International Journal of
Science, Architecture, Technology and Environment

Detecting and Defending Against Adversarial Attacks on AI Powered Security Systems

Saeed Hubairik Aliyu[1*], Abosede Oluwabunmi Adelore[2], Chijioke C. Ekechi[3], Hanafi Musa Olayinka[4], Oluwasola Abiodun Elijah[5], Unuigbokhai Peter Afegbai[6]

[1]*Department of Mechatronics and Robotics Engineering, South Ural State University, Russia*

[2]*Department of Computer Science, University of Ghana, Ghana*

[3]*Department of Electrical and Computer Engineering, Tennessee Technological University, USA*

[4]*Department of Computer Science & Engineering Technology, University of Houston Downtown , USA*

[5]*Department of Mechanical Engineering , University of Ilorin*

[6]*Auchi Polytechnic, Edo State, Nigeria.*

[*]*Corresponding author*

DOI: https://doi.org/10.63680/ijsate0725029.16

Abstract

Adversarial attacks pose a critical threat to the security and reliability of artificial intelligence (AI) powered systems, exploiting model vulnerabilities to manipulate outcomes and compromise sensitive operations. These attacks (ranging from evasion and data poisoning to model extraction and prompt injection) undermine the integrity of AI models across high stakes applications, including healthcare, autonomous systems, and cybersecurity. This paper explores the taxonomy of adversarial attacks, their operational impact, and the current landscape of defence strategies designed to detect and mitigate them. Emphasis is placed on proactive and adaptive approaches such as adversarial training, real time monitoring, input preprocessing, and explainable AI techniques. While notable progress has been made, the evolving nature of these threats continues to challenge existing defences, demanding more scalable, interpretable, and resource efficient solutions. The study underscores the urgency of cross disciplinary collaboration to develop resilient AI systems capable of withstanding adversarial manipulation in increasingly complex environments.

Keywords: Adversarial Attacks, AI Security Evasion Attacks, Data Poisoning, Model Extraction, Prompt Injection, Intrusion Detection, Adversarial Training, Input Preprocessing, Explainable AI (XAI), Machine Learning, Robustness, AI Threat Mitigation, Secure AI Systems

1 Introduction

Adversarial attacks on artificial intelligence (AI) systems represent a significant and growing threat to the

security and reliability of AI powered technologies. These malicious techniques exploit inherent vulnerabilities within AI models, resulting in harmful consequences such as privacy violations, operational failures, and potential endangerment of public safety. As AI systems are increasingly integrated into critical sectors (ranging from healthcare to autonomous vehicles), the need to detect and defend against these attacks has become paramount, highlighting the pressing challenges that developers and researchers face in securing AI technologies against malicious exploitation. [1], [2].

Adversarial attacks can be categorised into various forms, including evasion attacks, data poisoning, model extraction, and prompt injection. Each type presents unique challenges, with evasion attacks capable of misleading AI models into making incorrect predictions and data poisoning compromising the integrity of training datasets. Notably, the emergence of model extraction techniques enables attackers to replicate AI models, further exacerbating the security risks associated with adversarial tactics. The dynamic nature of these threats necessitates ongoing advancements in detection and defence mechanisms, which must adapt to the evolving landscape of cyber threats while maintaining the performance and accuracy of AI systems [[3], [4], [5].

The importance of AI security transcends mere technical concerns, as the implications of successful adversarial attacks extend to economic impacts, privacy breaches, and the disruption of essential services. Consequently, stakeholders across various industries are increasingly prioritising the development of robust defence strategies that encompass both proactive detection and effective mitigation techniques. These strategies include real time monitoring, adversarial training, input preprocessing, and continuous model optimisation, all aimed at fortifying AI systems against potential adversarial influences.[6][7][8]

Despite these efforts, significant challenges remain in effectively detecting and defending against adversarial attacks. Attackers continuously evolve their methods to bypass existing defences, leading to a perpetual cycle of adaptation between adversaries and defenders. Additionally, the complexity of models required to counteract these threats can introduce computational burdens, complicating deployment in resource constrained environments. Addressing these challenges requires ongoing research and collaboration within the AI community to establish resilient, scalable defence frameworks that can safeguard the future of AI technologies against adversarial manipulation.[9][10][11]

2 Background

Adversarial attacks on artificial intelligence (AI) systems have emerged as a critical concern in the domain of AI security. These attacks exploit unique vulnerabilities inherent to AI, potentially leading to severe consequences such as privacy violations, operational disruptions, and threats to public safety. Understanding the landscape of these attacks is vital for the development of robust defence mechanisms and the overall enhancement of AI system security.

