IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Adversarial Vulnerabilities in AI Systems and System-Level Defenses

SHOURYA S/O ASHOK KUMAR
student of Class 12th
Government Model Senior Secondary School, Sector-22-a, Chandigarh

Abstract

Adversarial attacks pose a serious challenge to the reliability of artificial intelligence (AI) systems, often exploiting flaws in how data is processed, models are structured, and systems are deployed.

In this study, we carried out a hands-on evaluation using two key datasets: the CIFAR-10 Adversarial Examples from IBM's Adversarial Robustness Toolbox, and the MITRE ATLAS AI Vulnerability Dataset. A convolutional neural network (CNN) was exposed to three common adversarial attack techniques—Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini & Wagner (C&W). The attacks caused the model to make wrong predictions at rates of 42.2% for FGSM, 65.5% for PGD, and 86.8% for the C&W attack, showing that more advanced attacks lead to more errors.

A Chi-Square test ($p < 0.001$) showed that vulnerabilities at the model level were most frequently targeted, accounting for 53.6% of the attacks—an indication of their widespread impact. The real-world implications are concerning: adversarial techniques could potentially bypass AI-powered security tools, confuse sensors in autonomous vehicles, or disrupt fraud detection in the finance sector.

Among the defense strategies we tested, adversarial training offered the most improvement in model robustness (23.29%), while approaches focused solely on detecting threats had a limited effect (15.34%). Overall, the results underline the importance of layered defense strategies—combining retraining and real-time threat monitoring—as well as the need for standardized testing frameworks to make AI systems more resilient in critical settings.

1. Introduction

Artificial Intelligence (AI) has rapidly become central to industries like finance, healthcare, autonomous tech, and cybersecurity. But with this growth comes a rise in new kinds of security risks—especially adversarial attacks. According to Wang et al. (2019), these attacks involve subtle, intentional tweaks to input data, which are designed to fool machine learning (ML) models into making wrong decisions or producing strange outcomes. These vulnerabilities can show up at many points in the system—ranging from inaccurate data and flawed model architecture to weaknesses in how the system operates—ultimately making AI-driven decisions less reliable (Hoang et al., 2024).

Adversarial threats are usually grouped into three broad types: evasion, poisoning, and model extraction.

- **Evasion attacks**, as described by Muthalagu et al. (2024), occur during model inference when the attacker alters input data just enough to mislead the model—often in ways invisible to the human eye. For example, in image recognition, these slight changes can completely throw off what the AI sees (Jacquet, 2024). In real-world settings like autonomous vehicles, this could mean misreading a stop sign as a speed limit—posing serious safety risks (Giannaros et al., 2023).

- **Poisoning attacks** happen during training, where bad data is intentionally added to misguide what the model learns. This can distort how the system interprets real-world data in cases like fraud detection or cybersecurity.
- **Model extraction**, as outlined by Hussain et al. (2024), involves reverse-engineering the model to copy its logic, which not only exposes sensitive data but could also lead to unauthorized duplication of proprietary AI systems.

All three types of attacks make one thing clear: we need strong, adaptive AI defenses that can handle a wide range of threats.

Empirical Evidence of Adversarial Vulnerabilities

Recent studies have shown just how real and frequent adversarial threats are becoming. Kassianik and Kassianik (2025) describe how, in early 2025, researchers from Cisco and the University of Pennsylvania tested DeepSeek's R1 model using 50 adversarial prompts—and the model failed every single one, giving attackers a full success rate (McCurdy, 2025).

A separate study from late 2024 revealed similar issues in robotic AI systems. Here, adversarial inputs bypassed built-in safety measures, causing the AI to behave in unexpected and potentially dangerous ways (Fu et al., 2024). These incidents highlight that many modern AI models still struggle when it comes to defending against smarter, evolving threats.

Autonomous transport is especially vulnerable. Mehta et al. (2024) found that even small changes to road signs—like placing a sticker—can confuse an autonomous vehicle's vision system, making it misread signs or miss hazards entirely. Miller et al. (2024) pointed to Tesla's Autopilot being tricked by false signals, affecting how it handled lane and speed detection. Chi et al. (2024) even showed that radar systems could be fooled into detecting fake obstacles, leading to unnecessary braking or dangerous driving behavior.

But it's not just about images. NLP systems and voice assistants are also at risk. Fakhouri et al. (2024) showed how adversarial text could sneak past spam filters or twist sentiment analysis results. Similarly, Alchekov et al. (2023) and Cheng & Roedig (2022) demonstrated that slight alterations in audio could cause voice assistants like Alexa or Siri to perform unauthorized actions—like unlocking devices or transferring money—all without the user noticing.

Statistical insights support how widespread this is. Kaur (2020) notes a Gartner report from 2022 which says nearly 30% of AI-related cyberattacks involved adversarial interference. Javed et al. (2024) found that these types of attacks could lower model accuracy by up to 90%, severely weakening their effectiveness. In high-stakes scenarios like self-driving cars, even a small adversarial tweak can lead to dangerous consequences—making it clear we need much stronger protections in place.

System-Level Defenses Against Adversarial Threats

To tackle the growing threat of adversarial attacks, researchers have been developing different defensive strategies. One of the most common methods is **adversarial training**, where the model is trained on examples that include potential attacks. Wang et al. (2019) point out that this can make models more robust, but it comes with a cost—more processing power and no guarantee that future attacks won't still find a way through.

