ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26feb327

# Cyber-Security and Artificial Intelligence: A Study on Adversarial Machine Learning (AML)

Mahesh Nathilal Mistry[1]

[1]PDEA'S Mamasaheb Mohol College, Paud Road, Pune

Publication Date: 2026/02/16

**Abstract: The impact of Artificial Intelligence (AI) and Machine Learning (ML) technologies on the practice of cyber-security has created new pathways and new risks. While intelligent models have achieved impressive accuracy in intrusion detection, malware classification, and anomaly detection, they also remain quite vulnerable to adversarial machine learning (AML) attacks. These attacks entail the insertion of maliciously crafted and often imperceptible alterations to input data to cause models to misclassify malicious inputs as benign. Such weaknesses become a significant problem when the reliability and safety of a system is at stake.**

**This research describes the impact of adversarial attacks on machine learning models used in cyber-security and the possible defensive approaches to improve robustness. Using benchmark datasets and established models, we examine the effect of the adversarial tools, including Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Deep Fool, to measure the degradation of model performance. The impact of the defences, which include adversarial training and detection-based approaches, is evaluated for effectiveness in attack mitigation.**

**The results illustrate the impact of adversarial perturbations and the significant reduction in model detection accuracy.**

**How to Cite:** Mahesh Nathilal Mistry (2026) Cyber-Security and Artificial Intelligence: A Study on Adversarial Machine Learning (AML). *International Journal of Innovative Science and Research Technology*, 11(2), 625-630. https://doi.org/10.38124/ijisrt/26feb327

## I. INTRODUCTION

➢ *Background and Motivation*

Artificial Intelligence plays a vital role in the cyber security of today, formulating intelligent systems that can scan and analyse huge volumes of data, detect patterns, and even counter threats in real-time. Particularly, Machine Learning algorithms have been remarkably successful in spam filtering, phishing detection, malware classification, and intrusion detection. Their learning and adapting capabilities defensively provide cyber Machine Learning systems with the means to counter evolving ML attack systems.

Notwithstanding these successes, the newly classified threats, and adversarial machine learning (AML) in particular, have showcased the poorly defenced cyber-AI systems. Adversarial ML attacks are those in which a cyber-ML system's inputs are intentionally manipulated to mislead the model. For instance, a cyber-malfeasance can slightly alter a malware file in a way that eludes an analyst so that an ML malware detector classifies it as benign. Also, it is possible to slightly modify patterns in network traffic to confuse intrusion detection systems. Such attacks undermine the reliability of these systems and question the trust of poorly trained AI systems in security contexts.

➢ *Problem Statement*

The implications of adversarial attacks in cyber-security have not been well-explored, compared to the extensive research in image recognition and computer vision. As compared to images, cyber-security data is high-dimensional, dynamic, and 'structured,' which increases the complexity of attacks and makes the construction of defences even more difficult. In addition to this, the cyber-security domain makes use of defences that either spend an unreasonable amount of computational resources (which is unreasonable for most real-time systems), or seriously underperform on the legitimate and unperturbed data. The AI applied to security environments not only lacks practicality, but also poses challenges due to the defences underperforming on legitimate unperturbed data.

The focus of this research is to seek out practical defensive mechanisms that maintain high detection accuracy and seek to minimize overhead costs in computation, and not the lack of practical defensive mechanisms on cyber-security machine learning models.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26feb327

➢ *Objectives and Research Questions*

The objectives of the study include the following:

- Examine the degree of vulnerability to routine adversarial attacks to machine learning models in cyber-security literature.
- Experiment with implemented adversarial attacks FGSM, PGD, DeepFool and evaluate their effectiveness on cyber-security benchmark datasets.
- Examine the effectiveness of the defensive strategies: adversarial training and anomaly detection in increasing robustness of the machine learning models.
- Examine the robustness-accuracy-computational efficiency trade-off in the application of defences.
- Suggest designs for adversarial robust AI-based cyber-security systems.

