


# THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút):  
<https://youtu.be/SXOHuV5y3XM>
- Link báo cáo đề cương(dạng .pdf đặt trên Github):  
<https://github.com/huuquyen2606/CS2205.CH1702.APR2023/DeCuongNghienCuu.pdf>
- Link slides (dạng .pdf đặt trên Github):  
<https://github.com/huuquyen2606/CS2205.CH1702.APR2023/SlideBaoCaoDeCuong.pdf>

|                                                                                                                                                                                              |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ul style="list-style-type: none"><li>● Họ và Tên: Nguyễn Hữu Quyền</li><li>● MSHV: CH220202021</li></ul>  | <ul style="list-style-type: none"><li>● Lớp: CS2205.CH1702.APR2023</li><li>● Tự đánh giá (điểm tổng kết môn): 9/10</li><li>● Số buổi vắng: 0</li><li>● Số câu hỏi QT cá nhân: 9</li><li>● Số câu hỏi QT của cả nhóm: 3</li><li>● Link Github:<br/><a href="https://github.com/huuquyen2606/CS2205.CH1702.APR2023/">https://github.com/huuquyen2606/CS2205.CH1702.APR2023/</a></li><li>● Mô tả công việc và đóng góp của cá nhân cho kết quả của nhóm:<ul style="list-style-type: none"><li>○ Đề xuất và lên ý tưởng cải tiến bài toán</li><li>○ Viết phần đề cương nghiên cứu</li><li>○ Làm slide báo cáo</li><li>○ Làm Poster</li><li>○ Làm video YouTube</li></ul></li></ul> |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

# ĐỀ CƯƠNG NGHIÊN CỨU

## TÊN ĐỀ TÀI (IN HOA)

HỆ THỐNG PHÁT HIỆN XÂM NHẬP PHI TẬP TRUNG KHẢ DIỄN GIẢI HỖ TRỢ HỌC TIỆM TIẾN

## TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

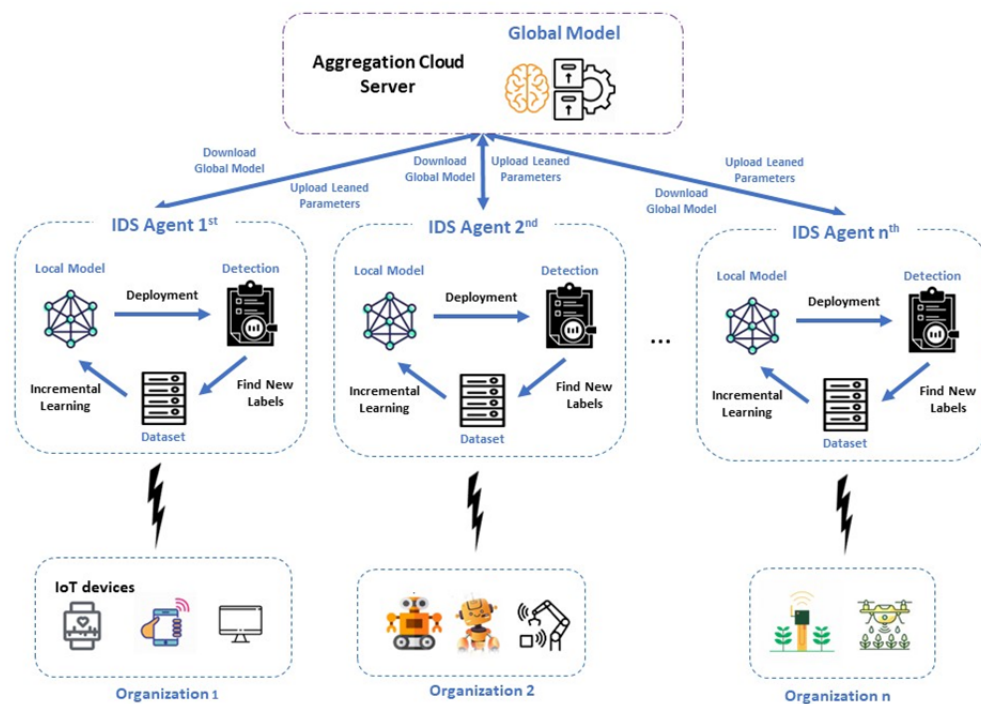
FedXI-IDS: AN EXPLAINABLE FEDERATED INCREMENTAL LEARNING APPROACH FOR INTRUSION DETECTION SYSTEM