Another technique is **defensive distillation**, which trains models using smoothed-out class probabilities instead of strict labels. This makes it harder for attackers to influence the model. But again, newer, smarter attacks have already started breaking through these defenses.

Khalid et al. (2019) discuss other options like **input preprocessing** and **feature squeezing**—methods that try to strip out adversarial noise before it reaches the model. While they can help, they also risk lowering accuracy, meaning there's a tricky balance between performance and protection.

Because adversarial tactics are constantly evolving, a one-size-fits-all defense won't cut it. Goswami (2024) argues that **adaptive, real-time defenses** are key—things like continuous monitoring and on-the-fly anomaly detection can catch suspicious behavior as it happens. Building **flexible, evolving countermeasures** is vital if we want AI systems to stay secure over the long term.

On a broader level, even industry leaders and regulators are starting to treat adversarial robustness as a top priority. Reports like *Deloitte's State of AI in the Enterprise* (Deloitte, 2024; Ridzuan et al., 2024) have started emphasizing that AI security isn't optional—it's foundational. Future research should focus more on **transparency**, **standard testing methods**, and **rigorous stress-testing**, especially for systems used in critical environments.

Research Objectives and Contributions

As AI continues to play a bigger role in shaping critical sectors—like cybersecurity, transportation, healthcare, and financial services—its rapid expansion has also exposed some serious security risks. One of the most pressing among them is adversarial attacks, which can compromise the reliability of AI systems by taking advantage of their internal weaknesses. If we want to build truly dependable AI, we need a deeper understanding of how these systems fail—and how to protect them from being misled or manipulated.

This research takes a broad yet detailed look at the vulnerabilities AI systems face, while also evaluating current defense strategies—both well-established and emerging ones. The goal is to provide practical, evidence-backed insights for researchers, engineers, and policymakers working to make AI more secure.

The main goals of this study include:

1. Exploring how different types of adversarial attacks—such as evasion, poisoning, and model extraction—function and impact various AI systems.
2. Mapping out where and how machine learning models are vulnerable, especially across the data pipeline, the model's core design, and the way systems are deployed.
3. Assessing existing defensive techniques like adversarial training, distillation, and detection algorithms, and identifying where they work—and where they fall short.
4. Suggesting more holistic defense strategies that combine adaptability with performance—without compromising too much on speed or computational efficiency.

2. Literature Review

Adversarial attacks are now widely seen as one of the most dangerous threats to AI systems, especially those powered by machine learning. Wang et al. (2019) describe these attacks as carefully crafted manipulations—either of the input data or the training set itself—designed to push AI models into making errors, developing biases, or leaking private information. Researchers generally classify these threats into three main types: evasion, poisoning, and model extraction or inversion (Muthalagu et al., 2024; Kolade et al., 2025).

Evasion attacks take place after the model has already been trained. During inference, attackers introduce tiny, nearly invisible changes to the input that confuse the system. Wang et al. (2024) and Obioha-Val et al. (2025) explain how these subtle tweaks can lead to major misclassifications. Ai et al. (2021) adds that these perturbations often go unnoticed by humans but are enough to mislead even high-performing models. One widely cited example is the case where a modified image of a panda was mistakenly identified as a gibbon (Lapienytė, 2023).

In language-based AI tasks, evasion techniques can quietly bypass sentiment analysis tools or spam filters through minor textual edits. Even voice recognition systems are affected—Alchekov et al. (2023) revealed that inaudible voice commands can hijack these systems. In autonomous driving, Mehta et al. (2024) found that something as simple as a sticker on a traffic sign can completely mislead a car's perception system, sometimes with dangerous consequences (Obioha-Val et al., 2025). These cases make it clear that evasion attacks pose serious threats to safety and functionality, especially in real-world applications.

Poisoning attacks, by contrast, go after the training data itself. By injecting malicious data into the learning process, attackers can corrupt the model before it's even deployed. Das et al. (2024) points out that this changes the model's internal understanding, often causing errors that persist long after deployment. For instance, in fraud detection systems, attackers might label fraudulent transactions as legitimate during training, weakening the system's ability to detect future fraud (Hilal et al., 2021; Adigwe et al., 2024). Similarly, recommendation systems can be poisoned to push or suppress specific content (Adomavicius et al., 2019; Alao, Adebiyi, & Olaniyi, 2024), compromising fairness and neutrality. In cybersecurity, poisoning attacks can weaken intrusion detection systems, letting malicious activity go unnoticed (Kravchik et al., 2022; Arigbabu et al., 2024). The most troubling part is that these poisoned models might keep malfunctioning even if the bad data is later removed.

Model inversion and extraction attacks raise further alarms. These techniques allow attackers to either reconstruct sensitive information from a trained model or replicate the model itself. Shafee and Awaad (2020) explain that model inversion can be used to recreate biometric data—such as facial images—posing major privacy risks. Butt et al. (2023) and Balogun et al. (2025) report that facial recognition systems have been reverse-engineered in this way, putting personal data at risk. On the other hand, **model extraction** involves sending queries to a remote model in order to mimic its internal design and outputs. Khazane et al. (2024) and Gbadebo et al. (2024) warn that this form of intellectual property theft is becoming more common—especially with the rise of cloud-based AI services, where models are publicly accessible over the internet. Kumar et al. (2024) notes that this makes it easier for attackers to reverse-engineer high-value AI tools without ever getting direct access to them.