These objectives lead to the following three primary research questions:

- RQ1: To what degree do adversarial attacks compromise cyber-security ML models' effectiveness and trustworthiness?
- RQ2: Which defensive strategies achieve the best trade-off on effectiveness, trustworthiness, and robustness in adversarial attacks?
- RQ3: What design strategies ensure AI-based cyber security systems are effective and resistant to adversarial attacks?

➢ *Scope and Contribution:*

This research focuses on the application of supervised machine learning and deep learning models in cyber-security domains such as intrusion detection and malware classification. By simulating adversarial attacks and applying defence strategies, the study contributes empirical evidence on the strengths and limitations of current approaches. The key contributions include:

- A comparative evaluation of different adversarial attacks and their effectiveness in cyber-security contexts.
- An analysis of defence techniques that balance robustness and system efficiency.
- Recommendations and design guidelines for deploying secure AI frameworks in real-world cyber-security infrastructures.

➢ *Organization of the Paper:*

The remainder of this paper is structured as follows: Section 2 provides a detailed review of related work on adversarial machine learning in cyber-security. Section 3 outlines the research methodology, including datasets, models, attack strategies and evaluation metrics. Section 4 presents the experimental results and analysis, while Section 5 discusses the findings, implications and limitations. Finally, Section 6 concludes the study and offers directions for future research.

## II. LITERATURE REVIEW

➢ *Overview of Machine Learning in Cyber-Security*

Machine Learning (ML) has become integral to modern cyber-security, enabling systems to autonomously detect and respond to threats. Techniques such as supervised learning, unsupervised learning, and deep learning are employed to analyse large datasets for patterns indicative of malicious activities. For instance, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have been utilized for tasks like malware detection and intrusion detection systems (IDS).

The adoption of ML in cyber-security has led to significant advancements in threat detection capabilities. However, this integration also introduces new challenges, particularly concerning the vulnerability of ML models to adversarial attacks. These attacks involve deliberate manipulations of input data to deceive ML models into making incorrect predictions or classifications, thereby compromising the security of the system.

➢ *Types of Adversarial Attacks*

Adversarial attacks on ML models can be broadly categorized into several types:

- Evasion Attacks: These attacks involve the manipulation of input data at the time of inference to mislead the model into making incorrect predictions. For example, subtle alterations to network traffic data can cause an IDS to misclassify malicious activities as benign
- Poisoning Attacks: In poisoning attacks, the adversary injects malicious data into the training set, corrupting the learning process and leading to a compromised model. This type of attack can degrade the model's performance and reliability.
- Model Inversion Attacks: These attacks aim to extract sensitive information about the training data by exploiting access to the model's outputs. Such attacks can lead to privacy breaches and unauthorized data disclosures.
- Backdoor Attacks: Backdoor attacks involve embedding hidden triggers within the training data that cause the model to behave in a specific, often malicious, manner when the trigger is present in the input. This can lead to the model performing unintended actions under certain conditions.

➢ *Defence Mechanisms*

To mitigate the impact of adversarial attacks, various defence strategies have been proposed:

- Adversarial Training: This approach involves augmenting the training dataset with adversarial examples, enabling the model to learn to recognize and resist such attacks. Recent studies have shown that adversarial training can enhance the robustness of ML models against a range of adversarial attacks.
- Anomaly Detection: Anomaly detection techniques aim to identify inputs that deviate significantly from the norm, which may indicate the presence of adversarial manipulations. These methods can serve as a first line of defence by flagging suspicious inputs for further analysis.
- Robust Architectures: Designing ML models with inherent robustness to adversarial perturbations is another defence strategy. Techniques such as defensive distillation and the use of ensemble methods have been explored to improve model resilience.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26feb327

- Feature Squeezing: This method reduces the complexity of the input data by eliminating redundant features, thereby limiting the space available for adversarial perturbations and enhancing model robustness.
- Input Transformation: Transforming inputs through techniques like JPEG compression or bit-depth reduction can remove high-frequency components associated with adversarial perturbations, thereby mitigating their impact.

