www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 10, October 2025, pp : 93-95

e-ISSN : 2583-1062
Impact Factor : 7.001

ADVERSARIAL ATTACKS IN CYBERSECURITY: HOW AI MODELS CAN BE FOOLED AND METHODS TO MAKE THEM ROBUST

Dr. I. Carol[1], Abirami M[2]

[1]Assistant Professor, Department of Information Technology St. Joseph's College, Trichy, India.

[2]PG Student, Department of Information Technology, St. Joseph's College, Trichy, India.

DOI: https://www.doi.org/10.58257/IJPREMS44069

ABSTRACT

Artificial Intelligence (AI) is widely used in cybersecurity for intrusion detection, malware classification, and phishing prevention. However, these models are vulnerable to adversarial attacks, where small changes in input data can mislead the system. This paper studies common adversarial attack techniques, such as FGSM, PGD, and CW, and evaluates defense methods including adversarial training and preprocessing. Experiments show that attacks significantly reduce model accuracy, while defenses improve robustness but do not fully eliminate risks. The work highlights the need for stronger, more reliable AI models in cybersecurity applications.

Keywords: Adversarial Attacks, Cybersecurity, Machine Learning, Deep Learning, Robustness, Intrusion Detection.

1. INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) have become important tools in cybersecurity, supporting tasks such as network intrusion detection, malware classification, and phishing prevention. These systems provide faster and more accurate detection compared to traditional rule-based methods.

However, ML models are not fully secure. They are vulnerable to adversarial perturbations—small, carefully designed changes in input data that can cause a model to misclassify. Such weaknesses can be exploited by attackers in real-world scenarios, leading to serious risks like evasion of intrusion detection systems or poisoning of training data.

This paper addresses these challenges with the following objectives:

1. To survey common adversarial attack techniques.
2. To demonstrate their impact on cybersecurity ML models.
3. To evaluate and compare defense methods for improving robustness.

By achieving these objectives, this work aims to provide insights into building more secure and trustworthy AI-based cybersecurity systems.

2. LITERATURE REVIEW / RELATED WORK

Machine learning has been widely applied in cybersecurity, with models used for intrusion detection, malware classification, and spam or phishing detection. These systems have shown high accuracy and efficiency compared to traditional rule-based approaches.

Recent studies, however, reveal that ML-based cybersecurity systems are vulnerable to adversarial attacks. Prior studies demonstrate that even minor modifications in input data can compromise machine learning models in cybersecurity. For instance, IDS models can be evaded through subtle alterations in network traffic patterns [Alotaibi & Rassam, 2023], malware classifiers may be deceived into recognizing harmful files as harmless [Li, 2024], and phishing detection systems can be bypassed by modifying URL or email content [Adversarial Challenges in NIDS, 2024]. Despite these findings, there are notable gaps in the literature. Many studies focus on demonstrating attacks but provide limited evaluation of defense strategies. While proposed defense strategies have shown effectiveness in experimental setups, their scalability and reliability in real-world, dynamic cybersecurity environments remain limited [Tafreshian & Zhang, 2025]. Furthermore, very few works combine adversarial robustness with explainable AI, leaving a gap in developing models that are both interpretable and secure.

This motivates the need for further research on practical adversarial defenses tailored to cybersecurity applications.

3. METHODOLOGY

This study evaluates the impact of adversarial attacks on machine learning models for cybersecurity and tests defense strategies to improve robustness.

Dataset: For experimentation, this study employs the NSL-KDD dataset, which provides labeled examples of both normal and malicious traffic. Due to its balanced design and improvements over KDD'99, it has become a standard benchmark for intrusion detection research [Tavallaee et al., 2009].

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 10, October 2025, pp : 93-95

e-ISSN : 2583-1062
Impact Factor : 7.001

Model: A baseline Random Forest classifier is trained on the dataset. In addition, a simple deep learning model (Convolutional Neural Network) is implemented for comparison.

Adversarial Attacks Tested:

- **FGSM (Fast Gradient Sign Method):** Generates adversarial examples by adding small perturbations based on the gradient of the loss function.
- **PGD (Projected Gradient Descent):** Iteratively refines adversarial samples to create stronger attacks.
- **CW (Carlini & Wagner):** A powerful optimization-based attack that minimizes perturbation while forcing misclassification.

Defense Mechanisms:

- **Adversarial Training:** Incorporating adversarial samples during training to improve robustness.
- **Defensive Distillation:** Using softened labels to make models less sensitive to small perturbations.
- **Feature Squeezing/Preprocessing:** Reducing input complexity (e.g., rounding values, noise reduction) to filter adversarial noise.

Tools: The experiments are implemented in Python using Scikit-learn, TensorFlow, and the Adversarial Robustness Toolbox (ART).

Workflow:

1. Train baseline models on clean data.
2. Generate adversarial samples using FGSM, PGD, and CW attacks.
3. Measure model performance before and after attack.
4. Apply defense methods and evaluate improvements.

4. RESULTS AND DISCUSSION

The experiments were carried out on a system with an Intel i5 processor, 8GB RAM, and Python 3.10. The models were implemented using Scikit-learn, TensorFlow, and the Adversarial Robustness Toolbox (ART).

Performance Metrics: Model evaluation was based on Accuracy, Precision, Recall, and F1-score.

