Emerging AI/ML Attacks and Risks

Navigating Emerging AI/ML Threats in Application Security

Edited by **MOHAN DAS VISWAM**

In the relentless pursuit of innovation, applications are harnessing the remarkable capabilities of Artificial Intelligence (AI) and Machine Learning (ML). These technologies promise transformative advances, from intelligent recommendation systems and chatbots for precise data analysis. But these systems can also introduce vulnerabilities that can be exploited by malicious actors. As AI and ML become indispensable in our digital ecosystem, understanding the risks and attacks they can be used for is pivotal. This article embarks on a journey through the intricacies of these evolving risks, unveiling the mechanisms behind AI/ML-related vulnerabilities. By understanding these, organizations can fortify their defenses, ensuring that the promise of AI/ML technology is embraced while guarding against its potential perils

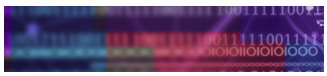
Emerging AI/ML Threats and Risks

As AI and ML continue to permeate various sectors, they also bring forth a range of new cybersecurity threats and attacks that must be addressed to protect application security and cybersecurity as a whole. Here, we outline some of the prominent AI/ML-related threats and attacks and provide insights into strategies to mitigate them.

Tasiruddin Ahmed
Sr. Technical Director
asm-tasir@nic.in

Bronjon Gogoi
Scientist-C
asm-bronjon@nic.in

In the era of relentless innovation driven by Artificial Intelligence (AI) and Machine Learning (ML), applications are gaining remarkable capabilities. From intelligent chatbots to precise data analysis, these technologies hold immense promise. However, they also bring vulnerabilities that malicious actors can exploit. Understanding these threats is crucial for organizations to fortify their defenses and embrace the potential of AI/ML while safeguarding against potential perils. The convergence of security and innovation is shaping the future of application security.



Data Poisoning

Data poisoning is a stealthy attack that involves altering the training data fed into AI models. This manipulation can subtly skew the model's understanding, resulting in

incorrect predictions or classifications. For instance, malicious actors could poison the data used to train medical diagnosis AI systems, leading to erroneous disease diagnoses and potentially severe consequences.

Mitigation strategies for data poisoning include rigorous data validation and anomaly detection mechanisms to identify and exclude tainted data during model training.

Adversarial Attacks

Adversarial attacks are a significant concern, where manipulated data is injected into AI systems to manipulate their outcomes, leading to misclassification or incorrect predictions. These attacks can be performed by insiders or external attackers. (See Figure 2) For example, a self-driving car can misinterpret a stop sign for a speed limit sign due to subtle alterations in the visual data.

Mitigating adversarial attacks requires developing robust anomaly detection systems and continually updating AI models to adapt to evolving threat.

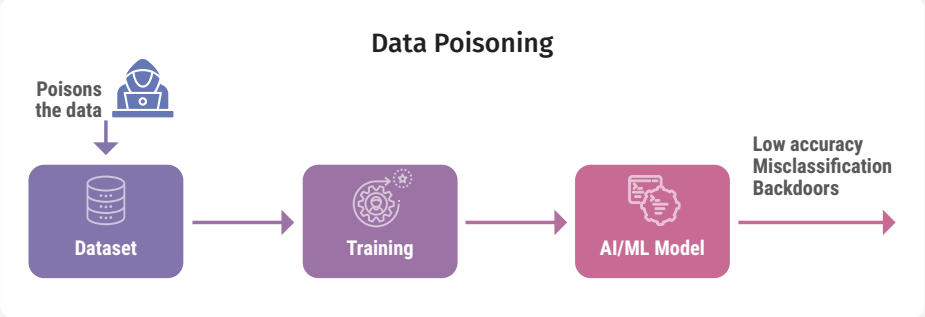
Evasion of Detection Systems

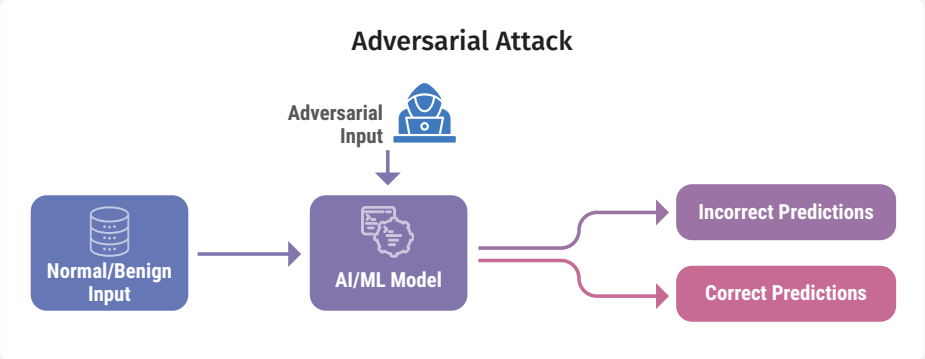
In Evasion of Detection Systems attackers identify blind spots in the system and exploit them to evade detection. A simple diagram of evasion of detection is shown in Figure 3. Here the attacker modifies the input data to trick the model into not detecting something it should detect, for example, a security system not detecting malware due to clever modification of the malware's data and code. It may seem identical to adversarial but they differ in objectives and the process by which it achieves them. An attacker manipulating malware / virus / trojan to avoid recognition by an AI-based antivirus program is an example of an evasion of detection attack.