## TÓM TẮT (Tối đa 400 từ)

Học cộng tác (Federated Learning - FL) đã thu hút được sự chú ý của các nhà nghiên cứu thông qua việc huấn luyện các mô hình học máy (Machine Learning - ML) với bộ dữ liệu phi tập trung. Đặc biệt, trong ngữ cảnh an toàn thông tin của các mạng thiết bị IoTs, các hệ thống phát hiện xâm nhập mạng phi tập trung (Federated Intrusion Detection System - FedIDS) được quan tâm và phát triển nhằm bảo vệ quyền riêng tư của dữ liệu cho các bên tham gia. Tuy nhiên, hầu hết các hệ thống hiện có chưa đạt được hiệu quả cao khi áp dụng vào thực tế. Cụ thể, các FedIDS phải huấn luyện tiệm tiến (Incremental Learning) liên tục để có thể phát hiện được các lớp (classes) tấn công mới (zero-day). Nhưng các bên cộng tác thường không đủ khả năng lưu trữ được lớp dữ liệu mới lẫn lớp dữ liệu cũ, dẫn tới việc tái huấn luyện mô hình nhưng lại lãng quên kiến thức cũ (Catastrophic forgetting - CF). Hơn nữa, các bên cộng tác mới tham gia vào quy trình huấn luyện cộng tác mà chưa có các lớp dữ liệu mới chưa từng thấy hay các bên huấn luyện với dữ liệu mất cân bằng (NonIID) sẽ làm trầm trọng thêm tình trạng CF của mô hình toàn cục. Để giải quyết những thách thức, chúng tôi đề xuất một hệ thống phát hiện xâm nhập mạng phi tập trung khả diễn giải hỗ trợ học tiệm tiến (**FedXI-IDS**) để giảm sự ảnh hưởng của CF lên hiệu suất hệ thống. Đặc biệt, để giải quyết tình trạng quên cục bộ do mất cân bằng lớp tại các bên, chúng tôi áp dụng học khả diễn giải (Explainable Artificial Intelligence - XAI) vào trong huấn luyện mô hình cục bộ giúp chất lọc quan hệ ngữ nghĩa giữa lớp và giá trị mất mát để cân bằng sự lãng quên của các lớp dữ liệu cũ với các lớp dữ liệu mới. Hệ thống FedXI-IDS được chúng tôi thực nghiệm và so sánh với các hệ thống mới hiện có. Các thực nghiệm dựa trên ba bộ dữ liệu CSE-CIC-IDS2018, ToN-IoT, Bot-IoT là các bộ dữ liệu mới nhất và mạng tin cậy.

## GIỚI THIỆU (Tối đa 1 trang A4)

Sự phát triển nhanh chóng trong lĩnh vực Internet và IoT đã dẫn đến sự bùng nổ và gia tăng đáng kể về quy mô mạng cũng như dữ liệu mạng. Tuy nhiên, dữ liệu mạng chứa đựng nhiều thông tin nhạy cảm nên sự bùng nổ cũng mang đến nhiều thách thức trong việc bảo mật lượng dữ liệu mạng tránh khỏi bị kẻ gian đánh cắp. Ngày càng nhiều các IDS được xây dựng giúp phát hiện và ngăn chặn các hành động tấn công và xâm nhập mạng. Đặc biệt, các FedIDS được nghiên cứu nhằm tận dụng nguồn tài nguyên dữ liệu để xây dựng và huấn luyện mô hình mà vẫn đảm bảo được tính riêng tư cho dữ liệu của các bên tham gia. Cụ thể, một số nghiên cứu gần đây [1][2] đã chỉ ra được sự hiệu quả của mô hình FedIDS trong phát hiện xâm nhập mạng. Tuy nhiên, lượng lớn các tham số trong mạng nơron làm cho các mô hình IDS áp dụng mạng nơron hiện nay chưa thể giải thích được lý do cho những quyết định, hành động được đưa ra. Việc giải thích các dự đoán của mô hình đóng vai trò quan trọng tăng độ tin cậy của mô hình. Mặt khác, các lớp tấn công mới (zero-day) xuất hiện ngày càng nhiều gây ra khó khăn lớn các mô hình phân loại chưa được huấn luyện đầy đủ. Hơn nữa, dữ liệu mạng có số lượng rất lớn và đang gia tăng mỗi ngày tạo ra áp lực lớn lên các hệ thống lưu trữ. Bên cạnh đó, các bên tham gia trong huấn luyện cộng tác thương không đảm bảo dữ liệu huấn luyện cân bằng về số lớp. Trong nghiên cứu của mình, Jiahua Dong [3] cũng đã chỉ ra được ảnh hưởng của dữ liệu và các lớp dữ liệu mới ở các bên cộng tác lên hiệu suất của mô hình chung. Nhóm tác giả cũng đã đề xuất phương pháp huấn luyện GLFC giúp giải quyết vấn đề trên mà không cần gia tăng thêm tài nguyên phần cứng cho quá trình lưu trữ.



