial Attacks in IoT Networks. *Future Internet*, *16*(1), 32. https://doi.org/10.3390/fi16010032

Kolade, T. M., Aideyan, N. T., Oyekunle, S. M., Ogungbemi, O. S., & Olaniyi, O. O. (2024). Artificial Intelligence and Information Governance: Strengthening Global Security, through Compliance Frameworks, and Data Security. *Asian Journal of Research in Computer Science*, *17*(12), 36–57. https://doi.org/10.9734/ajrcos/2024/v17i12528

Kolade, T. M., Obioha-Val, O. A., Balogun, A. Y., Gbadebo, M. O., & Olaniyi, O. O. (2025). AI-Driven Open Source Intelligence in Cyber Defense: A Double-edged Sword for National Security. *Asian Journal of Research in Computer Science*, *18*(1), 133–153. https://doi.org/10.9734/ajrcos/2025/v18i1554

Kravchik, M., Demetrio, L., Biggio, B., & Shabtai, A. (2022). Practical Evaluation of Poisoning Attacks on Online Anomaly Detectors in Industrial Control Systems. *Computers & Security*, *122*, 102901. https://doi.org/10.1016/j.cose.2022.102901

Kumar, P. (2024). Adversarial attacks and defenses for large language models (LLMs): methods, frameworks & challenges. *International Journal of Multimedia Information Retrieval*, *13*(3). https://doi.org/10.1007/s13735-024-00334-8

Kumar, S., Dwivedi, M., Kumar, M., & Gill, S. S. (2024). A comprehensive review of vulnerabilities and AI-enabled defense against DDoS attacks for securing cloud services. *Computer Science Review*, *53*, 100661–100661. https://doi.org/10.1016/j.cosrev.2024.100661

Lapienytė, J. (2023). *AI mistakes a panda for a gibbon. Why does it matter?* | *Cybernews*. Cybernews. https://cybernews.com/tech/ai-mistakes-panda-for-gibbon/

Li, Z., Yu, D., Wu, M., Chan, S., Yu, H., & Han, Z. (2024). Revisiting single-step adversarial training for robustness and generalization. *Pattern Recognition*, *151*, 110356. https://doi.org/10.1016/j.patcog.2024.110356

Liu, H., Li, D., & Li, Y. (2021). Poisonous Label Attack: Black-Box Data Poisoning Attack with Enhanced Conditional DCGAN. *Neural Processing Letters*, *53*(6), 4117–4142. https://doi.org/10.1007/s11063-021-10584-w

Malatji, M., & Tolah, A. (2024). Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00427-4

Malik, J., Muthalagu, R., & Pawar, P. M. (2024). A Systematic Review of Adversarial Machine Learning Attacks, Defensive Controls, and Technologies. *IEEE Access*, *12*, 99382–99421. https://doi.org/10.1109/access.2024.3423323

McCarthy, A., Ghadafi, E., Andriotis, P., & Legg, P. (2022). Functionality-Preserving Adversarial Machine Learning for Robust Classification in Cybersecurity and Intrusion Detection Domains: A Survey. *Journal of Cybersecurity and Privacy*, *2*(1), 154–190. https://doi.org/10.3390/jcp2010010

McCurdy, W. (2025). *DeepSeek Fails Every Safety Test Researchers Throw at It*. PCMAG; PCMag. https://www.pcmag.com/news/deepseek-fails-every-safety-test-thrown-at-it-by-researchers

Mehta, A., Padaria, A. A., Bavisi, D., Ukani, V., Thakkar, P., Geddam, R., Kotecha, K., & Abraham, A. (2024). Securing the Future: A Comprehensive Review of Security Challenges and Solutions in Advanced Driver Assistance Systems. *IEEE Access*, *12*, 643–678. https://doi.org/10.1109/access.2023.3347200

Miller, T., Durlik, I., Kostecka, E., Borkowski, P., & Łobodzińska, A. (2024). A Critical AI View on Autonomous Vehicle Navigation: The Growing Danger. *Electronics*, *13*(18), 3660. https://doi.org/10.3390/electronics13183660

Mintoo, A. A., Nabil, A. R., Alam, M. A., & Ahmad, I. (2024). Adversarial Machine Learning In Network Security: A Systematic Review Of Threat Vectors And Defense Mechanisms. *Innovatech Engineering Journal*, *1*(01), 80–98. https://doi.org/10.70937/itej.v1i01.9

Montagnani, M. L., Najjar, M.-C., & Davola, A. (2024). The EU Regulatory approach(es) to AI liability, and its Application to the financial services market. *Computer Law & Security Review*, *53*, 105984–105984. https://doi.org/10.1016/j.clsr.2024.105984