2.1Types of Adversarial Attacks

Adversarial attacks can be classified into several categories, each posing distinct threats to AI systems.

Evasion Attacks: These attacks occur when an adversary manipulates the input data to deceive an AI model into making incorrect predictions or classifications. This can be particularly dangerous in systems like self driving cars, where misclassifications can lead to catastrophic outcomes [1][3].

Data Poisoning Attacks: In these attacks, malicious actors inject harmful data into the training dataset to corrupt the model's learning process, ultimately compromising its performance. This type of attack can

significantly undermine the reliability of AI systems that depend on accurate training data [6][3]. Model Extraction Attacks: Here, an attacker queries an AI model (often through a public API) to infer its structure and parameters, effectively creating a replica of the model without direct access to its internal workings. This type of attack can facilitate further exploits, such as data theft or evasion [4][3].

Prompt Injection Attacks: These involve embedding harmful instructions into input prompts, leading AI systems to perform unintended or harmful actions. They can bypass security measures and execute unauthorised commands, posing risks across various applications [4].

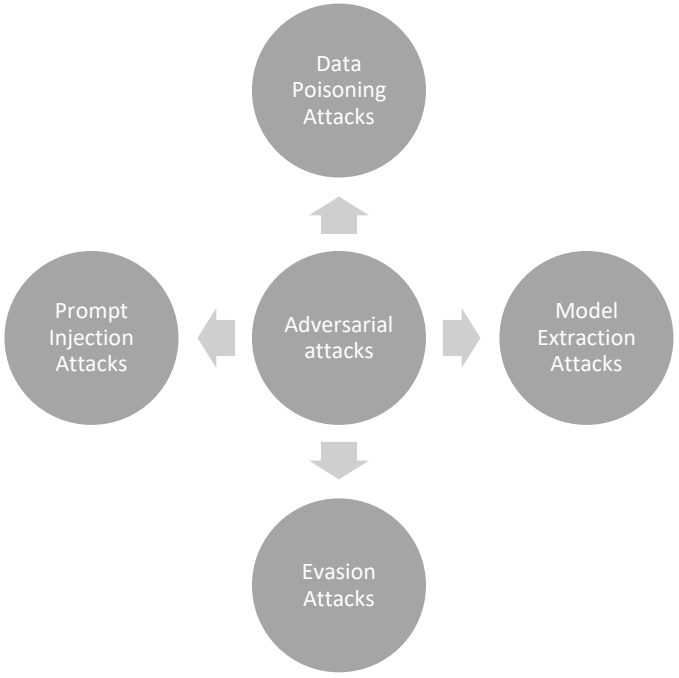


Figure 1 Classification of adversarial attacks on AI systems.

The Importance of AI Security As AI systems take on increasingly sensitive roles in society, the implications of adversarial attacks extend beyond mere technical failures. The stakes are high, as attackers may exploit these vulnerabilities for economic gain, privacy breaches, or even malicious disruptions of critical services. Consequently, enhancing AI security has become paramount not only for developers and researchers but also for stakeholders across various sectors involved in deploying and managing AI technologies [2][7]. Furthermore, a comprehensive understanding of adversarial attacks and their potential impacts on the confidentiality, integrity, and availability of AI systems is essential. This awareness is crucial for devising effective strategies to mitigate risks and protect AI systems from malicious actions [12][1].

2.2 Detection Methods

Detection methods are crucial in identifying and mitigating adversarial attacks on AI systems. These techniques serve as a primary line of defence, allowing systems to flag potential adversarial inputs before they can cause harm to the classification process.

2.2.1 Real Time Monitoring

Real time monitoring involves the continuous observation of model inputs and outputs to detect anomalies that may indicate an adversarial attack. This method utilises anomaly detection systems to flag unexpected patterns or deviations from expected behaviour in the data being processed. For example, sudden changes in the distribution of input data or unanticipated shifts in model predictions can trigger alerts or defence mechanisms, effectively enhancing the model's resilience against potential threats [5][4].