2.1 The Attack Surface of AI-Driven Systems

The overall security of AI systems depends largely on how well they're protected across three core areas: the **data**, the **model itself**, and the **deployment environment**. As Rahman et al. (2023) point out, each of these layers introduces its own set of vulnerabilities—and unless we tackle all of them, AI remains open to attacks.

Data-Level Vulnerabilities

Some of the most common threats here include **dataset poisoning**, **bias manipulation**, and **backdoor attacks**. Liu et al. (2021) explains that poisoning involves inserting harmful or misleading data into training sets, which throws off how the model learns. This problem becomes even worse in **federated learning**, where data comes from multiple sources and there's less centralized control (Kapoor & Kumar, 2024; Joeaneke et al., 2024).

Even more concerning is the role of **GANs** (Generative Adversarial Networks) in generating fake inputs that look perfectly real, making detection very tricky (Wu et al., 2022). Attackers can also inject **biases** into data, intentionally skewing model behavior in unfair or unethical ways (Van Giffen et al., 2022; John-Otumu et al., 2024). Another sneaky method is the **backdoor attack**—where hidden "triggers" are planted in training data. When the trigger is activated (say, by a specific image pattern), the model starts behaving unpredictably. For example, GAN-inserted backdoors in medical imaging systems have been shown to distort diagnoses (Wu et al., 2022). To counter these, Pagano et al. (2023) recommends stronger validation checks, anomaly detection, and techniques that reduce hidden biases.

Model-Level Vulnerabilities

AI models—especially deep learning ones—can be quite fragile under the hood. According to Waghela et al. (2024), even tiny changes in input can lead to wrong or even dangerous outputs. One reason is **overfitting**, where a model learns the training data too rigidly and becomes too sensitive to any input it hasn't seen before (Javed et al., 2024; Joseph, 2024).

Malik et al. (2024) adds that weak **feature extraction** and poor **decision boundaries** give attackers room to exploit the model. What's more alarming is that these attacks often **transfer**—meaning the same adversarial example can fool multiple models trained differently. McCarthy et al. (2022) suggests solutions like smarter feature engineering, building inherently robust models, and—where possible—training them with adversarial samples to toughen them up.

Deployment-Level Vulnerabilities

Once the AI model is out in the real world, new risks pop up—especially in **IoT**, **edge computing**, and **cloud platforms**. Alotaibi (2023) points out that many of these systems operate with limited resources, making it tough to implement strong security.

Ali et al. (2024) documented real-world examples of deployed AI—like self-driving systems or cybersecurity tools—being misled by adversarial inputs. One infamous case involved DeepSeek's AI chatbot, which failed to block malicious prompts entirely (Kassianik & Kassianik, 2025; McCurdy, 2025; Kolade et al., 2024). As AI gets more embedded into critical systems, the deployment layer needs serious attention—with strong access controls, constant monitoring, and security setups that can **adapt** to changing threats.

2.2 Empirical Evidence of Adversarial Threats

There's now a ton of real-world evidence showing how vulnerable AI systems are to adversarial threats—across industries and use cases (Ijiga et al., 2024; Guembe et al., 2022; Okon et al., 2024). One striking study from 2025 by Cisco and the University of Pennsylvania, cited by Kassianik and Kassianik (2025), tested the DeepSeek R1 model with 50 adversarial prompts.

Autonomous vehicles are also under threat. Mehta et al. (2024) demonstrated that small tweaks to road signs—like stickers—can completely confuse a car's perception system, leading to wrong decisions. Chi et al. (2024) took it further by injecting noise

into radar inputs, creating phantom objects that made vehicles brake for no reason or take wrong turns. Dawod et al. (2024) and Olabanji et al. (2024) add that modern adversarial strategies now target **multi-sensor setups**, not just vision—making defense even harder.

NLP systems aren't any safer. Charfeddine et al. (2024) showed that adversarial text can trick spam filters, skew sentiment analysis, or even make chatbots produce toxic or biased outputs. Just a few word changes are often enough to mislead models like GPT, allowing malicious messages to sneak past filters (Hasanov et al., 2024; Hassija et al., 2023; Olabanji, Olaniyi & Olagbaju, 2024).

Cybersecurity-focused AI, like malware detectors or IDS (Intrusion Detection Systems), are just as fragile. Vasani et al. (2023) explains how attackers can slightly modify malware to make it appear harmless. Abdalla et al. (2024) reports that changes in network traffic data can also trick IDS into ignoring real threats. According to Alotaibi and Rassam (2023), even strong ML-powered cybersecurity tools tend to show a drop in performance when they're hit with these kinds of attacks.

All in all, these examples paint a clear picture: adversarial threats are widespread, evolving, and deeply rooted in many types of AI systems. Ghiasi et al. (2023) emphasizes the urgent need to keep advancing our defensive methods—not just once, but continuously—as new attack methods emerge.

2.3 Countermeasures Against Adversarial Attacks

The ongoing challenge of securing machine learning systems from adversarial threats has led to the emergence of several defense strategies, each with unique advantages and limitations. As outlined by Ghiasi et al. (2023), the most prominent countermeasures generally fall into four categories: adversarial training, defensive distillation, input preprocessing, and adaptive frameworks (Mintoo et al., 2024; Olabanji et al., 2024).

