

ABSTRAK

Integrasi *Artificial Intelligence* (AI) di berbagai sektor telah memperluas permukaan serangan siber sehingga ancaman terhadap sistem AI berkembang menjadi serangan berantai yang memerlukan kerangka klasifikasi untuk mendukung identifikasi dan mitigasi. Penelitian ini bertujuan menganalisis jenis dan pola serangan berantai pada Sistem AI menggunakan *Cybersecurity Attack Dataset 2025* serta merumuskan strategi mitigasi berbasis *MITRE ATLAS*. Dari 14.133 data serangan, diperoleh 1.492 data terkait AI yang terdiri atas 502 data AI Data *Leakage & Privacy Risks*, 496 data AI/ML *Security*, dan 494 data AI *Agents & LLM Exploits*. Rekonstruksi *Attack Steps* menunjukkan tiga pola serangan dominan, yaitu Data *Exfiltration* via *Prompt Injection* 10 langkah, *Intercept WebView Cookies* 12 langkah, dan *Prompt-In-Attribute Trick* 7 langkah. Hasil pemetaan menunjukkan bahwa Data *Exfiltration* via *Prompt Injection* sepenuhnya terpetakan ke tiga teknik AML dengan enam rekomendasi mitigasi resmi, *Intercept WebView Cookies* tidak memiliki padanan dalam *MITRE ATLAS* karena menargetkan lapisan aplikasi, sedangkan *Prompt-In-Attribute Trick* hanya terpetakan sebagian karena melibatkan kerentanan aplikasi web berupa DOM XSS. Temuan ini menunjukkan bahwa *MITRE ATLAS* efektif mendukung identifikasi teknik serangan berbasis AI dan penyusunan strategi mitigasi berlapis.

Kata kunci : Sistem AI, serangan berantai, *MITRE ATLAS*, taksonomi ancaman AI, strategi mitigasi

ABSTRACT

The integration of Artificial Intelligence (AI) across various sectors has expanded the cyberattack surface, making threats to AI systems increasingly evolve into multi-stage attacks that require a structured framework for effective identification and mitigation. This study analyzes the types and chained attack patterns within the AI ecosystem using the Cybersecurity Attack Dataset 2025 and formulates mitigation strategies based on the MITRE ATLAS framework. From 14,133 attack records, 1,492 AI-related records were identified, comprising 502 AI Data Leakage & Privacy Risks, 496 AI/ML Security, and 494 AI Agents & LLM Exploits records. Reconstruction of the Attack Steps revealed three dominant attack patterns: Data Exfiltration via Prompt Injection 10 steps, Intercept WebView Cookies 12 steps, and Prompt-In-Attribute Trick 7 steps. The mapping results showed that Data Exfiltration via Prompt Injection was fully mapped to three AML techniques with six official mitigations, Intercept WebView Cookies had no corresponding MITRE ATLAS technique because it targeted the application layer, whereas Prompt-In-Attribute Trick was only partially mapped due to DOM XSS web application vulnerabilities. These findings demonstrate that MITRE ATLAS effectively supports the identification of AI-related attack techniques and the development of layered mitigation strategies.

Keywords: AI ecosystem, chained attacks, MITRE ATLAS, mitigation strategy, AI threat taxonomy