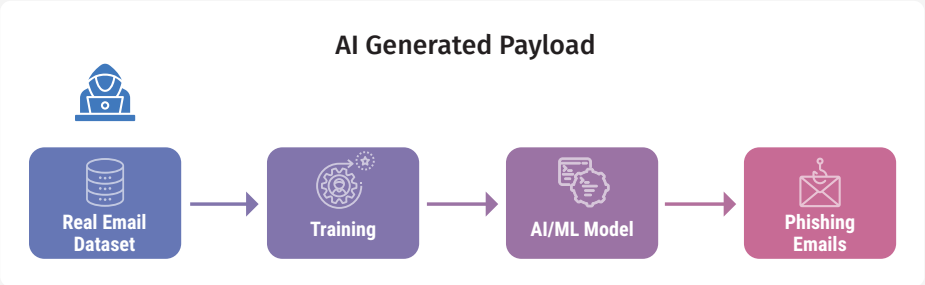
In order to mitigate such attacks, organizations should regularly evaluate and update their detection mechanisms.

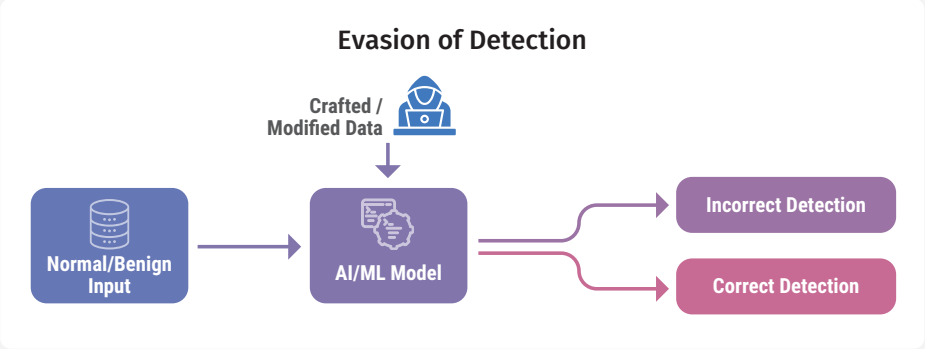
AI-Generated Payloads

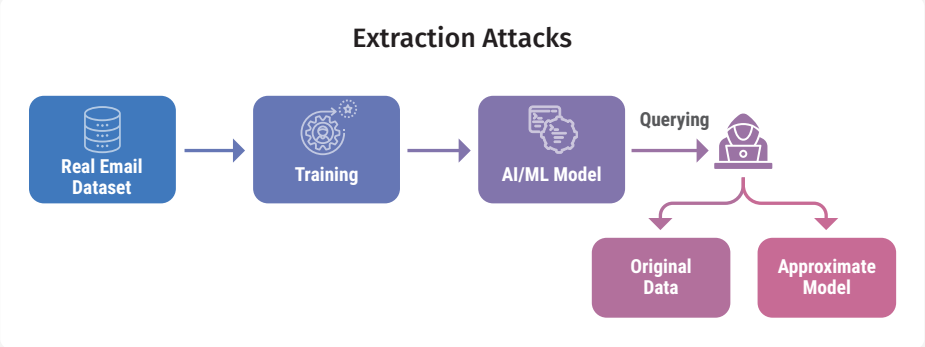
Malicious actors can leverage AI/ML techniques to generate attack payloads that evade traditional defenses. An example is shown in Figure 4 where an attacker trains an AI/ML model to generate phishing emails by providing











real email communications as the dataset. The phishing emails generated by this model will mimic genuine communication, increasing the likelihood of successful phishing attacks.

To counter AI-generated payloads, organizations need to enhance email filtering systems with AI-driven anomaly detection to identify and block such malicious messages.

AI Enhanced social engineering

Attackers leverage AI-driven insights to craft highly personalized and convincing social engineering attacks, increasing the likelihood of successful phishing or impersonation attacks. For example, Attackers use AI to analyze a target's social media activity and preferences to craft tailored phishing emails that appear more convincing.

Mitigating AI-enhanced social engineering attacks necessitates user education, multi-factor authentication, and advanced behavior analytics to identify suspicious activities.

Extraction Attacks

Attackers employ AI/ML methods to extract sensitive data or confidential information from AI models or datasets. An illustration of the extraction attack is shown in Figure 5. The attacker queries the AI model providing inputs, and from the responses of the models, the attacker can reconstruct the original data set used for training the model or create an approximate model. For instance, attackers can use machine learning algorithms to reverse engineer proprietary algorithms or extract personal information from a machine learning model.

Protecting against extraction attacks involves implementing robust access controls, encryption, and monitoring mechanisms to safeguard sensitive data and model outputs.

Conclusion

In conclusion, while the integration of AI and ML in applications empowers innovation, it also introduces new security challenges. Recognizing and mitigating these emerging threats is imperative. With vigilant understanding and proactive defenses, organizations can harness the transformative potential of AI/ML while safeguarding their digital landscapes.

As technology evolves, the synergy between security and innovation becomes paramount, shaping the future of application security. It is essential for organizations to stay ahead of these AI/ML-related risks and continuously adapt their cybersecurity strategies to defend against evolving threats.

Contact for more details

Bronjon Gogoi
Scientist-C
RCoEAS, Jayanagar, Beltola
Guwahati - 781022
Email: asm-bronjon@nic.in, Phone: 9365558235