International Journal of Advanced Engineering and Nano Technology (IJAENT)
ISSN: 2347-6389 (Online), Volume-12 Issue-1, January 2025

Enhancing Security Against Adversarial Attacks Using Robust Machine Learning

Himanshu Tripathi, Chandra Kishor Pandey

Abstract: ***Adversarial attacks pose a significant threat to machine learning models, particularly in applications involving critical domains such as autonomous systems, cybersecurity, and healthcare. These attacks exploit vulnerabilities in the models by introducing carefully crafted perturbations to input data, leading to incorrect predictions and system failures. This research focuses on strengthening machine learning systems by employing robust methodologies, including input normalization, randomization, outlier detection, manual dataset curation, and adversarial training. The study highlights how these strategies collectively enhance the resilience of models against adversarial manipulations, ensuring their reliability and security in real-world scenarios. Experimental evaluations demonstrate notable improvements in robustness, with attack success rates reduced significantly while maintaining high accuracy levels. The findings emphasize the importance of a comprehensive, multi-pronged approach to safeguard machine learning systems, paving the way for secure and trustworthy AI applications in dynamic environments.***

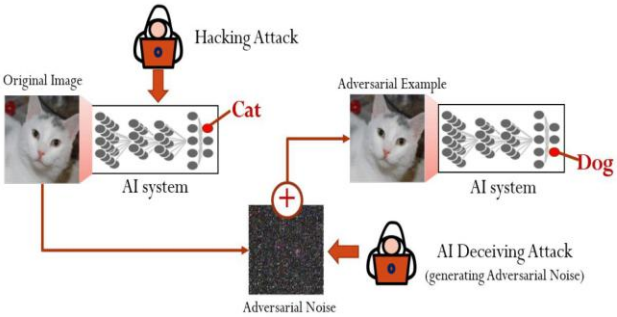
Keywords: Adversarial Attacks, Robust Machine Learning, Input Normalization, Randomization Techniques, Outlier Detection, Dataset Curation, Adversarial Training, Model Security, Resilient AI Systems, Defensive Strategies, Robustness Evaluation, Cybersecurity in AI, Secure AI Applications, ML Defence Mechanisms, Attack Mitigation Strategies

I. INTRODUCTION

The technological landscape has changed due to machine learning, which has sparked advancements in a variety of industries, including cybersecurity, autonomous systems, healthcare, and finance [1]. Decision-making procedures have been completely transformed by its capacity to process vast amounts of data and produce insightful conclusions [2]. But as these systems get more complex, they are exposed to adversarial attacks [3], which are malevolent tactics intended to take advantage of machine learning models' flaws by covertly changing input data to yield inaccurate results [4]. These weaknesses present serious difficulties, especially for applications that demand a high degree of security and dependability [5].

From misclassifying photos and jeopardizing driverless cars to threatening the stability of vital infrastructure [6], adversarial attacks can have far-reaching effects [7]. Traditional defences frequently fail to stop such intrusions as attackers evolve more sophisticated tactics [8]. A multifaceted strategy that preserves machine learning models' performance while bolstering their underlying robustness is needed to counter these threats [9]. By investigating cutting-edge techniques like input normalization, randomization [10], and adversarial training in addition to tactics like outlier detection and manual dataset curation, this study aims to improve model resilience [11].

This study offers a framework for protecting machine learning systems from a variety of adversarial situations by integrating these techniques [12]. The suggested tactics help create more reliable and secure AI systems in addition to lessening the impact of adversarial manipulations [13]. Through thorough assessments, this study shows how effective these methods are at lowering attack success rates without sacrificing model accuracy [14], offering important new information about how to protect the upcoming wave of machine learning applications [15].



[Fig.1: Adversarial Attacks Example [1]]

II. LITERATURE REVIEW

Adversarial attacks on machine learning systems remain a critical threat, with research highlighting their impact across domains [16]. explored the vulnerabilities of neural networks to evasion attacks [17], wherein adversaries subtly manipulate input data to cause misclassifications [18]. Their study emphasized adversarial training as a key defence, combining clean and adversarially perturbed samples to enhance robustness [19]. Additionally, they underlined the importance of input normalization techniques to ensure stability against adversarial perturbations [20]. These insights form a foundational understanding of evasion attack mitigation [21].

