TÊN ĐỀ TÀI – NGHIÊN CỨU VÀ XÂY DỰNG KHUNG HỌC SÂU TUẦN TỰ: NÂNG CAO TÍNH BỀN VỮNG VÀ TÍNH KIÊN CƯỜNG CHO HỆ THỐNG PHÁT HIỆN XÂM NHẬP

Nguyễn Duy Hải - 250202007

Tóm tắt

- Lớp: CS2205.RM
- Link Github: https://github.com/haindtdt/CS2205.RM
- Link YouTube video: https://www.youtube.com/watch?v=Bty2G0HX7ck&t=43s

Nguyễn Duy Hải - 250202007

Giới thiệu

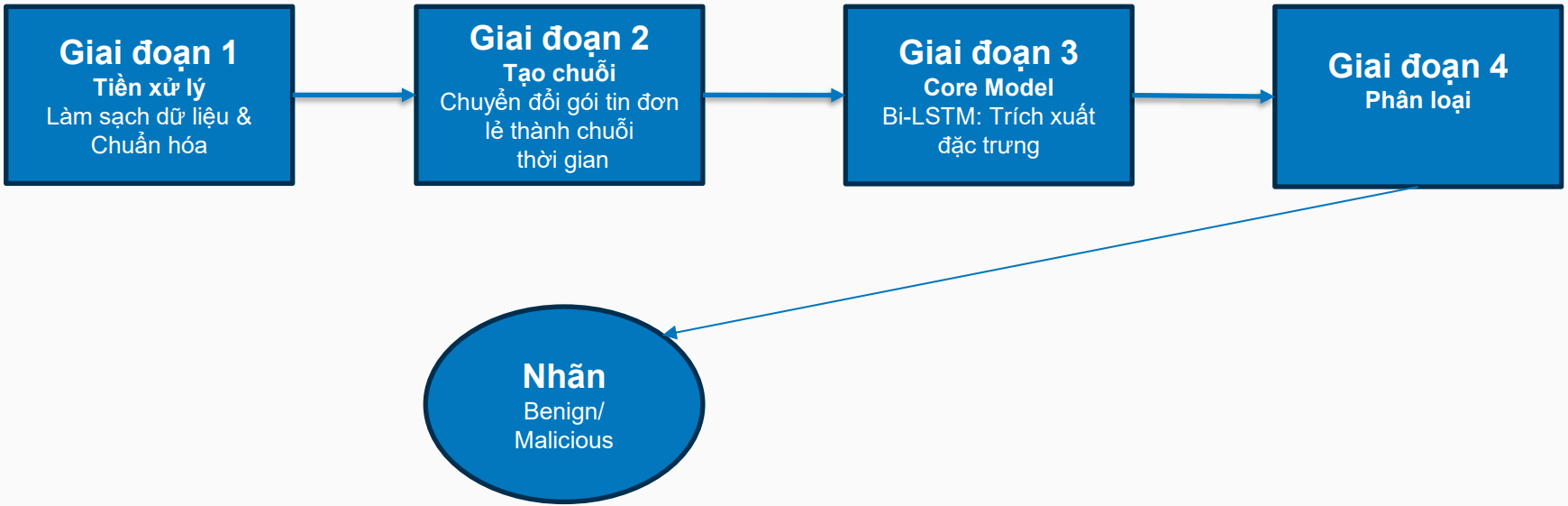
- **Bối cảnh:** Các cuộc tấn công mạng hiện đại (APT, Malware đa hình) ngày càng tinh vi, diễn ra âm thầm trong thời gian dài.
- **Vấn đề tồn tại:**
  - ✓ **Thiếu ngữ cảnh:** Các mô hình NIDS cũ xử lý gói tin rời rạc, không thấy được sự liên kết chuỗi thời gian (Sequential patterns).
  - ✓ **Lỗ hổng AI:** Các mô hình Học sâu (Deep Learning) rất dễ bị đánh lừa bởi các mẫu đối kháng (Adversarial Examples) – chỉ cần thêm nhiễu nhỏ vào gói tin là AI dự đoán sai.
- **Giải pháp:** Cần một hệ thống "Sequential" (để hiểu chuỗi) và "Resilient" (để chống tấn công).