2.2.2 Statistical Anomaly Detection

Statistical anomaly detection techniques analyse the statistical properties of inputs to identify those that deviate from normal distributions. This approach includes methods such as behavioural monitoring, which tracks the model's performance over time, enabling the quick identification of unexpected changes that may signify malicious actions. By comparing the model's outputs to expected norms, statistical analysis can uncover outliers that might have been manipulated by an adversary [5][9].

2.2.3 Gradient Based Detection

Gradient based detection examines the gradients computed during model predictions to identify unusual patterns that could indicate adversarial perturbations. For instance, high gradients may suggest that the model is overly sensitive to minor changes in the input data, which can signal the presence of adversarial manipulations [8].

2.2.4 Feature Squeezing and Input Transformation

Feature squeezing techniques reduce the precision of input features, thereby minimising the impact of subtle manipulations that adversaries may employ. Additionally, input transformation methods involve rigorous filtering and validation processes for incoming data to enhance overall security [9][13].

2.2.5 Machine Learning Monitoring

Continuous monitoring of machine learning model performance is essential for effective detection of adversarial attacks. This method includes tracking metrics related to model behaviour and performance, enabling the identification of shifts that could indicate adversarial influence. Ensemble methods can also be employed, where inputs are processed through multiple models to check for consistency in predictions, thereby enhancing robustness against adversarial tactics [8][4].

2.2.6 Regular Updates and Model Optimisation

To maintain effectiveness, detection algorithms should be regularly updated and optimised. This ongoing process involves adjusting model parameters and detection strategies in response to new attack vectors and evolving data characteristics. For example, financial institutions have successfully integrated autoencoder algorithms to enhance the detection of abnormal transaction data, significantly reducing the missed detection rate of malicious transactions [14]. By implementing these detection methods, AI powered security systems can improve their resilience against adversarial attacks, ensuring more reliable performance in real world applications.

2.3 Defence Mechanisms

To defend against adversarial attacks on AI driven security systems, a variety of strategies are employed to improve model robustness and mitigate risks related to model extraction and manipulation. Techniques like adversarial training enhance resilience by incorporating adversarial examples into the training dataset. Input preprocessing transforms or filters data to eliminate adversarial perturbations before they reach the

model. Ensemble methods, which combine multiple models, add an extra layer of security by complicating attacks on individual models. While these defences have their strengths, they also present challenges such as higher computational costs and potential reductions in accuracy, highlighting the need for ongoing research to balance effectiveness and efficiency.

2.3.1 Adversarial Training and Model Hardening

Adversarial training is one of the most prominent techniques for fortifying models against adversarial attacks. This approach involves training the model using a dataset that includes adversarial examples, thus enabling it to learn to recognise and resist these manipulations effectively [1][15]. In addition, model hardening techniques focus on reinforcing models to improve their resistance to adversarial inputs by adjusting the learning process and parameters [15].

2.3.2 Input Preprocessing Techniques

Input preprocessing is a crucial preventative technique for protecting AI systems from adversarial attacks. This method modifies or cleans data before it is processed by the model, effectively sanitising the input and removing potential vulnerabilities [16][5]. JPEG Compression: This technique smooths out subtle adversarial perturbations, making it harder for attackers to manipulate inputs successfully.

Feature Squeezing: By reducing colour depth and applying spatial smoothing, this method limits the amount of detail available for adversarial exploitation.

Randomisation: Adding unpredictable noise to inputs can further obscure patterns that attackers might exploit [16][17]. These preprocessing methods enhance the model's robustness, reliability, and interpretability while complicating the attacker's ability to find successful manipulation strategies [5].

2.3.3 Model Obfuscation and Secure Aggregation

In the context of federated learning, where models are trained across multiple devices, model obfuscation serves as a defence mechanism that alters the model's output to make it more difficult for adversaries to extract sensitive information [18]. Secure aggregation protocols are also employed to protect both local updates and the overall model during collaborative training, thus maintaining privacy and security in a distributed environment [18].