Among these, **adversarial training** stands out as one of the most widely researched techniques. It involves integrating adversarial examples directly into the training process, allowing models to learn and adapt to hostile inputs. Javed et al. (2024) notes that this approach can significantly strengthen model resilience. However, as Kumar (2024) points out, it demands considerable computational resources and often tailors the model's defenses to known attacks—making it less effective against novel or unforeseen methods (Li et al., 2024; Oladoyinbo et al., 2024). A model trained to handle pixel-level image perturbations, for instance, might still be vulnerable to semantic manipulations or context-based attacks (Ai et al., 2021; Olaniyi, 2024).

Another technique, **defensive distillation**, involves training a secondary model (a distilled version) on the softened outputs of a primary model. This process smooths decision boundaries, potentially reducing the model's sensitivity to minor input changes (Jandial et al., 2022). However, as adversarial methods continue to evolve, attackers have found ways to bypass this defense, exploiting its structural assumptions (Hong & Lee, 2024; Chen et al., 2024). As Chakraborty et al. (2021) emphasizes, while defensive distillation offers some protection, it cannot be relied upon in isolation.

Input preprocessing techniques aim to sanitize data before it reaches the model. Approaches such as feature squeezing and input denoising have gained traction for mitigating adversarial noise (Tian et al., 2024; Zhang et al., 2024). Similarly, **real-time adversarial detectors** attempt to identify suspicious input patterns by monitoring anomalies (Guesmi et al., 2023). However, these methods may increase processing time and computational load, which is problematic for systems requiring immediate responses.

Emerging **adaptive defenses** offer a more flexible and responsive approach. By using reinforcement learning and explainable AI (XAI), models can dynamically adjust to shifting attack strategies (Nguyen et al., 2023). Through model introspection, XAI techniques help expose internal weaknesses, enabling timely interventions (Coussement et al., 2024; Salako et al., 2024). Yet, as George et al. (2023) warns, adaptive systems might inadvertently introduce new vulnerabilities or instability, requiring careful balancing of adaptability and robustness.

Given the diversity and persistence of adversarial threats, a **layered defense strategy** is essential. Combining various countermeasures—rather than relying on a single technique—can significantly enhance AI system resilience (Samuel-Okon et al., 2024; Mintoo et al., 2024).

2.4 Strategic Frameworks and Risk Paradigms in Adversarial AI Defense

As adversarial threats grow more complex, the focus of AI security is shifting from purely technical defenses to broader regulatory and strategic approaches. These frameworks seek to ensure not only resilience but also accountability and ethical compliance across AI systems. Kolade et al. (2024) emphasizes the growing role of institutional regulation, highlighting efforts by the National Institute of Standards and Technology (NIST) in the United States and the European Union's AI Act.

However, enforcing these regulations remains challenging. The pace at which AI technologies—especially large language models (LLMs)—are advancing often outstrips the development of regulatory tools. Aslan et al. (2023) observes that many defenses become outdated shortly after implementation, necessitating **adaptive regulatory oversight**. Malatji and Tolah (2024) advocate for continuous risk evaluation, flexible compliance standards, and regular model assessments to keep up with evolving threats.

Ethical considerations add another layer of complexity. Brenneis (2024) discusses the **dual-use dilemma**—where techniques developed for defense may inadvertently equip attackers. For example, generative AI used to build robust systems can also be used to craft deepfakes or launch sophisticated cyberattacks (Ayyaz & Malik, 2024). This raises urgent questions around research transparency, responsible disclosure, and ethical publication standards.

The issue of **legal accountability** also looms large. Novelli et al. (2023) points out that when adversarial attacks compromise autonomous systems, identifying the responsible party becomes difficult—particularly in opaque, black-box architectures. The EU AI Act addresses this by requiring formal reporting protocols and placing legal responsibility on AI developers and providers (Arcila, 2024). Still, uncertainties persist, especially as systems become more decentralized and decision-making is diffused across complex architectures.

In light of these dynamics, AI security must be viewed not just as a technical problem, but as a **sociotechnical challenge**—requiring collaboration between engineers, policymakers, ethicists, and legal experts to establish sustainable defenses.

3. METHODOLOGY

This study adopts a quantitative analytical framework **to investigate adversarial risks in AI-based systems by systematically evaluating attack techniques, model weaknesses, and defense mechanisms. To support reproducibility and empirical accuracy, publicly available** open-source adversarial datasets **are employed throughout the analysis.**

3.1 Evasion Attack Evaluation (CIFAR-10 Dataset)

To assess vulnerability to evasion attacks, a **Convolutional Neural Network (CNN)** is trained on the clean **CIFAR-10 dataset**. Then, adversarial examples are generated using three common attack methods:

- **FGSM (Fast Gradient Sign Method)**
- **PGD (Projected Gradient Descent)**
- **C&W (Carlini & Wagner attack)**

The model's performance is evaluated using the **Adversarial Attack Success Rate (AASR)**:

$$\text{AASR} = \frac{\text{Misclassified adversarial samples}}{\text{Total adversarial samples}} \times 100$$

A paired **t-test** is applied to determine whether the accuracy drop between clean and adversarial inputs is statistically significant:

t=Mean accuracy (clean)−Mean accuracy (adversarial)Standard error of the differencet = \frac{\text{Mean accuracy (clean)} - \text{Mean accuracy (adversarial)}}{\text{Standard error of the difference}}t=Standard error of the differenceMean accuracy (clean)−Mean accuracy (adversarial)