Mục tiêu

- **Xây dựng mô hình Học sâu tuần tự:** Thiết kế kiến trúc mạng nơ-ron hồi quy hai chiều (Bi-Directional LSTM) để phân loại lưu lượng mạng dựa trên ngữ cảnh quá khứ và tương lai.
- **Phát triển cơ chế phòng thủ:** Tích hợp kỹ thuật Huấn luyện đối kháng (Adversarial Training) để mô hình có khả năng nhận diện ngay cả khi dữ liệu đầu vào bị kẻ tấn công làm nhiễu.
- **Đánh giá thực nghiệm:** Kiểm chứng hiệu năng trên tập dữ liệu chuẩn CIC-IDS-2017 và đánh giá khả năng "sống sót" của hệ thống trước các thuật toán tấn công như FGSM.

Nội dung và Phương pháp – Kiến trúc tổng thể

- **Mô hình đề xuất gồm 4 giai đoạn:**



Nội dung và Phương pháp – Chiến lược phòng thủ

- **Vấn đề:** Hacker dùng thuật toán FGSM (Fast Gradient Sign Method) để tạo mẫu đối kháng nhằm qua mặt NIDS.
- **Phương pháp tăng cường tính Kiên cường (Resilience):**

  ✓ **Bước 1:** Giả lập tấn công - Sử dụng thư viện *Adversarial Robustness Toolbox (ART)* để sinh ra tập dữ liệu đối kháng từ tập gốc.

  ✓ **Bước 2:** Adversarial Training - Trộn tập dữ liệu đối kháng vào tập huấn luyện.

  ✓ **Bước 3:** Retrain - Huấn luyện lại mô hình để nó "học" được hình dạng của các mẫu tấn công đã bị méo mó.

Nội dung và Phương pháp – Công cụ & Dữ liệu

- **Tập dữ liệu: CIC-IDS-2017** (Canadian Institute for Cybersecurity).
  - ✓ Bao gồm đầy đủ các loại tấn công: DDoS, PortScan, Botnet, Web Attack
- **Công nghệ triển khai:**
  - ✓ Ngôn ngữ: **Python 3.9**.
  - ✓ Deep Learning Framework: **TensorFlow/Keras** hoặc **PyTorch**.
  - ✓ Thư viện bảo mật AI: **IBM ART** hoặc **CleverHans**.
  - ✓ Môi trường: **Google Colab Pro** (Tesla T4 GPU).

Kết quả dự kiến

- **Về độ chính xác (Robustness):**
  - ✓ Mô hình đạt **F1-Score > 95%** trên tập dữ liệu gốc.
  - ✓ Vượt trội hơn các mô hình truyền thống (SVM, RF) trong việc phát hiện các tấn công chuỗi.
- **Về tính kiên cường (Resilience):**
  - ✓ Khi bị tấn công bằng FGSM: Độ chính xác của mô hình đề xuất duy trì ở mức **> 85%**.
  - ✓ (Trong khi đó, mô hình thường dự kiến sẽ giảm xuống dưới 50% - bị vô hiệu hóa hoàn toàn).
- **Sản phẩm:** Mã nguồn (Source code) hoàn chỉnh và Báo cáo phân tích chi tiết.

Tài liệu tham khảo

[1]. Soumyadeep Hore, Jalal Ghadermazi, Ankit Shah, Nathaniel D. Bastian: A sequential deep learning framework for a robust and resilient network intrusion detection system. Comput. Secur. 144: 103928 (2024)

[2]. Zongwen Yu, Ping Yi, Jiacheng Ye, Wei Wang, Yuming Qu: NIDS-Vis: Improving the generalized adversarial robustness of network intrusion detection systems. Comput. Secur. 139: 103700 (2024)

[3]. Vivek Kumar, Kamal Kumar, Maheep Singh, Neeraj Kumar: NIDS-DA: Detecting functionally preserved adversarial examples for network intrusion detection system using deep autoencoders. Expert Syst. Appl. 270: 126513 (2025)

[4]. Mohamed Amine Ferrag, Leandros A. Maglaras: Constraining Adversarial Attacks on Network Intrusion Detection Systems: Transferability and Defense Analysis. IEEE Access 12: 24500-24515 (2024)

[5]. S. M. Udin, M. A. Al-Garadi: SiamIDS: A Novel Cloud-Centric Siamese Bi-LSTM Framework for Interpretable Intrusion Detection in Large-Scale IoT Networks. IEEE Internet of Things Journal (Early Access) (2025)