2.3.4 Weight Regularisation

Weight regularisation is another vital defence strategy, which introduces penalties for high model weights during the training process. This approach encourages the model to prioritise simpler, more interpretable structures and reduces sensitivity to noise, ultimately leading to improved generalisation and robustness against adversarial examples [5]. Techniques like L1 and L2 regularisation are commonly employed to achieve these objectives.

2.4 Monitoring and Detection

Effective monitoring and detection systems play a significant role in defending against adversarial attacks. These systems can identify anomalous patterns in input data that may indicate an ongoing attack, allowing for timely interventions and adjustments to the model [15].

2.4.1 Challenges in Detection and Defence

The landscape of adversarial AI attacks presents numerous challenges for detection and defence mechanisms. As attackers continuously evolve their strategies to exploit vulnerabilities in AI models,

maintaining effective security systems becomes increasingly complex.

2.4.2 Evasion Attacks and Detection Difficulties

Evasion attacks pose significant challenges to detection systems. These attacks are designed to modify inputs in a manner that confuses AI models, making it difficult for traditional security measures to recognise malicious activity. In black box attacks, for instance, attackers have limited or no knowledge of the model's architecture and parameters, which complicates the detection of such threats. Since these attackers rely on probing the model to understand its behaviour, conventional detection methods often fall short in identifying their tactics [19][20].

2.4.3 Continuous Evolution of Attack Strategies

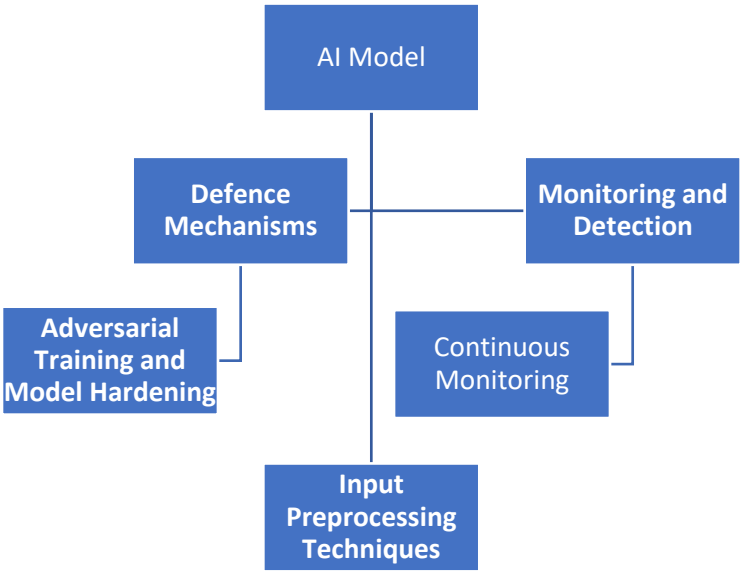
The dynamic nature of adversarial attacks adds to the complexity of defence efforts. As detection mechanisms improve, attackers adapt by developing new techniques to evade these defences, creating a perpetual cat and mouse game between adversaries and defenders [5][21]. This constant evolution necessitates that organisations remain vigilant and proactive in monitoring and updating their defence strategies.

2.4.4 Resource Constraints and Model Complexity

Defending against a wide array of adversarial techniques often requires increasing the complexity of the model. However, this heightened complexity can lead to greater computational burdens, making it challenging to deploy these models in resource constrained environments [10][14]. Furthermore, as defences are enhanced to resist adversarial attacks, there is often a trade off with the model's standard accuracy, potentially reducing its performance on non adversarial inputs [10][22].

2.5 Importance of Continuous Monitoring

To effectively mitigate the risks associated with adversarial attacks, organisations must implement robust monitoring and detection systems. This involves systematic observation of model behaviour, analysis of input output patterns, and evaluation of outliers to quickly respond to suspicious activity [5][1]. Additionally, continuous updating of models and detection systems is essential to adapt to new attack strategies and improve resilience against evolving threats [21][11].



3 Case Studies

3.1 Overview of Red Teaming in AI Security

Case studies highlight how businesses across various industries have effectively utilised red teaming to enhance their security readiness against adversarial attacks. Red teaming involves ethical hackers simulating real world cyberattacks to evaluate an organisation's capability to detect, respond to, and recover from sophisticated threats [23]. This approach transcends traditional penetration testing by mimicking malicious tactics in an unrestricted manner, thus providing a comprehensive understanding of an organisation's security vulnerabilities [23][24].