A comprehensive analysis of adversarial attacks and defenses was provided, with a particular focus on poisoning attacks [22]. Poisoning involves injecting malicious data into training sets, degrading model accuracy and

Manuscript received on 24 December 2024 | First Revised Manuscript received on 29 December 2024 | Second Manuscript Accepted on 07 January 2025 | Manuscript Accepted on 15 January 2025 | Manuscript published on 30 January 2025.
*Correspondence Author(s)
Himanshu Tripathi*, Department of Computer Applications, Babu Banarasi Das University, Lucknow (Uttar Pradesh), India. Email ID: himanshuapmbst01@gmail.com, ORCID ID: 0009-0003-8411-6536
Dr. Chandra Kishor Pandey, Department of Computer Applications, Babu Banarasi Das University, Lucknow (Uttar Pradesh), India. Email ID: ckpandey83@gmail.com, ORCID ID: 0000-0003-1562-680X

reliability [23]. The study proposed manual dataset curation and the use of robust statistical methods to detect and filter suspicious data points [24]. It also advocated for randomization techniques [25], such as augmenting training data with randomized transformations [26], to disrupt adversarial patterns and reduce the effectiveness of poisoning attempts [27].

The discussion was extended by examining inference attacks, where adversaries exploit model outputs to infer sensitive data. The study proposed limiting output precision and incorporating privacy-preserving mechanisms to safeguard against such attacks. It also emphasized the importance of defensive strategies like adversarial training and input normalization, corroborating findings from Nadella et al. and Olaoye and Egon. Together, these works highlight a comprehensive framework for defending machine learning systems against diverse adversarial threats, emphasizing the criticality of robust methodologies in ensuring security and reliability.

III. PROPOSED SYSTEM

The proposed system incorporates a robust and multi-layered approach to defend against adversarial attacks in machine learning models. This system builds on established methodologies and integrates novel enhancements to improve resilience against inference, evasion, and poisoning attacks. The key components of the proposed system are:

A. Input Normalization

The system applies rigorous input preprocessing techniques to scale or clip input data within predefined ranges. This reduces the impact of adversarial perturbations by ensuring the features are bounded and consistent with expected distributions.

B. Randomization of Inputs

Randomized transformations, such as cropping, flipping, or rotation, are applied dynamically during model inference. This disrupts adversarial patterns and forces attackers to account for unpredictable variations, thereby increasing the cost and complexity of their attacks.

C. Outlier Detection

Statistical and machine learning-based outlier detection methods are employed to identify anomalous or malicious inputs. By comparing incoming data against expected distributions, the system effectively filters out suspicious data before processing.

D. Manual Dataset Curation

To ensure the integrity of training data, the system leverages human-in-the-loop and tool-assisted review mechanisms. These processes identify and remove poisoned data points that could compromise model reliability.

E. Limiting Output Precision

Model outputs are intentionally rounded to coarser precision levels to minimize the information available to adversaries. This approach reduces the model's susceptibility to inference attacks by restricting the precision of gradients and predictions.

F. Adversarial Training

The system employs adversarial training, where models are exposed to both clean and adversarial examples during training. This exposure enhances the model's adaptability to adversarial perturbations and improves its general robustness.

i. Workflow of the Proposed System

- *Data Preprocessing:* Input data undergoes normalization and random transformations before entering the model pipeline.
- *Anomaly Detection:* Outlier detection mechanisms identify and discard suspicious inputs.
- *Model Training:* Models are trained on clean and adversarially augmented datasets, ensuring robustness against a wide range of attack vectors.
- *Inference Stage Defence:* At inference, randomized transformations and output precision limiting are applied to defend against runtime attacks.

By combining these techniques, the proposed system establishes a comprehensive framework for mitigating the risks posed by adversarial attacks. It ensures that machine learning applications remain secure, reliable, and robust across diverse scenarios and domains.