**Hình 1: Hệ thống FedXI-IDS**

Cùng với đó, qua khảo sát các nghiên cứu liên quan trong lĩnh vực an toàn thông tin và cụ thể là bài toán phát hiện xâm nhập, chúng tôi nhận thấy rằng vẫn chưa có nghiên cứu nào giải quyết được hết các vấn đề nêu trên. Vì vậy, trong định hướng nghiên cứu này, chúng tôi tập trung nghiên cứu xây dựng một hệ thống phát hiện xâm nhập phi tập trung khả diễn giải hỗ trợ học tiệm tiến (**FedXI-IDS**), giải quyết các vấn đề kể trên. Đầu vào (**Input**) của hệ thống sẽ là các dữ liệu mạng X và đầu ra (**Output**) của hệ thống là lớp tấn công tương ứng của dữ liệu mạng X. Cụ thể hơn, hệ thống được mô tả như ở **Hình 1**.

## MỤC TIÊU

*(Viết trong vòng 3 mục tiêu, lưu ý về tính khả thi và có thể đánh giá được)*

- ❖ Xây dựng và đánh giá hiệu năng của mô hình phát hiện xâm nhập khả diễn giải tập trung (**XAI-based IDS**) dựa trên ba bộ dữ liệu CSE-CIC-IDS2018[4], ToN-IoT[5], Bot-IoT[6].
- ❖ Xây dựng và đánh giá hiệu năng của mô hình phát hiện xâm nhập khả diễn giải phi tập trung (**XAI-based FedIDS**) so với mô hình **XAI-based IDS** dựa trên ba bộ dữ liệu như trên.
- ❖ Đề xuất mô hình phát hiện xâm nhập phi tập trung khả diễn giải (**FedXI-IDS**) và so sánh hiệu suất với mô hình **XAI-based FedIDS** dựa trên ba bộ dữ liệu như trên.

## NỘI DUNG VÀ PHƯƠNG PHÁP

*(Viết nội dung và phương pháp thực hiện để đạt được các mục tiêu đã nêu)*

**Nội dung 1:** Nghiên cứu xây dựng mô hình phát hiện xâm nhập khả diễn giải.

### Mục tiêu:

- ❖ Tiến hành nghiên cứu và xây dựng mô hình phát hiện xâm nhập khả diễn giải dựa trên học máy và học khả diễn giải (**XAI-based IDS**). Nhằm diễn giải được mức độ ảnh hưởng của các đặc trưng dữ liệu lên hiệu suất mô hình.

### Phương pháp:

- ❖ Chúng tôi tập trung vào việc xây dựng mô hình học sâu và học khả diễn giải trên 3 bộ dữ liệu CSE-CIC-IDS2018, ToN-IoT, Bot-IoT. Bên cạnh đó, để kết quả được tối ưu, các mô hình suy luận ngôn ngữ này sẽ được chúng tôi nghiên cứu kết hợp với các mô hình biểu diễn từ (word representations) như Word2Vec, n-grams, BERT để trích xuất các đặc trưng về ngữ cảnh và cấu trúc của dữ liệu mạng.
- ❖ Kiểm tra đánh giá hiệu suất của mô hình được nghiên cứu. Với độ chính xác kì vọng của mô hình là trên 95%

**Nội dung 2:** Nghiên cứu đề xuất mô hình phát hiện xâm nhập phi tập trung khả diễn giải.

**Mục tiêu:**

- ❖ Tiến hành nghiên cứu và đề xuất mô hình phát hiện xâm nhập phi tập trung khả diễn giải (**XAI-based FedIDS**) dựa trên mô hình **XAI-based IDS** và **học cộng tác**. Nhằm giải quyết được các vấn đề liên quan đến bảo mật thông tin cho dữ liệu huấn luyện, cũng như giảm bớt được chi phí trong quá trình huấn luyện.

