International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2229-5518
SJIF (2025): 6.623

Vulnerabilities of AI in Cybersecurity

Shahabuddin Shahid

Kardan University, Faculty of Computer, 40 Meter Road, Kabul 1007, Afghanistan
Email: *shahid.kontakt[at]gmail.com*

Abstract: ***Artificial intelligence create new ways of attack that adversaries could take advantages of as it becomes more connected to cybersecurity systems. Two important vulnerabilities are evaluated in this paper: Model poisoning and adversarial attacks. We look at how these issues can damage overall security, dependability and integration of technologies that AI is used in, especially in cybersecurity applications. The risks of releasing AI in an adversarial environment are increasingly recognized by new researchers and developers.***

Keywords: Artificial Intelligence, Cybersecurity, adversarial attacks, Model poisoning, Security Weaknesses, Machine Learning Security, Threat Detection

1. Introduction

Artificial intelligence (AI) has significantly enhanced cybersecurity by enabling automated behavior analysis and more effective thread detection. However, the integration of AI into cybersecurity system also introduces new vulnerabilities. Adversaries now exploit AI itself by manipulating algorithms, data inputs and decision-making processes to deceive or bypass security measures. This poses serious risks to critical components such as intrusion detection system (IDS), malware classifiers and network protection tools. To counter these emerging threats, it is essential to develop robust defenses, including techniques like adversarial training, continuous monitoring and the implementation of multi-layered security strategies.

Strengthening AI-powered cybersecurity systems will require ongoing research and adaptive measure to stay ahead of malicious actors exploiting AI vulnerabilities.

2. Methodology Approach

This paper takes a serious alternative and analytical approach, based on an overall view of academic literature, white paper, practical and real-world examples. The study doesn't use real data to experiments. Instead, it uses information from a number of reliable resources to show the main problem with today's AI systems used in cybersecurity, the aims of this method is to provide a theoretical but useful overview of weaknesses and possible defense.

3. Literature Review

Several strong studies have explored the weaknesses of AI models in challenging environments. Goodfellow et aI. (2014) proved how minimal answers could misdirect deep neural networks, introducing the concepts of adversarial examples. Steinhardt et aI. (2017) expanded on this by identifying how poison training data could compromise machine models. Biggio and Roli (2018) provided a comprehensive decade-long review of adversarial machine learning, showing the persistent risks. The foundational works highlighted the important of secure AI model designed in cybersecurity applications.

4. Core Vulnerabilities

4.1 Adversarial Attacks

Adversarial attacks manipulate inputs in a way that causes AI models to misunderstand them while appearing normal to human observers. In cybersecurity, this may include covering up malicious behavior in a way that bypasses detection by AI-based intrusion systems or malware filters. These attacks use the mathematical structure of AI algorithms to misuse their hidden weaknesses.

4.2 Model Poisoning

Model poisoning occurs when attackers influence the training data to insert weaknesses to the artificial intelligence model. This leads the model to incorrectly operate under specific conditions.

Poisoned models might ignore dangerous behavior or answer inaccurately to threats and weakening the purpose of their deployment in cybersecurity.

4.3 Figure

The diagram below shows the vulnerability of a pipeline in AI systems in cybersecurity. It highlights the stages form data collection and model training to deployment, where adversarial attacks or poisoning may affect system functionalities.

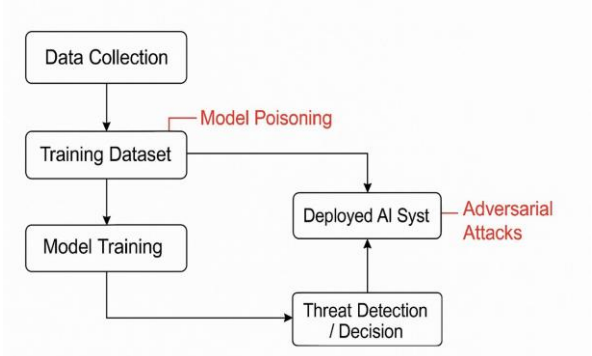

Figure 1: AI Vulnerabilities in Cybersecurity Pipeline

4.4 Table

Table 1: Adversarial vs. Poisoning: Key Traits

Aspect	Adversarial Attacks	Model Poisoning
Aim	Mislead the AI system during Interface	Corrupt the model during training
Impact	Evasion of detection	Long-term behavioral manipulation
Defense	Adversarial Training, input filtering	Data Validation, robust learning techniques

5. Discussion and Future Works

The challenges caused by adversarial attacks and model poisoning are not only technical challenges but also rise concerns about trust, reliability, and governance in AI systems. With every progress, artificial intelligence used to protect digital infrastructure, these vulnerabilities also grow stronger. The methods of learning in certified defenses and explainable is important work used to provide certified defenses, transparent Artificial intelligence behavers and learning structures. In addition, collaboration of cybersecurity experts and artificial intelligence practitioners are vital to stop future threats before they can be exploited.

6. Conclusion

Artificial intelligence not only improve cybersecurity power but also introduce unique vulnerabilities. This study shows how adversarial attacks and model poisoning present sustainable risks to the stability and functionality of artificial intelligence powered sorority systems. Dealing with these problems is vital for creating reliable artificial intelligence applications that can protect important digital assets in challenging environment.

Acknowledgment
The author would like to express warm appreciation to the Faculty of Computer Science at Kardan University, Kabul, Afghanistan, for their academic guidance, encourages and prolonged supports during this research effort.

References

[1] Goodfellow, I., Shlens, J., & Szegedy, C. (2014). *Explaining and harnessing adversarial examples*. arXiv preprint arXiv:1412.6572.
[2] Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317–331.
[3] Steinhardt, J., Koh, P. W., & Liang, P. (2017). Certified defenses for data poisoning attacks. Advances in Neural Information Processing Systems, 30, 3517–3529.
[4] Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B. (2018). Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In 2018 IEEE Symposium on Security and Privacy (pp. 19–35). IEEE.
[5] Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I. P., & Tygar, J. D. (2011). Adversarial machine learning. In Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence (pp. 43–58). ACM.
[6] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy* (pp. 372–387). IEEE.

Author Profile

Shahabuddin Shahid is a first-year Bachelor of Computer Science student at Kardan University, Kabul, Afghanistan. His academic interests include artificial intelligence, cybersecurity web and software development. He has strong motivation to explore Artificial intelligence applications in strengthen digital security. This research shows his early and strong contribution to the field, focusing on the vulnerabilities of AI in cybersecurity.