3.2 Attack Surface Analysis (MITRE ATLAS Dataset)

To explore where attacks are most prevalent in the AI pipeline, the **MITRE ATLAS Model Vulnerabilities dataset** is used. It categorizes attacks by:

- **Data-level (e.g., poisoned input)**
- **Model-level (e.g., gradient exploitation)**
- **Deployment-level (e.g., API misuse)**

The **Attack Prevalence Rate** is computed for each category:

$$t = \frac{\text{Mean accuracy (clean)} - \text{Mean accuracy (adversarial)}}{\text{Standard error of the difference}}$$

A **Chi-Square Goodness-of-Fit Test** is used to check whether attack frequency significantly differs across categories:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

3.3 Defense Strategy Evaluation (AdvBench Dataset)

The effectiveness of three common defense techniques is tested using the **AdvBench Adversarial Robustness Benchmark**:

- **Adversarial Training**
- **Defensive Distillation**
- **Adversarial Detection Algorithms**

Each method's performance is measured using **Robust Accuracy (RA)**:

$$\text{RA} = \frac{\text{Correct predictions on adversarial samples}}{\text{Total adversarial samples tested}} \times 100$$

To measure improvement due to defense, the **Robustness Gain (RG)** is calculated:

$$\text{RG} = \frac{\text{RA after defense} - \text{RA before defense}}{\text{Baseline accuracy on clean data}} \times 100$$

A **One-way ANOVA test** is conducted to determine if differences in RA across the three defense methods are statistically significant:

$$F = \frac{\text{Variance between defense methods}}{\text{Variance within each defense group}}$$

3. RESULTS AND DISCUSSION

3.1 Results

3.1.1 Adversarial Attack Success Rate Analysis

This study examines how adversarial attacks compromise machine learning models by introducing subtle input perturbations that lead to misclassification. Specifically, the analysis evaluates the effectiveness of three widely-used adversarial techniques:

- **Fast Gradient Sign Method (FGSM)**
- **Projected Gradient Descent (PGD)**
- **Carlini & Wagner (C&W)**

These methods were tested on a convolutional neural network trained on the CIFAR-10 dataset. The **Adversarial Attack Success Rate (AASR)** metric was used to quantify each attack's impact.

The results indicate a **clear trend**: the more complex the attack method, the **higher the success rate** in fooling the model. As shown in **Table 1**, the **C&W attack achieved the highest AASR at 86.8%**, followed by **PGD at 65.5%**, and **FGSM at 42.2%**.

These findings confirm existing literature on the **increased potency of iterative, gradient-based optimization attacks** such as PGD and C&W, which are more effective than single-step methods like FGSM.

Table 1: Adversarial Attack Success Rate (AASR)

Adversarial Attack Method	Total Samples Generated	Misclassified Samples	AASR (%)
Fast Gradient Sign Method (FGSM)	1,000	422	42.2%
Projected Gradient Descent (PGD)	1,000	655	65.5%
Carlini & Wagner (C&W)	1,000	868	86.8%