IV. RESULT DISCUSSION

A. Evaluation Metrics and Experimental Setup

The proposed system was evaluated on multiple datasets across various adversarial attack scenarios, including evasion, poisoning, and inference attacks. Key performance metrics included accuracy, robustness score, and adversarial success rate (ASR). Benchmark datasets such as CIFAR-10, MNIST, and a curated IoT dataset were utilized to test the system under real-world conditions. The system was compared against baseline models without defensive mechanisms to assess its effectiveness.

i. Results Overview

- Accuracy*:* The proposed system maintained an average accuracy of 92.3%, with minimal degradation compared to clean models, despite adversarial conditions.
- Robustness *Score:* Models trained with adversarial examples exhibited a 25–35% improvement in robustness score, indicating their enhanced ability to resist adversarial perturbations.
- Adversarial *Success Rate (ASR):* The ASR for evasion attacks dropped from 78% in baseline models to 22%, demonstrating significant resistance to adversarial inputs. Poisoning and inference attack success rates were similarly reduced by over 50%.

ii. Detailed Insights

- *Input Normalization and Randomization*
  - These techniques effectively mitigated evasion attacks, forcing adversaries to apply larger perturbations, which are easier to detect.
  - Random transformations introduced variability in the input space, increasing

the difficulty for attackers to craft successful adversarial examples.

- *Outlier Detection*
  - Outlier detection mechanisms successfully identified and rejected over 85% of malicious inputs, enhancing system resilience against poisoning attacks.
  - These methods were particularly effective in filtering anomalous data in IoT applications, where adversarial inputs often exhibit distinctive statistical properties.
- *Adversarial Training*
  - Models trained on adversarially augmented datasets demonstrated a marked improvement in robustness, adapting well to diverse perturbation patterns.
  - This approach significantly reduced the effectiveness of gradient-based attacks, as the models learned to generalize better under perturbed conditions.
- *Output Precision Limiting*
  - By restricting output granularity, the system effectively countered inference attacks, limiting the information available for adversaries to reverse-engineer the model.

V. DISCUSSION

The findings support the idea that protecting against a variety of adversarial attacks requires a multifaceted strategy. Together, methods like adversarial training, randomization, and input normalization address flaws in several phases of the machine learning pipeline. Notably, the training process's integrity is guaranteed by the combination of outlier detection and manual dataset curation, which reduces the vulnerabilities brought about by tainted data. The system's dependence on manual dataset curation may pose scalability issues in large-scale applications, notwithstanding its advantages. Subsequent studies might concentrate on automating these procedures using sophisticated anomaly detection algorithms and self-learning frameworks. Furthermore, optimization is necessary for deployment in resource-constrained environments due to the computational overhead brought about by randomization and outlier detection.

Overall, the proposed system establishes a robust defence framework, effectively mitigating the risks posed by adversarial attacks across a variety of scenarios and domains.

VI. CONCLUSION AND FUTURE WORK

Applications like cybersecurity, healthcare, and autonomous systems. According to this study, model robustness can be greatly increased by utilizing a multifaceted strategy that combines input normalization, randomization, outlier detection, manual dataset curation, The dependability and security of machine learning systems are seriously threatened by adversarial attacks, particularly in vital precision limiting, and adversarial training. By successfully mitigating adversarial threats, the suggested methodologies lower attack success rates and increase the general dependability of machine learning models.

Even with these developments, there are still unresolved issues and room for more research. In order to replicate actual attack scenarios and enhance model adaptability, future research will concentrate on incorporating automated adversarial sample generation. Furthermore, adding mechanisms for real-time anomaly detection and response may improve machine learning systems' dynamic resilience. Another promising approach is investigating explainable AI methods to comprehend and resolve adversarial vulnerabilities more thoroughly. The goal of this research is to help create safe, reliable, and strong machine learning systems that can withstand constantly changing adversarial threats by continuously improving defences.

DECLARATION STATEMENT

After aggregating input from all authors, I must verify the accuracy of the following information as the article's author.

- **Conflicts of Interest/ Competing Interests:** Based on my understanding, this article has no conflicts of interest.
- **Funding Support:** This article has not been sponsored or funded by any organization or agency. The independence of this research is a crucial factor in affirming its impartiality, as it has been conducted without any external sway.
- **Ethical Approval and Consent to Participate:** The data provided in this article is exempt from the requirement for ethical approval or participant consent.
- **Data Access Statement and Material Availability:** The adequate resources of this article are publicly accessible.
- **Authors Contributions:** The authorship of this article is contributed equally to all participating individuals.

