Incident Response Plan

Prepared by: Isabelle Jaber
Date: September 18th, 2024

Table of Contents:

EXECUTIVE SUMMARY

This Incident Response Plan focuses on VISION's Artificial Intelligence (AI) infrastructure, which powers critical computer vision models for healthcare diagnostics. As machine learning systems are vulnerable to a variety of specific threats, such as model poisoning, adversarial attacks, and data manipulation, this Incident Response Plan is designed to respond to incidents targeting VISION's AI models, data pipelines, and underlying infrastructure. The plan ensures that VISION can quickly mitigate security threats to its Machine Learning (ML) services and maintain reliable service delivery for clients.

INTRODUCTION

Machine learning systems at VISION play a crucial role in analyzing medical images such as MRIs and CT scans. The potential vulnerabilities unique to ML, such as adversarial inputs, data poisoning, and model theft, pose significant risks to both the integrity of AI models and the healthcare data they process. This Incident Response Plan is specifically tailored to address security incidents affecting machine learning models and the associated infrastructure, ensuring that the AI solutions remain secure, compliant, and operational during and after an incident.

INCIDENT RESPONSE PLAN GOALS

The goals of this Incident Response Plan are as follows:

1. **Quick Identification**: Rapid detection of machine learning-specific threats (e.g., model poisoning, adversarial attacks) to ensure timely response.
2. **Containment**: Limit the spread of security incidents affecting ML models or data pipelines.
3. **Mitigation and Remediation**: Protect ML model integrity by mitigating vulnerabilities and remediating attacks.
4. **Recovery**: Restore affected ML systems with minimal downtime while ensuring data and model accuracy.
5. **Compliance**: Ensure adherence to healthcare regulations, such as HIPAA, when responding to ML-related incidents.

INCIDENT RESPONSE PLAN PROCEDURES

PREPARATION

- **Security Monitoring**: Deploy machine learning-specific monitoring tools such as model activity logs, feature manipulation detection, and adversarial input monitoring.
- **Computer Security Incident Response Team (CSIRT)**: Establish a specialized ML-focused IRT trained to identify and respond to model and data pipeline-related incidents.

DETECTION AND ANALYSIS

- **Monitor for Adversarial Attacks**: Use techniques from MITRE's ATLAS framework to detect adversarial inputs designed to trick or manipulate model outputs (e.g., modifying healthcare images).
- **Anomaly Detection**: Set thresholds and use behavior analysis to identify irregular activity in training data pipelines and model performance.
- **Source Verification**: Continuously verify the integrity and source of training data to prevent data poisoning.
- **Model Drift Detection**: Implement tools to track model performance drift, which could indicate external tampering or adversarial attacks.

CONTAINMENT

- **Isolate Compromised Models**: Immediately isolate any machine learning models suspected of being compromised.
- **Disable Vulnerable Components**: Disable interfaces, APIs, or external systems interacting with the affected models to prevent further damage or data manipulation.
- **Restrict Access**: Limit access to the ML pipeline, including restricting privileged access to data, model artifacts, and training resources.

ERADICATION

- **Adversarial Defense**: Apply defensive techniques (e.g., adversarial training or data sanitization) to remove poisoned data or adversarial inputs from the system.
- **Model Retraining**: If necessary, retrain affected models using verified clean data to restore performance and accuracy.

RECOVERY

- **Revalidate Models**: Run validation tests on restored models to ensure they function as expected and are not compromised.
- **Reintroduce Services**: Gradually reintroduce ML models into production following stringent testing and monitoring.
- **Review Dependencies**: Ensure that all dependencies, such as data pipelines, are functioning securely post-incident.

POST-INCIDENT REVIEW

- **Incident Analysis**: Conduct a thorough analysis of the incident to identify the root cause and affected systems.
- **Update Response Protocols**: Integrate lessons learned from the incident into future threat detection and response mechanisms.
- **Report to Stakeholders**: Provide detailed reports to internal stakeholders and regulatory bodies, ensuring compliance with healthcare regulations such as HIPAA.

SECURITY TOOLING

To support the Incident Response Plan, VISION will leverage the following tools:

1. **Adversarial Detection Tools**: Use tools recommended by OWASP for identifying adversarial inputs and attempts to manipulate ML models.
2. **Model Integrity Monitoring**: Leverage solutions from MITRE’s ATLAS to continuously monitor the integrity of machine learning models and detect unusual patterns.
3. **Data Pipeline Security**: Implement automated pipeline verification and integrity checks to detect any unauthorized changes to training or input data.
4. **SIEM with ML Support**: Use a Security Information and Event Management (SIEM) system integrated with machine learning-specific threat detection capabilities.

CONCLUSION

The Incident Response Plan is a crucial component of VISION's overall security strategy, aimed at ensuring VISION's machine learning models remain secure, reliable, and compliant in the face of evolving threats. By focusing on ML-specific vulnerabilities and mitigation strategies, VISION can effectively manage incidents, restore operations, and protect sensitive healthcare data. Continuous testing, improvement, and training will ensure that VISION can adapt to evolving threats while maintaining its clients' trust and regulatory compliance.