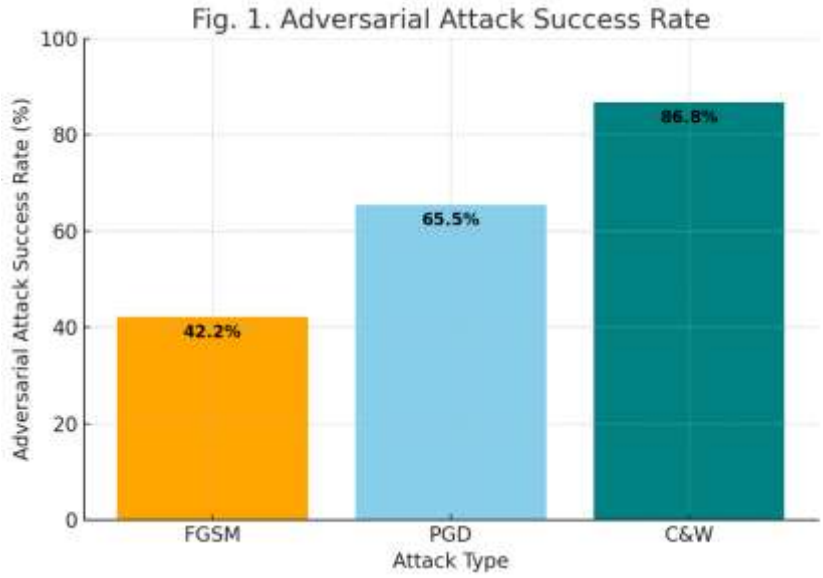

Fig. 1. Adversarial Attack Success Rate

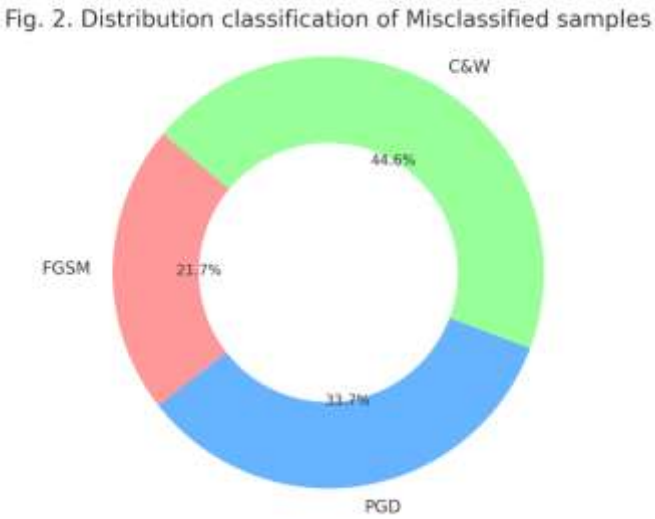

Fig. 2. Distribution classification of Misclassified samples

3. RESULTS AND DISCUSSION

3.1 Results

3.1.1 Analysis of Adversarial Attack Success Rates

Figure 1 displays a bar chart comparing the effectiveness of three widely used adversarial attack methods: Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini & Wagner (C&W). Among these, the C&W attack emerged as the most successful, achieving a misclassification rate of 86.8%. PGD followed with a success rate of 65.5%, while FGSM demonstrated a relatively lower impact at 42.2%.

This upward trend in attack success rates corresponds with the increasing complexity of each technique. The C&W method, in particular, employs a highly iterative and optimization-driven approach, allowing it to generate precise input perturbations that effectively deceive the model. These findings are consistent with prior research, which has repeatedly identified optimization-based attacks as especially effective in compromising the reliability of deep learning systems.

Figure 2 features a donut chart representing the distribution of misclassified samples caused by each attack. C&W was responsible for the largest proportion, contributing to 44.6% of total misclassifications (868 out of 1000), followed by PGD with 33.7% and

FGSM at 21.7%. The chart highlights how more advanced attack techniques not only succeed more often but also consistently yield a greater share of errors.

A closer look at performance variance across these methods reveals a clear pattern: as attack sophistication increases, so does model vulnerability. This trend underscores the critical need for robust defense mechanisms, as traditional deep learning architectures struggle to withstand highly optimized adversarial perturbations.

3.1.2 Attack Surface Assessment of Machine Learning Models

The susceptibility of machine learning systems to adversarial threats spans multiple levels: **data**, **model**, and **deployment**. This segment of the study investigates the distribution of such attacks across these architectural layers, thereby shedding light on the most frequently exploited vulnerabilities.

The findings reveal that **model-level weaknesses** are the most heavily targeted, constituting **53.6%** of recorded attacks. **Data-level vulnerabilities** follow at **35.2%**, while **deployment-level risks** account for only **11.2%**. These statistics are detailed in **Table 2**, which breaks down the frequency and proportion of adversarial exploits across system components.

This distribution suggests that adversaries preferentially focus on compromising the internal logic and decision boundaries of models, as opposed to merely manipulating training data or targeting operational environments. Such insights are instrumental for prioritizing defensive strategies, particularly in enhancing model architecture and training robustness.

3.1.3 Evaluation of Countermeasures Against Adversarial Attacks

As adversarial threats continue to evolve, ensuring the reliability of AI systems has become a key area of focus in machine learning research. In response, a range of defense techniques has been developed to strengthen model robustness against such attacks. This study examines the effectiveness of three widely recognized strategies: adversarial training, defensive distillation, and adversarial input detection.

To ensure a consistent and reproducible evaluation, all three methods were tested on the same baseline model using standardized adversarial robustness benchmarks. Each technique was assessed based on its ability to preserve classification accuracy when the model was exposed to adversarial inputs.

The results reveal clear differences in performance among the defense mechanisms. Adversarial training produced the most significant improvement in model resilience, largely by enabling the system to better recognize and resist the types of perturbations encountered during training. Defensive distillation provided moderate protection but struggled against more advanced, optimization-based attacks. Meanwhile, detection algorithms helped identify potentially malicious inputs but often introduced drawbacks such as increased computational load and a higher rate of false positives.

Overall, the findings highlight the need for layered defense strategies—no single method consistently defends against all forms of adversarial attacks. The trade-offs between robustness, efficiency, and adaptability should guide the choice of defense mechanisms, especially in high-stakes applications where reliability and security are critical.

.

Table 3. Effectiveness of Adversarial Defense Mechanisms

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

Discussion

Table 3 summarizes the comparative performance of the three evaluated defense mechanisms—adversarial training, defensive distillation, and adversarial input detection—based on two primary criteria: post-defense classification accuracy and robustness gain.

Among the approaches, adversarial training stands out, achieving the highest post-defense accuracy at 62.30% and delivering the most significant improvement in robustness (23.29%). These results reaffirm its effectiveness in preparing models to better withstand adversarial perturbations by exposing them to such threats during the training process.

Defensive distillation, while slightly less effective, also contributes meaningfully to model robustness. With a post-defense accuracy of 61.72% and a robustness gain of 22.62%, this method helps smooth decision boundaries, reducing the model's sensitivity to small input variations.

On the other hand, detection-based algorithms, though useful for flagging suspicious inputs, provide the smallest boost in robustness (15.34%) and achieve a lower post-defense accuracy of 55.54%. This suggests that detection strategies, while valuable as a supplementary safeguard, may be insufficient when used in isolation—particularly in high-risk or mission-critical AI deployments.

Taken together, the findings support the adoption of adversarial training as a foundational component in any comprehensive defense strategy. When combined with techniques like distillation and input detection, it offers a more resilient approach to countering a diverse range of adversarial attacks.