Muthalagu, R., Malik, J., & Pawar, P. M. (2024). Detection and prevention of evasion attacks on machine learning models. *Expert Systems with Applications*, *266*, 126044. https://doi.org/10.1016/j.eswa.2024.126044

Nguyen, S., O'Keefe, G., Arisian, S., Trentelman, K., & Alahakoon, D. (2023). Leveraging explainable AI for enhanced decision making in humanitarian logistics: An Adversarial CoevoluTION (ACTION) framework. *International Journal of Disaster Risk Reduction*, *97*, 104004. https://doi.org/10.1016/j.ijdrr.2023.104004

Novelli, C., Taddeo, M., & Floridi, L. (2023). Accountability in artificial intelligence: what it is and how it works. *AI & SOCIETY*, *39*(1). https://doi.org/10.1007/s00146-023-01635-y

Obioha-Val, O. A., Gbadebo, M. O., Olaniyi, O. O., Chinye, N. C., & Balogun, A. Y. (2025). Innovative Regulation of Open Source Intelligence and Deepfakes AI in Managing Public Trust. *Journal of Engineering Research and Reports*, *27*(2), 136–156. https://doi.org/10.9734/jerr/2025/v27i21400

Obioha-Val, O. A., Lawal, T. I., Olaniyi, O. O., Gbadebo, M. O., & Olisa, A. O. (2025). Investigating the Feasibility and Risks of Leveraging Artificial Intelligence and Open Source Intelligence to Manage Predictive Cyber Threat Models. *Journal of Engineering Research and Reports*, *27*(2), 10–28. https://doi.org/10.9734/jerr/2025/v27i21390

Obioha-Val, O. A., Olaniyi, O. O., Gbadebo, M. O., Balogun, A. Y., & Olisa, A. O. (2025). Cyber Espionage in the Age of Artificial Intelligence: A Comparative Study of State-Sponsored Campaign. *Asian Journal of Research in Computer Science*, *18*(1), 184–204. https://doi.org/10.9734/ajrcos/2025/v18i1557

Okon, S. U., Olateju, O. O., Ogungbemi, O. S., Joseph, S. A., Olisa, A. O., & Olaniyi, O. O. (2024). Incorporating Privacy by Design Principles in the Modification of AI Systems in Preventing Breaches across Multiple Environments, Including Public Cloud, Private Cloud, and On-prem. *Journal of Engineering Research and Reports*, *26*(9), 136–158. https://doi.org/10.9734/jerr/2024/v26i91269

Olabanji, S. O., Marquis, Y. A., Adigwe, C. S., Abidemi, A. S., Oladoyinbo, T. O., & Olaniyi, O. O. (2024). AI-Driven Cloud Security: Examining the Impact of User Behavior Analysis on Threat Detection. *Asian Journal of Research in Computer Science*, *17*(3), 57–74. https://doi.org/10.9734/ajrcos/2024/v17i3424

Olabanji, S. O., Olaniyi, O. O., & Olagbaju, O. O. (2024). Leveraging Artificial Intelligence (AI) and Blockchain for Enhanced Tax Compliance and Revenue Generation in Public Finance. *Asian Journal of Economics, Business and Accounting*, *24*(11), 577–587. https://doi.org/10.9734/ajeba/2024/v24i111577

Olabanji, S. O., Oluwaseun Oladeji Olaniyi, O. O., & Olaoye, O. O. (2024). Transforming Tax Compliance with Machine Learning: Reducing Fraud and Enhancing Revenue Collection. *Asian Journal of Economics Business and Accounting*, *24*(11), 503–513. https://doi.org/10.9734/ajeba/2024/v24i111572

Oladoyinbo, T. O., Olabanji, S. O., Olaniyi, O. O., Adebiyi, O. O., Okunleye, O. J., & Alao, A. I. (2024). Exploring the Challenges of Artificial Intelligence in Data Integrity and its Influence on Social Dynamics. *Asian Journal of Advanced Research and Reports*, *18*(2), 1–23. https://doi.org/10.9734/ajarr/2024/v18i2601

Olaniyi, O. O. (2024). Ballots and Padlocks: Building Digital Trust and Security in Democracy through Information Governance Strategies and Blockchain Technologies. *Asian Journal of Research in Computer Science*, *17*(5), 172–189. https://doi.org/10.9734/ajrcos/2024/v17i5447

Olaniyi, O. O., Omogoroye, O. O., Olaniyi, F. G., Alao, A. I., & Oladoyinbo, T. O. (2024). CyberFusion Protocols: Strategic Integration of Enterprise Risk Management, ISO 27001, and Mobile Forensics for Advanced Digital Security in the Modern Business Ecosystem. *Journal of Engineering Research and Reports*, *26*(6), 32. https://doi.org/10.9734/JERR/2024/v26i61160