Baseline Performance: On the NSL-KDD dataset, the Random Forest classifier achieved high accuracy on clean data, while the CNN model showed slightly lower performance but stronger generalization.

Impact of Adversarial Attacks: After applying FGSM, PGD, and CW attacks, both models showed significant drops in accuracy. Even small perturbations were enough to mislead the classifiers, highlighting the vulnerability of ML models in cybersecurity.

Defense Results: When defense strategies were applied, performance improved:

- **Adversarial Training** provided the best results, restoring most of the lost accuracy.
- **Defensive Distillation** reduced sensitivity to small perturbations but at the cost of higher training time.
- **Feature Squeezing** was simple and effective against weaker attacks but less reliable for strong ones like CW.

Key Findings:

1. Experimental findings show that **the accuracy of ML-based intrusion detection systems may drop by over 30% under strong adversarial perturbations**, consistent with prior studies [Penmetsa et al., 2025; Goodfellow et al., 2015].
2. Defense mechanisms improve robustness, but none completely eliminate vulnerabilities.
3. Adversarial training is the most effective, while lightweight methods like feature squeezing may be useful in resource-limited systems.

5. CONTRIBUTION OF THE WORK

This paper makes the following contributions:

1. **Comparative Study of Adversarial Attacks:** A detailed analysis of popular adversarial techniques (FGSM, PGD, and CW) on machine learning models applied to cybersecurity tasks.
2. **Evaluation of Defense Mechanisms:** Systematic testing of adversarial training, defensive distillation, and feature squeezing on the same dataset, highlighting their strengths and weaknesses.
3. **Practical Insights:** Demonstrates the real impact of adversarial perturbations on intrusion detection models, offering guidance for deploying ML securely in cybersecurity systems.

4. **Future-Oriented Approach:** Suggests combining robustness techniques with explainable AI to improve both the reliability and interpretability of security models.

6. CHALLENGES AND FUTURE WORK

This work highlights several challenges in applying adversarial robustness to cybersecurity systems. A major limitation is the **computational cost** of defenses such as adversarial training, which require extensive resources and longer training times. Another concern is **dataset bias**, as benchmark datasets like NSL-KDD may not fully represent real-world network traffic.

For future research, three directions are proposed:

1. **Real-Time Testing:** Extending experiments to live network traffic for more realistic evaluation.
2. **Hybrid Defense Models:** Combining AI with traditional rule-based security methods to enhance resilience.
3. **Federated Learning with Robustness:** Developing distributed, privacy-preserving models that are resistant to adversarial attacks.

These directions will help move adversarial defense research closer to practical deployment in real-world cybersecurity environments.

7. CONCLUSION

Adversarial attacks present a serious threat to AI applications in cybersecurity, as even minor perturbations can significantly reduce model accuracy. The study shows that machine learning models, while effective under normal conditions, are highly vulnerable without appropriate defense mechanisms. Defense strategies including adversarial training, defensive distillation, and feature preprocessing, do enhance resilience, but current evidence suggests that no single method can ensure complete protection [Springer, 2025].

For real-world deployment, stronger adversarially trained models are needed, along with integration of explainable AI to ensure transparency and trust. Building resilient, interpretable, and scalable defense systems remains a key challenge for the future of secure AI in cybersecurity.

8. REFERENCES

[1] A. Alotaibi and M. A. Rassam, “Adversarial Machine Learning Attacks against Intrusion Detection Systems: A Survey on Strategies and Defense,” Future Internet, vol. 15, no. 2, p. 62, Jan. 2023. ResearchGate

[2] L. Li, “Comprehensive Survey on Adversarial Examples in Cybersecurity: Impacts, Challenges, and Mitigation Strategies,” arXiv, Dec. 2024. arXiv

[3] “Adversarial Challenges in Network Intrusion Detection Systems,” arXiv, Sept. 2024. arXiv+1

[4] B. Tafreshian and S. Zhang, “A Defensive Framework Against Adversarial Attacks on Machine Learning-Based Network Intrusion Detection Systems,” arXiv, Feb. 2025. arXiv

[5] “Advanced gradient-based evasion attacks on IDS: Poisoning and evasion taxonomy,” Nature Scientific Reports, Apr. 2025. Nature

[6] M. Penmetsa et al., “Adversarial Machine Learning in Cybersecurity: A Review on Defending Against AI-Driven Attacks,” (rev. June 2025). ResearchGate

[7] R. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, “A Detailed Analysis of the KDD CUP 99 Data Set,” in Proc. IEEE Symp. Computational Intelligence for Security and Defense Applications (CISDA), 2009. (NSL-KDD dataset citation) University of New Brunswick

[8] “NSL-KDD Dataset,” Ridwan N. Wibowo et al., Semantic Scholar, 2019. Semantic Scholar

[9] “Adversarial machine learning: a review of methods, tools, and ...” (AML landscape including robustness and privacy), Springer, 2025. SpringerLink

[10] I. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and Harnessing Adversarial Examples,” Commun. ACM, 2015. (Classic foundational work on FGSM) Wikipedia

[11] “Adversarial Machine Learning Attacks Against Intrusion Detection Systems in In-Vehicle Networks…,” Sensors, 2024. MDPI

[12] World Health Organization, “Breast cancer cases and deaths are projected to rise globally,” commented via IARC analysis, Nature Medicine (2025). IARCThe Guardian