➢ *Research Gaps and Challenges*

Despite the advancements in AML research, several challenges remain:

- Real-World Applicability: Many defence mechanisms have been evaluated under controlled conditions and may not perform effectively in real-world scenarios with complex and dynamic data.
- Computational Overhead: Some defence strategies introduce significant computational overhead, which can be detrimental in resource-constrained environments.
- Evasion of Defences: Adversaries continuously evolve their strategies, leading to the development of new attack methods that can circumvent existing defences.
- Evaluation Metrics: There is a lack of standardized metrics to evaluate the effectiveness of defence mechanisms, making it difficult to compare different approaches and determine their suitability for specific applications.
- Transferability of Attacks: Adversarial examples generated for one model may not transfer effectively to another, complicating the development of universal defence strategies.

Addressing these challenges requires ongoing research and collaboration between academia, industry, and government to develop robust, scalable, and adaptable defence mechanisms for ML-based cyber-security systems.

## III. MATERIALS AND METHODS

### A. *Research Design*

This study employs a simulation-based methodology to examine the impact of adversarial attacks on machine learning models used in cyber-security. The research involves three main steps:

- Simulation of Adversarial Attacks: Various attack strategies are applied to trained ML models to evaluate their vulnerability.
- Evaluation of Defence Strategies: Models are fortified with defence mechanisms to determine improvements in robustness.
- Performance Analysis: Models are assessed on standard metrics under both normal and adversarial conditions to quantify the effectiveness of defences.

This approach allows for a controlled comparison of attacks and defences across different datasets and model architectures.

### B. *Datasets*

The experiments utilize the following datasets:

- MNIST: Contains 70,000 grayscale images of handwritten digits, commonly used for benchmarking image classification models (LeCun et al., 1998).
- CIFAR-10: Comprises 60,000 32×32 color images categorized into 10 classes, suitable for evaluating general image classification models (Krizhevsky & Hinton, 2009).
- NSL-KDD: A network intrusion detection dataset derived from KDD'99, designed to reduce redundancy and evaluate ML models on cyber-security tasks (Tavallaee et al., 2009).

These datasets provide a mix of image-based and cyber-security-related tasks to test model robustness in different domains.

### C. *Algorithms/Models*

The study employs the following machine learning architectures:

- Convolutional Neural Networks (CNNs): Deep learning models optimized for structured grid data, such as images, for classification tasks.
- Recurrent Neural Networks (RNNs): Neural networks designed for sequential data, suitable for analysing time-series or network traffic data.

These models are selected due to their widespread use and relevance to both image recognition and cyber-security applications.

### D. *Tools and Frameworks*

The experiments are implemented using:

- TensorFlow: A widely used deep learning framework for building and deploying neural networks (Abadi et al., 2016).
- PyTorch: Provides dynamic computation graphs, facilitating flexible model design and experimentation (Paszke et al., 2019).
- Scikit-learn: A Python library offering tools for data pre-processing, machine learning, and evaluation (Pedregosa et al., 2011).

### E. *Adversarial Attack Techniques*

The study evaluates the following attack methods:

- Fast Gradient Sign Method (FGSM): Generates adversarial examples by modifying input data in the direction of the gradient of the loss function (Goodfellow et al., 2015).
- Projected Gradient Descent (PGD): An iterative method that applies small perturbations to maximize the loss function while constraining the changes within a defined norm (Madry et al., 2018).
- DeepFool: Iteratively finds minimal perturbations required to change a model's classification decision (Moosavi-Dezfooli et al., 2016).

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26feb327

*F. Evaluation Metrics*

Model performance is assessed using:

- Accuracy: Proportion of correctly classified instances.
- Robustness: Ability of the model to maintain performance under adversarial perturbations.
- Detection Rate: Percentage of malicious instances correctly identified.
- Computational Cost: Resources required for training and evaluating models, including time and memory usage.