Olateju, O. O., Okon, S. U., Olaniyi, O. O., Samuel-Okon, A. D., & Asonze, C. U. (2024). Exploring the Concept of Explainable AI and Developing Information Governance

Standards for Enhancing Trust and Transparency in Handling Customer Data. *Journal of Engineering Research and Reports*, *26*(7), 244–268. https://doi.org/10.9734/jerr/2024/v26i71206

Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., Winkler, I., & Nascimento, E. G. S. (2023). Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. *Big Data and Cognitive Computing*, *7*(1), 15. https://doi.org/10.3390/bdcc7010015

Rahman, M. H., Wuest, T., & Shafae, M. (2023). Manufacturing cybersecurity threat attributes and countermeasures: Review, meta-taxonomy, and use cases of cyberattack taxonomies. *Journal of Manufacturing Systems*, *68*, 196–208. https://doi.org/10.1016/j.jmsy.2023.03.009

Ridzuan, N. N., Masri, M., Anshari, M., Fitriyani, N. L., & Syafrudin, M. (2024). AI in the Financial Sector: The Line between Innovation, Regulation and Ethical Responsibility. *Information*, *15*(8), 432. https://doi.org/10.3390/info15080432

Salako, A. O., Fabuyi, J. A., Aideyan, N. T., Selesi-Aina, O., Dapo-Oyewole, D. L., & Olaniyi, O. O. (2024). Advancing Information Governance in AI-Driven Cloud Ecosystem: Strategies for Enhancing Data Security and Meeting Regulatory Compliance. *Asian Journal of Research in Computer Science*, *17*(12), 66–88. https://doi.org/10.9734/ajrcos/2024/v17i12530

Samuel-Okon, A. D., Akinola, O. I., Olaniyi, O. O., Olateju, O. O., & Ajayi, S. A. (2024). Assessing the Effectiveness of Network Security Tools in Mitigating the Impact of

Deepfakes AI on Public Trust in Media. *Archives of Current Research International*, *24*(6), 355–375. https://doi.org/10.9734/acri/2024/v24i6794

Shafee, A., & Awaad, T. A. (2020). Privacy attacks against deep learning models and their countermeasures. *Journal of Systems Architecture*, *114*, 101940. https://doi.org/10.1016/j.sysarc.2020.101940

Tian, P., Poreddy, S., Danda, C., Gowrineni, C., Wu, Y., & Liao, W. (2024). Evaluating Impact of Image Transformations on Adversarial Examples. *IEEE Access*, *12*, 1–1. https://doi.org/10.1109/access.2024.3487479

Val, O. O., Kolade, T. M., Gbadebo, M. O., Selesi-Aina, O., Olateju, O. O., & Olaniyi, O. O. (2024). Strengthening Cybersecurity Measures for the Defense of Critical Infrastructure in the United States. *Asian Journal of Research in Computer Science*, *17*(11), 25–45. https://doi.org/10.9734/ajrcos/2024/v17i11517

Van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. *Journal of Business Research*, *144*(1), 93–106. https://doi.org/10.1016/j.jbusres.2022.01.076

Vasani, V., Bairwa, A. K., Joshi, S., Pljonkin, A., Kaur, M., & Amoon, M. (2023). Comprehensive Analysis of Advanced Techniques and Vital Tools for Detecting Malware Intrusion. *Electronics*, *12*(20), 4299. https://doi.org/10.3390/electronics12204299

Waghela, H., Sen, J., & Rakshit, S. (2024). Refining BERT Adversarial Attacks with Projected Gradient Descent. *IEEE*, 1–7. https://doi.org/10.1109/asiancon62057.2024.10837796

Wang, S., Ko, R. K. L., Bai, G., Dong, N., Choi, T., & Zhang, Y. (2024). Evasion Attack and Defense On Machine Learning Models in Cyber-Physical Systems: A Survey. *IEEE Communications Surveys and Tutorials*, *26*(2), 1–1. https://doi.org/10.1109/comst.2023.3344808

Wang, X., Li, J., Kuang, X., Tan, Y., & Li, J. (2019). The security of machine learning in an adversarial setting: A survey. *Journal of Parallel and Distributed Computing*, *130*, 12–23. https://doi.org/10.1016/j.jpdc.2019.03.003

Wu, A. N., Stouffs, R., & Biljecki, F. (2022). Generative Adversarial Networks in the built environment: A comprehensive review of the application of GANs across data types and scales. *Building and Environment*, *223*, 109477. https://doi.org/10.1016/j.buildenv.2022.109477

Zhang, H., Zhang, X., Sun, Y., & Ji, L. (2024). Detecting adversarial samples by noise injection and denoising. *Image and Vision Computing*, *150*, 105238. https://doi.org/10.1016/j.imavis.2024.105238