3.2 Industry Applications

3.2.1 Healthcare

In the healthcare sector, adversarial attacks pose significant risks, particularly to the accuracy of medical diagnoses. Case studies illustrate instances where manipulated medical images have led to incorrect treatment recommendations, jeopardising patient safety [8]. Organisations employing red teaming techniques have improved their systems' resilience against such vulnerabilities, thereby protecting sensitive patient data and ensuring reliable healthcare delivery.

3.2.2 Finance

The finance industry faces threats from adversarial attacks targeting fraud detection systems. Successful exploitation of these systems could facilitate substantial financial crimes, resulting in significant monetary losses [8]. By engaging in red teaming exercises, financial institutions have refined their defensive strategies, enhancing their capability to identify and mitigate fraudulent activities.

3.2.3 Transportation

Adversarial attacks on autonomous vehicles pose a significant risk by manipulating the vehicle's perception of its surroundings. This distortion can result in dangerous situations, potentially leading to accidents with serious, even fatal, outcomes [8]. Automotive companies have utilised red teaming to simulate various attack

scenarios. This approach allows them to strengthen their defences against potential risks, ultimately enhancing the overall safety of their autonomous systems.

3.3 Effective Strategies for Mitigation

Organisations that have successfully leveraged red teaming to bolster their defences have employed a range of strategies. These include continuous monitoring of model behaviours, implementation of anomaly detection algorithms, and regular assessments of system performance against baseline metrics [25][5]. By proactively identifying unusual model behaviours, organisations can detect adversarial attacks early and mitigate their impacts effectively [24][25]. Additionally, integrating adversarial training and robust feature extraction into the development of AI systems has proven beneficial. These defensive mechanisms enhance model robustness, making it more challenging for adversaries to exploit vulnerabilities through data poisoning and other tactics [11][15]. The dynamic interplay between attackers and defenders highlights the necessity of evolving security measures, which are informed by ongoing case studies and real world applications [26].

4 Future Directions

As adversarial attacks on AI systems become increasingly sophisticated, future research must focus on developing unified and scalable defence frameworks that can adapt to the evolving threat landscape [10][18]. The advancement of defence mechanisms is critically important, particularly in addressing the specific vulnerabilities present in diverse computing environments while ensuring ease of deployment. For instance, in cloud computing scenarios, defence strategies should evolve beyond mere query monitoring to include sophisticated adaptive response mechanisms. These advanced systems must be capable of identifying and countering complex data extraction attempts (such as those involving deep packet inspection or subtle API manipulation) without compromising overall service quality for users. Achieving this requires a delicate balance between robust security measures and an optimal user experience, ensuring that protective interventions do not detrimentally affect performance or accessibility [18].

4.1 Challenges in Defence Mechanisms

One of the primary challenges is striking the right balance between security and performance. It is essential to achieve low latency and high throughput while enforcing robust security measures. In edge computing environments, the challenge is to devise lightweight yet effective defence strategies that function optimally within stringent resource limitations. This necessitates the exploration of hardware assisted security features and efficient encryption techniques that do not compromise device performance or battery life [18].

4.2 Collaborative and Adaptive Strategies

To effectively combat various adversarial threats in machine learning, it is vital to adopt a multifaceted approach that incorporates diverse defensive techniques and proactive monitoring. Collaboration among researchers, practitioners, and industry stakeholders is essential as new attack types, such as evasion, data poisoning, Byzantine, and model extraction, continue to emerge. To ensure the robustness of machine learning systems, defence strategies must adapt and evolve, utilising methods such as adversarial learning, monitoring, defensive distillation, and differential privacy [11][27]. The need for adaptive defence mechanisms is paramount; static strategies often fail against new types of attacks, exposing significant vulnerabilities if attackers understand the defences in place.[27]

4.3 Ethical Considerations in AI Defence

Furthermore, the integration of ethical considerations into the development of AI defence strategies is becoming increasingly important. Ethical AI promotes transparency in decision making processes, aiming to create trustworthy systems that respect user privacy and avoid biases.[28][29] Future directions in adversarial defence should also encompass these ethical dimensions, ensuring that defence strategies do not inadvertently compromise the rights of individuals or reinforce existing inequalities.

