International Journal of Interdisciplinary Innovative Research & Development (IJIIRD)
ISSN: 2456-236X
Vol. **11** Special Issue **02** | **2026**

Securing Intelligence: Cyber Threats in AI Systems

Tina Borole[1], Snehal Pawar[2], Madhuree Raising[3], Pooja Patil[4]
[1,2,3,4]*Student CSE, VBK Malkapur, Maharashtra, India*

DOI: 10.5281/zenodo.19288479

ABSTRACT

Artificial Intelligence (AI) systems have become integral to modern digital infrastructures, powering applications across healthcare, finance, transportation, and national security. While AI enhances efficiency and decision-making, it also introduces new and complex cybersecurity challenges. AI systems are increasingly targeted by cyber threats such as data poisoning, adversarial attacks, model theft, inference attacks, and unauthorized access, which can compromise system integrity, confidentiality, and reliability. These threats exploit vulnerabilities in training data, algorithms, deployment environments, and human interactions, posing significant risks to both organizations and end users.

Keywords:- Artificial Intelligence (AI)Cybersecurity,AI Security,Adversarial Attacks,Data Poisoning,Model Inversion,Inference Attacks,Secure Machine Learning,Threat Mitigation,Security-by-Design,Trustworthy AI

1.INTRODUCTION

Artificial Intelligence (AI) systems are increasingly embedded in critical infrastructure, autonomous systems, and decision-making pipelines. As these systems grow in capability, they also become targets of sophisticated cyber-attacks that exploit vulnerabilities at data, model, and deployment layers. Securing AI is now a multidisciplinary research frontier spanning machine learning (ML), cyber security, and systems engineering.

2. THREAT TAXONOMY IN AI SYSTEM

1.1 Adversarial Machine Learning

Key Idea: Attackers craft inputs that cause misclassification or misbehavior.
Szegedy et al. (2014) — one of the first works to expose imperceptible perturbations that fool neural networks.
Goodfellow et al. (2015) — introduced the Fast Gradient Sign Method (FGSM) for generating adversarial samples.
Papernot et al. (2016) — explored transferability of adversarial examples across models.
research work Introduction related your research work Introduction related your research work Introduction related your research work Introduction related your research work Introduction related your research work.

1.2 Poisoning Attacks

Key Idea: Corrupt training data to influence learned behavior.
Biggio et al. (2012) — demonstrated poisoning attacks on Support Vector Machines.
Steinhardt et al. (2017) — formalized robust learning in presence of corrupt data.
Bagdasaryan et al. (2020) — showed targeted poisoning in federated learning.

Insights:
Poisoning can cause backdoors or stealthy biases.
Particularly serious in distributed training (e.g., federated systems).

1.3 Model Extraction & Inference Attacks

Key Idea: Attackers steal model parameters or infer sensitive training data.
Tramèr et al. (2016) — model extraction attacks via API queries.
Shokri et al. (2017) — membership inference attacks uncover if specific data was used in training.
Carlini et al. (2019) — amplified privacy risks in generative models.

Insights:
Black-box access can reveal model internals.
Privacy breach vectors grow with predictive APIs.

3. SECURITY ACROSS THE AI LIFECYCLE

Data Collection & Preprocessing
Attack surface includes unverified sensor feeds, scraped data, and user inputs.
Hendrycks et al. (2018) investigated robustness to corrupted data.
Dataset quality remains central to both performance and security.

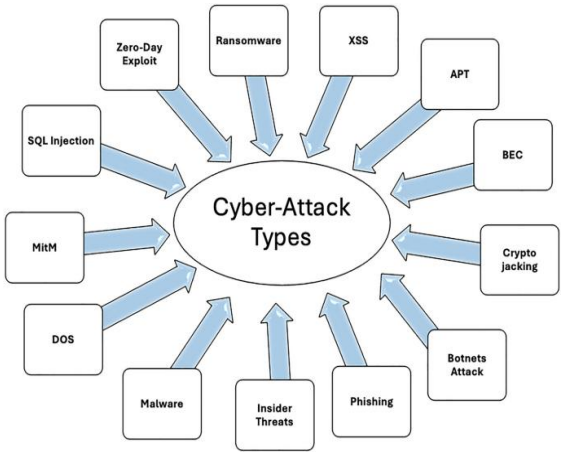
Open Challenges:
Automated detection of poisoned data.
Standards for trustworthy data curation.

Model Training & Optimization
Securing training involves resisting poisoning and ensuring gradient integrity.
Xie et al. (2020) developed Byzantine-resilient federated learning.

Research Gaps:
Scalable defenses against poisoned gradients.
Trade-offs between robustness and accuracy.

Deployment & Runtime Monitoring
Runtime threats include evasion attacks and sensor spoofing.
Eykholt et al. (2018) showed physical adversarial examples that fool cameras in autonomous vehicles.

Key Ideas:
Monitor runtime behavior for anomalies.
Incorporate ensemble and redundancy to detect attacks.



4. CONCLUSION

The security of AI systems is a multi-layered arms race. Research has mapped many threats — from adversarial inputs and poisoning to model theft and data leakage — but practical deployment often outpaces defense. Continued work on certified robustness, privacy preservation, and lifecycle security frameworks remains essential.

5. REFERENCES

[1]. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. CCS.
[2]. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. AISTATS
[3]Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against SVMs. ICML.
[4]. Carlini, N., Liu, C., Erlikhman, G., & Wagner, D. (2019). On evaluating adversarial robustness. arXiv.
[5]Papernot, N., McDaniel, P., Goodfellow, I., et al. (2016). Practical black-box attacks. ASIA CCS.
[6]Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks. IEEE S&P.
[7]Szegedy, C., Zaremba, W., Sutskever, I., et al. (2014). Intriguing properties of neural networks. ICLR.