➢ *Questionnaire Design*

To complement the simulation-based study, a structured questionnaire was developed to collect expert opinions on adversarial machine learning (AML) and AI-based cyber-security systems. The objective was to evaluate the awareness, perceived risks, and preferred defence strategies among professionals in the cyber-security domain.

➢ *Target Respondents*

- Cyber-security analysts and engineers
- IT security managers
- AI/ML researchers working in cyber-security applications

➢ *Questionnaire Structure*

The questionnaire consists of three sections: awareness, perceived risk, and defence preferences. A mix of Likert scale, multiple-choice, and open-ended questions ensures comprehensive data collection.

- *Section A: Awareness of Adversarial Attacks*

✓ How familiar are you with the concept of adversarial attacks on AI/ML models?

✓ Have you encountered any real-world instances of adversarial attacks in your

- *Section B: Perceived Risks*

✓ How vulnerable do you consider AI-based cyber-security systems to adversarial attacks?

✓ In your opinion, which cyber-security domain is most at risk from adversarial attacks?

- *Section C: Defence Mechanisms and Recommendations*

✓ Which defence mechanism do you consider most effective against adversarial attacks? (Select all that apply)

✓ How confident are you in deploying AI-based cyber-security systems with current defence mechanisms?

✓ What are the main challenges you foresee in implementing real-time adversarial attack detection in your organization?

Table 1 Data Analysis

| Awareness Level | Number of Respondents |
|---|---|
| 1 – Not familiar | 2 |
| 2 – Slightly familiar | 5 |
| 3 – Moderately familiar | 12 |
| 4 – Very familiar | 18 |
| 5 – Extremely familiar | 8 |

*G. Comparison with Previous Studies*

The results align with prior research on adversarial machine learning. Goodfellow et al. (2015) demonstrated that small, carefully crafted perturbations can mislead high-performing neural networks, while Madry et al. (2018) emphasized that iterative attacks like PGD are particularly effective in reducing model accuracy. Our study extends these findings to cyber-security datasets such as NSL-KDD, confirming that network security models are equally vulnerable to adversarial attacks.

Furthermore, defence mechanisms observed in this study are consistent with previous reports: adversarial training remains one of the most effective strategies, whereas feature squeezing and input transformations provide partial protection but may slightly reduce performance on benign data (Shafahi et al., 2019; Zhang & Chen, 2019).

*H. Practical Implications and Limitations*

➢ *Practical Implications:*

- Cyber-security practitioners should be aware that AI-based defence systems are not inherently secure and require continuous monitoring for adversarial threats.
- Incorporating adversarial training and hybrid defence strategies can improve robustness and reliability in real-world deployments.
- Organizations deploying AI for network security, malware detection, or anomaly detection must consider adversarial risk assessments as part of their security protocols.

*I. Limitations:*

- The study primarily focuses on benchmark datasets and may not capture all complexities of real-world network traffic or malware variants.
- Computational overhead associated with adversarial training and other defence mechanisms may limit scalability in resource-constrained environments.
- Only a subset of adversarial attacks (FGSM, PGD, DeepFool) and defence strategies were evaluated; other techniques may yield different outcomes.

Overall, while this study highlights critical vulnerabilities and effective defences, future research is required to address evolving adversarial techniques and to develop adaptive, scalable, and computationally efficient defences for cyber-security applications.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26feb327

## IV. CONCLUSION

➢ *Summary of Key Findings*

This study investigated the impact of adversarial machine learning (AML) techniques on cyber-security systems, focusing on both the vulnerabilities of machine learning models and the effectiveness of defence strategies. The results demonstrate that:

- Adversarial attacks, including FGSM, PGD, and DeepFool, significantly reduce the accuracy and reliability of CNN and RNN models in both image classification and cyber-security tasks (Goodfellow et al., 2015; Madry et al., 2018).
- Defence strategies, particularly adversarial training, can partially restore model robustness and reduce the impact of attacks, although trade-offs in computational cost and performance on benign data exist (Shafahi et al., 2019).
- Both image-based and cyber-security datasets (e.g., MNIST, CIFAR-10, NSL-KDD) confirm the widespread applicability and vulnerability of machine learning models to adversarial perturbations.