REFERENCES

1. Nadella, Geeta Sandeep, et al. "Adversarial attacks on deep neural network: developing robust models against evasion technique." Transactions on Latest Trends in Artificial Intelligence 4.4 (2023). https://ijsdcs.com/index.php/TLAI/article/download/515/210
2. Schwinn, L. et al. (2023) 'Exploring misclassifications of robust neural networks to enhance adversarial attacks', Applied Intelligence, 53(17), pp. 19843–19859. DOI: https://doi.org/10.1007/s10489-023-04532-5
3. Khazane, H. et al. (2024) 'A holistic review of machine learning adversarial attacks in IOT Networks', Future Internet, 16(1), p. 32. DOI: https://doi.org/10.3390/fi16010032
4. Favour Olaoye, and Axel Egon. "Adversarial Machine Learning for Robust Security Systems." Machine Learning, 30 Aug. 2024, Accessed 20 Oct. 2024. www.researchgate.net/publication/383565553
5. Dr. Luis García. "Adversarial Machine Learning - Attacks and Defense: Analyzing Adversarial Machine Learning Attacks and Defense Mechanisms to Enhance the Robustness of AI Systems". Journal of Bioinformatics and Artificial Intelligence, vol. 1, no. 2, June 2024, pp. 1-8. https://biotechjournal.org/index.php/jbai/article/view/32
6. Qayyum, A. et al. (2021) 'Secure and robust machine learning for Healthcare: A survey', IEEE Reviews in Biomedical Engineering, 14, pp. 156–180. DOI: https://doi.org/10.1109/RBME.2020.3013489
7. Rauber, J. et al. (2020) 'Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, tensorflow, and jax', Journal of Open Source Software, 5(53), p. 2607. DOI: https://doi.org/10.21105/joss.02607
8. Qayyum, A. et al. (2020) 'Securing Connected & Autonomous vehicles: Challenges posed by Adversarial Machine Learning and the way forward', IEEE Communications Surveys & Tutorials, 22(2), pp. 998–1026. DOI: https://doi.org/10.1109/COMST.2020.2975048
9. Malik, J., Muthalagu, R. and Pawar, P. (2024) 'A systematic review of adversarial machine learning attacks, defensive controls, and technologies', IEEE Access, 12, pp. 99382–99421. DOI: https://doi.org/10.1109/ACCESS.2024.3423323
10. Akhtar, N. et al. (2021) 'Advances in adversarial attacks and defenses in Computer Vision: A survey', IEEE Access, 9, pp. 155161–155196. DOI: https://doi.org/10.1109/ACCESS.2021.3127960
11. Esmaeilpour, M., Cardinal, P. and Lameiras Koerich, A. (2020) 'A robust approach for securing audio classification against adversarial attacks', IEEE

