COMPARATIVE ANALYZES SECURITY AI AND ML TRAINED MODELS

Bozorov Suhrobjon
TUIT named after Muhammad al-Khwarizmi

Abstract

Artificial Intelligence (AI) and Machine Learning (ML) are revolutionizing cybersecurity through advanced threat detection and prevention. This article compares the security challenges and methods used to safeguard trained models in AI and ML systems. It explores techniques for protecting model integrity, analyzes how these methods are applied to AI and ML, and presents a comparative analysis of their effectiveness in securing trained models.

Keywords: Artificial Intelligence, Machine Learning, Cybersecurity, Model Security, Threat Detection.

Introduction

Artificial Intelligence (AI) and Machine Learning (ML) are increasingly integrated into cybersecurity for tasks like intrusion detection, malware classification, and behavioral analysis. AI systems rely on ML models trained on large datasets to predict outcomes and detect anomalies. However, these models are vulnerable to security challenges, such as adversarial attacks and model inversion threats, which can compromise their integrity and confidentiality.

Protecting trained models is critical to ensuring the reliability of AI and ML systems. Techniques to safeguard these models, such as encryption, differential privacy, and trusted execution environments, are integral to addressing evolving threats. This article provides a comprehensive analysis of these methods, exploring their implementation and security implications.

Literature Review

Recent research highlights the importance of securing AI and ML models:

Mauri and Damiani (2022) discuss STRIDE-AI, a methodology for modeling threats to AI-ML systems and selecting effective security controls (Mauri & Damiani, 2022).

Ahmad and Prasad (2023) propose an AI-enabled cybersecurity framework utilizing Random Forest models for intrusion detection, achieving 95.97% accuracy (Ahmad & Prasad, 2023).

Xu et al. (2020) introduce MIDAS, a hardware-oriented solution to defend ML models against model inversion attacks using approximate memory systems (Xu et al., 2020).

Mothukuri et al. (2021) demonstrate the use of federated learning for anomaly detection in IoT networks, maintaining data privacy and enhancing security (Mothukuri et al., 2021).

Liu et al. (2022) employ Explainable AI (XAI) to understand the performance of ML-based malware classifiers and mitigate biases (Liu et al., 2022).

AI and ML: Description

Artificial Intelligence (AI): AI encompasses systems capable of simulating human intelligence, performing tasks such as reasoning, learning, and decision-making.

Machine Learning (ML): ML, a subset of AI, involves algorithms that learn patterns from data to make predictions or decisions without explicit programming. ML models include supervised learning (e.g., decision trees), unsupervised learning (e.g., clustering), and reinforcement learning.

Methods of Saving Trained Models

- Encryption: Encrypting model parameters ensures that even if intercepted, the model cannot be easily reverse-engineered.
- Differential Privacy: Adds noise to model outputs, protecting sensitive training data from reconstruction attacks.
- Federated Learning: Distributes training across devices, ensuring raw data remains decentralized and secure.
- Trusted Execution Environments (TEEs): Isolate models within secure hardware environments to prevent unauthorized access.
- Blockchain: Secures model integrity by recording updates and changes in an immutable ledger.

Methods Used by AI and ML

AI systems often rely on encryption and TEEs to secure large, complex models. ML models leverage federated learning and differential privacy to safeguard training data and model outputs.

Method	How It Works	Applications	Benefits	Challenges
Encryption	Encrypts model parameters and data during storage and transmission using algorithms like RSA and AES.	Critical applications in healthcare, finance, and other sensitive domains.	Provides strong protection against unauthorized access and tampering.	Computationally intensive for large models; may impact scalability.
Trusted Execution Environments (TEEs)	Executes models within secure hardware environments, isolating computations from external interference.	Edge computing, on-device AI, IoT devices.	Ensures confidentiality and security even in untrusted environments.	Limited availability of TEE-compatible hardware; potential vulnerabilities in implementations.
Federated Learning	Decentralizes training by keeping data on devices and sharing only model updates for aggregation.	Personalized recommendations, healthcare diagnostics, mobile apps.	Maintains data privacy; raw data never leaves the device; enables collaborative model training.	Bandwidth-intensive; requires robust aggregation mechanisms.
Differential Privacy	Adds noise to training data or outputs, masking individual contributions while preserving utility.	Scenarios involving sensitive data, like medical records or user activity logs.	Protects against data reconstruction attacks while allowing models to learn general patterns effectively.	Excessive noise can degrade model accuracy and performance.

Table 1. Comparison of methods used by AI and ML

Security Method	Advantages	Challenges	Use Cases
Encryption	Strong protection against unauthorized access.	Computationally intensive for large models.	Critical applications like finance and healthcare.
Differential Privacy	Protects sensitive training data.	Can degrade model accuracy if too much noise is added.	ML models handling personal data.
Federated Learning	Maintains data privacy by decentralizing training.	Requires high communication bandwidth and robust aggregation mechanisms.	IoT and mobile devices.
Trusted Execution Environments (TEEs)	Prevents tampering through secure hardware.	Limited by hardware availability and potential vulnerabilities in TEE implementations.	On-device AI and edge computing.
Blockchain	Immutable and transparent record of model changes.	Scalability issues and high energy costs for large-scale deployments.	Collaborative AI model training.

Table 2. Use cases of methods of saving trained models

Conclusion

Securing trained models is crucial for the reliable deployment of AI and ML systems in cybersecurity. While encryption, federated learning, and TEEs provide robust mechanisms, challenges such as computational costs and scalability must be addressed. By adopting tailored security methods, organizations can enhance the resilience of their AI and ML applications against evolving threats.

References

1. Mauri, L., & Damiani, E. (2022). Modeling Threats to AI-ML Systems Using STRIDE. Sensors. DOI: 10.3390/s22176662.

2. Ahmad, S. S., & Prasad, K. (2023). An Artificial Intelligence (AI) Enabled Framework for Cyber Security Using Machine Learning Techniques. International Research Journal. DOI: 10.61916/prmn.2023.v02i01.009.

3. Xu, Q., Arafin, M. T., & Qu, G. (2020). MIDAS: Model Inversion Defenses Using an Approximate Memory System. AsianHOST. DOI: 10.1109/AsianHOST51057.2020.9358254.

4. Mothukuri, V., Khare, P., & Srivastava, G. (2021). Federated-Learning-Based Anomaly Detection for IoT Security Attacks. IEEE IoT Journal. DOI: 10.1109/JIOT.2021.3077803.

5. Liu, Y., Tantithamthavorn, C., Li, L., & Liu, Y. (2022). Explainable AI for Android Malware Detection. ISSRE. DOI: 10.1109/ISSRE55969.2022.00026.