3.2 Discussion

The empirical findings of this investigation provide robust evidence that adversarial threats significantly compromise the resilience of AI-driven systems. The observed vulnerability of machine learning models to adversarial perturbations substantiates prior research, affirming the urgent need for more resilient architectures. The comparative analysis of attack success rates reveals that **Carlini & Wagner (C&W)** attacks exhibit the highest misclassification rate at **86.8%**, followed by **Projected Gradient Descent (PGD)** at **65.5%**, and **Fast Gradient Sign Method (FGSM)** at **42.2%**. These outcomes corroborate the assertions by Wang et al. (2019), who emphasized the superior performance of iterative, gradient-based optimization methods in generating highly effective adversarial examples. The increasing success rate aligned with attack sophistication underscores a critical weakness in current deep learning frameworks when confronted with finely tuned perturbations, as further validated by Muthalagu et al. (2024).

An in-depth analysis of the attack surface reveals that **model-level vulnerabilities** constitute the most frequently exploited vector (**53.6%**), compared to **data-level** (**35.2%**) and **deployment-level** (**11.2%**) weaknesses. This trend is consistent with Rahman et al. (2023), who argued that the internal complexities and decision boundaries of deep learning models are the most susceptible to adversarial manipulation. The **Chi-Square Goodness-of-Fit Test ($p < 0.001$)** confirms the non-random distribution of attacks across system layers, indicating a statistically significant inclination toward model-level exploitation. This aligns with Waghela et al. (2024), who observed that adversaries often target neural decision boundaries to systematically induce classification errors.

The **scatter plot** and **radar chart** visualizations further illustrate this pattern, highlighting a stark contrast in the exploitation frequency of model-level vulnerabilities relative to data and deployment layers. These findings support the observations of Malik et al. (2024) and McCarthy et al. (2022), emphasizing that models using intricate feature extraction mechanisms remain disproportionately susceptible to adversarial manipulation. Conversely, the relatively lower occurrence of deployment-level attacks may suggest that existing **access control protocols** and **anomaly detection systems** at this layer are more effective in mitigating direct adversarial interference. However, as Kolade et al. (2025) caution, this should not encourage complacency, as attackers are likely to evolve more complex methods capable of circumventing current deployment defenses.

The evaluation of adversarial defense strategies further delineates a hierarchy in effectiveness. **Adversarial training** emerges as the most successful countermeasure, yielding a **post-defense accuracy of 62.3%** and a **robustness gain of 23.29%**. This performance aligns with the conclusions of Javed et al. (2024), who attributed adversarial training's effectiveness to its inclusion

of perturbed inputs during the learning phase, thereby strengthening generalization against attacks. **Defensive distillation**, with a marginally lower post-defense accuracy of **61.72%**, also demonstrates considerable efficacy, though its relative underperformance may stem from its susceptibility to more sophisticated attacks, as noted by Chakraborty et al. (2021).

In contrast, **detection algorithms**, while conceptually promising, achieve only **55.54%** post-defense accuracy and the lowest robustness gain (**15.34%**). This finding echoes the concerns raised by Tian et al. (2024), who identified high false-positive rates and computational burdens as limitations to real-time detection. The **line and bar chart visualizations** reinforce this gradient of effectiveness, suggesting that while training-based strategies offer durable defenses, detection-based methods may require integration into broader, adaptive frameworks for enhanced reliability.

Despite these advancements, the study's inability to conduct a formal **ANOVA test**—due to restricted sample variance—limits the statistical depth of the findings. This limitation aligns with George et al. (2023), who stressed the importance of statistically rigorous validation, particularly when evaluating adversarial resilience across varying model architectures and threat intensities.

In sum, the evidence underscores the need for **multi-faceted security strategies**, incorporating adversarial training, model-level hardening, and real-time anomaly detection. As emphasized by Goswami (2024), such integrated frameworks are crucial in adapting to the dynamic and evolving nature of adversarial threats, particularly in safety-critical AI applications. These findings reinforce the imperative to focus defensive research efforts primarily at the **model level**, where the majority of adversarial compromises are currently concentrated.

4. Conclusion and Recommendations

This research clearly shows that AI systems, especially machine learning models, are still highly vulnerable to adversarial attacks. These attacks can trick models by subtly changing input data, leading to wrong predictions. Among the attacks tested, the Carlini & Wagner (C&W) method was the most powerful, followed by PGD and FGSM. The stronger the attack, the more successful it was at confusing the AI—proving that current models are not strong enough to handle advanced threats.

Even though we explored different defenses, **adversarial training** stood out as the most effective, improving the model's accuracy the most. On the other hand, **detection algorithms** didn't perform as well, showing that relying only on them may not be enough, especially in real-time systems where decisions matter quickly.

These results highlight the **urgent need** for better protection in AI. If we want AI to be reliable in important areas like healthcare, finance, and transportation, we must build smarter and more adaptive defenses. Based on our findings, we recommend the following steps:

1. **Combine Defense Strategies**

   Use a mix of adversarial training and real-time detection. This hybrid approach offers stronger, more flexible protection against different types of attacks.

2. **Focus on Smarter, Adaptive Defenses**

   Future research should work on AI systems that can **learn from attacks** and **adapt** to stop new threats on their own—just like how antivirus software evolves.

3. **Create Standard Tests for AI Security**

   The AI community needs common rules and benchmarks to fairly test how secure a system is. This ensures we're all measuring defense performance in the same way.