5.0 Conclusion

As AI powered systems become increasingly integral to critical domains such as healthcare, finance, and transportation, the risks posed by adversarial attacks have grown in scale, complexity, and consequence. This study has examined the diverse spectrum of adversarial threats including evasion attacks, data poisoning, model extraction, and prompt injection and reviewed a comprehensive set of detection and defense strategies ranging from gradient based detection and statistical anomaly analysis to adversarial training, input preprocessing, and model obfuscation.

The findings underscore that no single approach is sufficient; rather, a layered and adaptive defense strategy coupled with continuous monitoring and system updates is essential to secure AI models against evolving threats. The integration of techniques such as red teaming, explainable AI, and ensemble methods further enhances system resilience, but these must be deployed with awareness of resource constraints, ethical implications, and the risk of overfitting defenses to past attacks.

Despite considerable progress, the landscape remains dynamic, with adversaries constantly developing new evasion techniques. Future defenses must therefore prioritize scalability, adaptability, and transparency, ensuring that AI systems can respond not only to known threats but also to emergent attack vectors. Collaborative research, ethical governance, and the implementation of secure AI development lifecycles will be key to building trustworthy, resilient AI infrastructures in the face of adversarial manipulation.

Declaration of Conflicting Interests

The authors declare no potential conflicts of interest with respect to the research, authorship and publication of this article.

Funding

The author received no financial support for the research, authorship and publication of this article.

References

[1] Number Analytics. (2025). Adversarial attacks in AI: A comprehensive guide.
[2] Lumenova AI. (2025). Overcoming adversarial attacks in machine learning.
[3] Nightfall AI. (2025). Adversarial attacks and perturbations: The essential guide.
[4] Kiribuchi, N., Zenitani, K., & Semitsu, T. (2025). Securing AI systems: A guide to known attacks and impacts (Vol. 1).
[5] Practical DevSecOps. (2025). AI security system attacks in 2025.
[6] Rapid Innovation. (2025). Ethical AI development: Principles and best practices.
[7] BytePlus. (2025). The origins of adversarial machine learning.

[8] Data Science Salon. (2025). Defending against adversarial attacks in the era of generative AI.
[9] Lakera AI. (2025a). Outsmarting the smart: Intro to adversarial machine learning.
[10] Founding Minds. (2025). Overcoming adversarial threats and strengthening AI defense mechanisms.
[11] Founding Minds AI. (2025). Overcoming adversarial attacks in AI security.
[12] NFactor Technologies. (2025). Part 1: Navigating the threat of evasion attacks in AI.
[13] Wiz.io. (2025). The threat of adversarial AI.
[14] FPT Software. (2025). What is adversarial AI? Uncovering the risks and strategies to mitigate.
[15] Palo Alto Networks. (2025). What is adversarial AI in machine learning?
[16] Neptune.ai. (2025). Adversarial machine learning: Defense strategies.
[17] ISO. (2025). Building a responsible AI: How to manage the AI ethics debate.
[18] Transcend.io. (2025). Key principles for ethical AI development.
[19] National Institute of Standards and Technology (NIST). (2024). NIST identifies types of cyberattacks that manipulate behavior of AI systems.
[20] CCDCOE. (2019). Addressing adversarial attacks against security systems based on AI.
[21] Lakera AI. (2025b). Prompt injection and the rise of prompt attacks: All you need to know.
[22] OWASP. (2025). LLM01: Prompt injection. OWASP GenAI Security Project.
[23] IBM. (2025). What is a prompt injection attack?
[24] Wikipedia. (2025, July). Prompt injection.
[25] Google Security Blog. (2025, June). Mitigating prompt injection attacks with a layered defense strategy.
[26] Wired. (2023, September). Here come the AI worms.
[27] Wired. (2025, January). DeepSeek's R1 AI model fails jailbreak tests.
[28] McHugh, J., Šekrst, K., & Cefalu, J. (2025). Prompt Injection 2.0: Hybrid AI threats. arXiv preprint arXiv:2507.13169.
[29] Chen, S., Zharmagambetov, A., Wagner, D., Guo, C., & Fair, M. (2025). : A secure foundation LLM against prompt injection attacks.