**Phương pháp:**

- ❖ Với mô hình phát hiện xâm nhập khả diễn giải tập trung như ở Nội dung 1. Tập trung thử nghiệm mô hình trên với phương pháp học cộng tác và áp dụng công thức **FedAVG**[7] tổng hợp mô hình chung.
- ❖ Các kết quả thực nghiệm cũng được đánh giá trên 3 bộ dữ liệu CSE-CIC-IDS2018, ToN-IoT, Bot-IoT.
- ❖ Kiểm tra đánh giá hiệu suất của mô hình được đề xuất. Với độ chính xác kì vọng của mô hình là trên 95%
- ❖ Qua đó, so sánh đánh giá hiệu suất mang lại giữa mô hình huấn luyện tập trung và phi tập trung.

**Nội dung 3:** Nghiên cứu mô hình phát hiện xâm nhập khả diễn giải phi tập trung hỗ trợ học tiệm tiến.

**Mục tiêu:**

- ❖ Tiến hành nghiên cứu và đề xuất mô hình phát hiện xâm nhập phi tập trung khả diễn giải dựa trên các phương pháp học sâu hỗ trợ **học tiệm tiến (FedXI-IDS)**. Nhằm nâng cao hiệu suất của mô hình trong việc phát hiện các “lớp tấn công mới”. Hơn nữa, phương pháp này sẽ giải quyết được ảnh hưởng của lượng lớn dữ liệu khi liên tục cập nhật các dữ liệu mới, mà không làm tăng chi phí lưu trữ, cũng như tính toán trong quá trình tái huấn luyện.

**Phương pháp:**

- ❖ Với mô hình phát hiện xâm nhập khả diễn giải phi tập trung như ở Nội dung 2. Tập trung nghiên cứu và thử nghiệm huấn luyện mô hình trên với phương pháp học tiệm tiến.
- ❖ Các kết quả thực nghiệm cũng được đánh giá trên 3 bộ dữ liệu CSE-CIC-IDS2018, ToN-IoT, Bot-IoT.
- ❖ Kiểm tra đánh giá hiệu suất của phương pháp được đề xuất. Với độ chính xác kì vọng của mô hình là trên 95%
- ❖ Qua đó, so sánh đánh giá hiệu suất mang lại giữa **FedXI-IDS** được đề xuất so với phương pháp **GLFC**[3] và một số phương pháp State-of-the-art SOTA khác.

## KẾT QUẢ MONG ĐỢI

*(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)*

### Dự kiến kết quả nghiên cứu:

- ❖ Xây dựng hoàn thiện một hệ thống phát hiện xâm nhập phi tập trung khả diễn giải hỗ trợ học tiệm tiến đạt được hiệu suất phân loại cao với độ chính xác trên 95%.

### Báo cáo kết quả nghiên cứu:

- ❖ Dự kiến công bố bài báo khoa học tại Hội nghị chuyên ngành quốc tế Hạng B (theo CORE2021): 01 bài (Với nội dung 1 và 2).
- ❖ Dự kiến công bố bài báo khoa học tại Tạp chí chuyên ngành quốc tế Hạng Q1: 01 bài (Với nội dung 1, 2 và 3).

## TÀI LIỆU THAM KHẢO *(Định dạng DBLP)*

- [1]. Bimal Ghimire ,Danda B. Rawat: Recent Advances on Federated Learning for Cybersecurity and Cybersecurity for Federated Learning for Internet of Things. IEEE Internet of Things Journal 2022: 8229 - 8249
- [2]. Othmane Friha, Mohamed Amine Ferrag, Lei Shu, Leandros Maglaras, Kim-Kwang Raymond Choo, Mehdi Nafaa: FELIDS: Federated learning-based intrusion detection system for agricultural Internet of Things. Journal of Parallel and Distributed Computing 2022: 17-31
- [3]. Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, Qi Zhu: Federated Class-Incremental Learning. CVPR 2022: 10164-10173
- [4]. Iman Sharafaldin, Arash Habibi Lashkari, Ali A. Ghorbani: Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. ICISSP 2018: 108-116
- [5]. Abdullah Alsaedi, Nour Moustafa, Zahir Tari, Abdun Mahmood, Adnan Anwar: TON\_IoT Telemetry Dataset: A New Generation Dataset of IoT and IIoT for Data-Driven Intrusion Detection Systems. IEEE Access 2020: 165130-165150
- [6]. Jared M. Peterson, Joffrey L. Leevy, Taghi M. Khoshgoftaar: A Review and Analysis of the Bot-IoT Dataset. SOSE 2021: 20-27
- [7]. H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, Blaise Agueray Arcas: Communication-Efficient Learning of Deep Networks from Decentralized Data. AISTATS 2017:1273–1282