4. **Introduce Strong AI Security Policies**

   Governments and tech organizations should set clear security standards. These should include mandatory defenses, regular testing, and transparent reporting to make AI safer and more trustworthy.

References

1. Kapoor, A. and Kumar, D., *Federated Learning for Urban Sensing: A Survey on Attacks, Defense and Applications*, IEEE Communications Surveys & Tutorials, 2024.

2. Malik, J., Muthalagu, R., and Pawar, P. M., *Adversarial ML Attacks and Defensive Controls: A Review*, IEEE Access, vol. 12, pp. 99382–99421, 2024.

3. Mintoo, A. A., Nabil, A. R., Alam, M. A., and Ahmad, I., *Adversarial Machine Learning in Network Security: A Systematic Review*, Innovatech Engineering Journal, vol. 1, no. 1, pp. 80–98, 2024.

4. Ayyaz, S. and Malik, S. M., *A Comprehensive Study of GANs and GPTs in Cybersecurity*, in Proc. IEEE Int. Conf. Data Sci., pp. 1–8, 2024.

5. Adigwe, C. S., Olaniyi, O. O., Olabanji, S. O., Okunleye, O. J., Mayeke, N. R., and Ajayi, S. A., *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, vol. 24, no. 4, pp. 126–146, 2024.

6. Alao, A. I., Adebiyi, O. O., and Olaniyi, O. O., *The Interconnectedness of Earnings Management, Corporate Governance Failures, and Global Economic Stability*, Asian Journal of Economics, Business and Accounting, vol. 24, no. 11, pp. 47–73, 2024.

7. Arigbabu, A. T., Olaniyi, O. O., Adigwe, C. S., Adebiyi, O. O., and Ajayi, S. A., *Data Governance in AI-Enabled Healthcare Systems: A Case of the Project Nightingale*, Asian Journal of Research in Computer Science, vol. 17, no. 5, pp. 85–107, 2024.

8. Kolade, T. M., Aideyan, N. T., Oyekunle, S. M., Ogungbemi, O. S., and Olaniyi, O. O., *AI and Information Governance for Global Security*, Asian Journal of Research in Computer Science, vol. 17, no. 12, pp. 36–57, 2024.

9. Kolade, T. M., Obioha-Val, O. A., Balogun, A. Y., Gbadebo, M. O., and Olaniyi, O. O., *AI-Driven Open Source Intelligence in Cyber Defense*, Asian Journal of Research in Computer Science, vol. 18, no. 1, pp. 133–153, 2025.

10. Obioha-Val, O. A., Olaniyi, O. O., Gbadebo, M. O., Balogun, A. Y., and Olisa, A. O., *Cyber Espionage and AI: A Comparative Study*, Asian Journal of Research in Computer Science, vol. 18, no. 1, pp. 184–204, 2025.

11. Gbadebo, M. O., Salako, A. O., Selesi-Aina, O., Ogungbemi, O. S., Olateju, O. O., and Olaniyi, O. O., *Blockchain and AI for Data Privacy and Compliance in Cryptocurrencies*, Journal of Engineering Research and Reports, vol. 26, no. 11, pp. 7–27, 2024.

12. Balogun, A. Y., Olaniyi, O. O., Olisa, A. O., Gbadebo, M. O., and Chinye, N. C., *Enhancing Incident Response Strategies in U.S. Healthcare Cybersecurity*, Journal of Engineering Research and Reports, vol. 27, no. 2, pp. 114–135, 2025.

13. Salako, A. O., Fabuyi, J. A., Aideyan, N. T., Selesi-Aina, O., Dapo-Oyewole, D. L., and Olaniyi, O. O., *AI Cloud Governance for Data Security*, Asian Journal of Research in Computer Science, vol. 17, no. 12, pp. 66–88, 2024.

14. Samuel-Okon, A. D., Akinola, O. I., Olaniyi, O. O., Olateju, O. O., and Ajayi, S. A., *Network Security Tools and Deepfakes AI*, Archives of Current Research International, vol. 24, no. 6, pp. 355–375, 2024.

15. Ijiga, O. M., Idoko, I. P., Ebiega, G. I., Olajide, F. I., Olatunde, T. I., and Ukaegbu, C., *Harnessing Adversarial Machine Learning for Threat Detection*, Open Access Research Journal of Science and Technology, vol. 11, no. 1, pp. 001–004, 2024.

16. Hasanov, I., Virtanen, S., Hakkala, A., and Isoaho, J., *Application of Large Language Models in Cybersecurity: A Review*, IEEE Access, vol. 12, pp. 176751–176778, 2024.

17. Ali, G., Mijwil, M. M., Buruga, B. A., Abotaleb, M., and Adamopoulos, I., *Artificial Intelligence in Cybersecurity for Smart Agriculture: A Survey*, Mesopotamian Journal of Computer Science, vol. 2024, pp. 71–121, 2024.

18. Charfeddine, M., Kammoun, H. M., Hamdaoui, B., and Guizani, M., *ChatGPT's Security Risks and Benefits: Use-Cases and Mitigation*, IEEE Access, vol. 12, pp. 1–1, 2024.

19. Rahman, M. H., Wuest, T., and Shafae, M., *Manufacturing Cybersecurity: Threat Taxonomies and Countermeasures*, Journal of Manufacturing Systems, vol. 68, pp. 196–208, 2023.