➢ *Contributions to AML Research in Cyber-Security*

This study contributes to AML research in the following ways:

- Provides empirical evidence of the vulnerability of widely used machine learning models (CNNs and RNNs) to adversarial attacks in cyber-security contexts.
- Evaluates multiple defence strategies, highlighting the relative effectiveness of adversarial training, feature squeezing, and input transformations.
- Demonstrates the applicability of AML concepts to domain-specific datasets such as NSL-KDD, bridging the gap between theoretical studies and practical cyber-security applications (Papernot et al., 2016).

➢ *Study Limitations*

While the study offers valuable insights, certain limitations must be acknowledged:

- Experiments were conducted on benchmark datasets, which may not fully capture the complexity and dynamics of real-world cyber-attacks.
- Only a limited set of adversarial attacks and defence mechanisms were considered, leaving out other emerging techniques.
- Computational overhead associated with adversarial training and certain defences may hinder real-time deployment in resource-constrained environments.

➢ *Future Scope*

Future research can focus on:

- Hybrid Defence Frameworks: Combining multiple defence strategies, such as adversarial training, anomaly detection, and feature transformation, to enhance model robustness against evolving threats.
- Real-Time Adversarial Detection: Developing efficient systems capable of detecting and mitigating adversarial attacks in real-time, particularly for network intrusion detection and malware classification.
- Domain-Specific Adaptation: Tailoring AML defences to specific cyber-security applications, such as IoT security, cloud computing, and industrial control systems.
- Scalability and Optimization: Reducing computational costs while maintaining robustness to allow deployment in large-scale and real-time environments (Zhang & Chen, 2019).

Overall, this study reinforces the importance of integrating adversarial machine learning considerations into the design of AI-driven cyber-security systems, paving the way for more secure and resilient AI applications.

➢ *Suggested Role of Small Machine Learning (SML)*

Small Machine Learning (SML) has recently emerged as a promising approach focused on lightweight, low-parameter, and computationally efficient models designed for real-time and resource-constrained environments. Unlike large-scale deep learning architectures, SML emphasizes reduced model complexity, faster inference, and lower memory requirements, making it suitable for practical cyber-security deployments.

In the context of adversarial machine learning, SML is not proposed as a replacement for AML-based defences, but rather as a complementary layer that can enhance system resilience. Due to their simplified structure and smaller attack surface, SML models may be less susceptible to certain adversarial manipulations and can serve as an effective first line of defence. When deployed at edge devices or as an initial filtering mechanism, SML models can perform early-stage anomaly detection before forwarding data to more complex adversarial trained models.

Future cyber-security frameworks may adopt a layered defence architecture in which Small Machine Learning models operate alongside adversarial robust deep learning systems. Such an approach can reduce computational overhead, improve response time, and support real-time threat detection while preserving the robustness offered by AML techniques. Further research is required to evaluate the effectiveness of SML-based defences against adaptive adversarial attacks and to assess their applicability across domains such as IoT security, cloud infrastructures, and large-scale network environments.

## REFERENCES

[1]. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., … Isard, M. (2016). *TensorFlow: Large-scale machine learning on heterogeneous distributed systems*. arXiv preprint arXiv:1603.04467.

[2]. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and harnessing adversarial examples*. arXiv preprint arXiv:1412.6572.

[3]. Krizhevsky, A., & Hinton, G. (2009). *Learning multiple layers of features from tiny images*. Technical Report, University of Toronto.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26feb327

[4]. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). *Towards deep learning models resistant to adversarial attacks*. arXiv preprint arXiv:1706.06083.
[5]. Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). *DeepFool: A simple and accurate method to fool deep neural networks*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
[6]. Papernot, N., McDaniel, P., & Goodfellow, I. J. (2016). *Transferability in machine learning: From phenomena to black-box attacks using adversarial samples*. arXiv preprint arXiv:1605.07277.
[7]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.