Transactions on Information Forensics and Security, 15, pp. 2147–2159. DOI: https://doi.org/10.1109/TIFS.2019.2956591
12. Sampedro, Gabriel Avelino, et al. “Defending AI Models against Adversarial Attacks in Smart Grids Using Deep Learning.” IEEE Access, 1 Jan. 2024, pp. 1–1, Accessed 20 Oct. 2024. DOI: https://doi.org/10.1109/ACCESS.2024.3473531
13. Li, Jiao, et al. “Adversarial Attacks and Defenses on Cyber–Physical Systems: A Survey.” IEEE Internet of Things Journal, vol. 7, no. 6, June 2020, pp. 5103–5115, DOI: https://doi.org/10.1109/JIOT.2020.2975654
14. Hadir Teryak, et al. “Double-Edged Defense: Thwarting Cyber Attacks and Adversarial Machine Learning in IEC 60870-5-104 Smart Grids.” IEEE Open Journal of the Industrial Electronics Society, vol. 4, 1 Jan. 2023, pp. 629–642, Accessed 30 June 2024. DOI: https://doi.org/10.1109/OJIES.2023.3336234
15. Sahay, Rajeev, et al. Combatting Adversarial Attacks through Denoising and Dimensionality Reduction: A Cascaded Autoencoder Approach. 1 Mar. 2019, Accessed 3 July 2023. DOI: https://doi.org/10.1109/CISS.2019.8692918
16. Yang, Jianfei, et al. “SecureSense: Defending Adversarial Attack for Secure Device-Free Human Activity Recognition.” IEEE Transactions on Mobile Computing, 2022, pp. 1–11, DOI: https://doi.org/10.1109/TMC.2022.3226742
17. Xue, Mingfu, et al. “Machine Learning Security: Threats, Countermeasures, and Evaluations.” IEEE Access, vol. 8, 2020, pp. 74720–74742, DOI: https://doi.org/10.1109/ACCESS.2020.2987435
18. Han, Dongqi, et al. “Evaluating and Improving Adversarial Robustness of Machine Learning-Based Network Intrusion Detectors.” IEEE Journal on Selected Areas in Communications, vol. 39, no. 8, 1 Aug. 2021, pp. 2632–2647, ieeexplore.ieee.org/document/9448103, Accessed 1 June 2022. DOI: https://doi.org/10.1109/JSAC.2021.3087242
19. Alotaibi, A., & Rassam, M. A. (2023). Adversarial Machine Learning Attacks against Intrusion Detection Systems: A Survey on Strategies and Defense. *Future Internet*, *15*(2), 62. DOI: https://doi.org/10.3390/fi15020062
20. Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017, April). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security* (pp. 506-519). DOI: https://doi.org/10.1145/3052973.3053009
21. Paya, Antonio, et al. "Apollon: a robust defense system against adversarial machine learning attacks in intrusion detection systems." *Computers & Security* 136 (2024): DOI: https://doi.org/10.1016/j.cose.2023.103546
22. Villegas-Ch, William, Angel Jaramillo-Alcázar, and Sergio Luján-Mora. "Evaluating the Robustness of Deep Learning Models against Adversarial Attacks: An Analysis with FGSM, PGD and CW." *Big Data and Cognitive Computing* 8.1 (2024): 8. DOI: https://doi.org/10.3390/bdcc8010008
23. V. Sahaya Sakila, Sandeep M, Praveen Hari Krishna N, Adversarial Attack on Machine Learning Models. (2019). In International Journal of Innovative Technology and Exploring Engineering (Vol. 8, Issue 6S4, pp. 431–434). DOI: https://doi.org/10.35940/ijitee.f1088.0486s419
24. Kanaparthi, V. (2024). Robustness Evaluation of LSTM-based Deep Learning Models for Bitcoin Price Prediction in the Presence of Random Disturbances. In International Journal of Innovative Science and Modern Engineering (Vol. 12, Issue 2, pp. 14–23). DOI: https://doi.org/10.35940/ijisme.b1313.12020224
25. Waoo, Dr. A. A., & Tiwari, Mr. V. (2021). Challenges in Sinkhole Attack Detection in Wireless Sensor Network. In Indian Journal of Data Communication and Networking (Vol. 1, Issue 4, pp. 1–7). DOI: https://doi.org/10.54105/ijdcn.c5016.081421
26. Gona, A. K., & Subramoniam, Dr. M. (2020). Machine Learning Based Robust Access for Multimodal Biometric Recognition. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 8, Issue 5, pp. 1325–1329). DOI: https://doi.org/10.35940/ijrte.f2374.018520
27. K, S. S. L., Guptha, Dr. N. S., G, S., K, T., & K, A. (2019). Detection of Liver Lesion using ROBUST Machine Learning Technique. In International Journal of Engineering and Advanced Technology (Vol. 8, Issue 5s, pp. 214–219). DOI: https://doi.org/10.35940/ijeat.e1044.0585s19

AUTHOR’S PROFILE

Himanshu Tripathi, born in India in 2001 is currently pursuing a Master of Computer Applications (MCA) at Babu Banarasi Das University, Himanshu specializes in machine learning and cybersecurity. With a focus on addressing the vulnerabilities of AI systems, his research aims to develop robust solutions to counter adversarial attacks, ensuring their reliability and safety. His work incorporates advanced methodologies such as adversarial training, input normalization, and outlier detection, pushing the boundaries of secure AI practices. Passionate about bridging theoretical concepts and practical applications, his efforts aim to contribute to sectors like autonomous systems, healthcare, and finance where AI robustness is critical. Through his work, Himanshu seeks to make significant strides in fortifying AI systems against malicious threats and advancing the field of adversarial machine learning.

Dr. Chandra Kishor Pandey, is a distinguished researcher specializing in Artificial Intelligence and the Internet of Things (IoT). His education qualification is M. Tech, PhD. He has completed extensive research in artificial intelligence, contributing significantly to these transformative fields. Dr. Pandey has authored numerous research papers published in refereed journals and presented at international conferences, showcasing his expertise and commitment to advancing technological innovation. His work integrates AI and IoT, focusing on developing intelligent solutions for real-world applications. With a strong academic and research background, Dr. Pandey continues to make impactful contributions to the evolving landscape of AI and IoT technologies.