BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

Tống Anh Tuấn

NGHIÊN CỨU CẢI TIẾN MỘT SỐ MÔ HÌNH HỌC MÁY VÀ HỌC SÂU ÁP DỤNG CHO BÀI TOÁN PHÂN LOẠI DGA BOTNET

TÓM TẮT LUẬN ÁN TIẾN SĨ HỆ THỐNG THÔNG TIN

Hà Nội - 2023

Công trình được hoàn thành tại: Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam

Người hướng dẫn khoa học:

1. PGS. TS. Hoàng Việt Long, Trường Đại học Kỹ thuật - Hậu cần CAND, Bộ Công an.
2. PGS. TS. Nguyễn Việt Anh, Viện Công nghệ thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Phản biện 1: PGS. TS. Bùi Thu Lâm, Học viện Kỹ thuật mật mã, Ban Cơ yếu Chính phủ.
Phản biện 2: PGS. TS. Nguyễn Hà Nam, Đại học Điện lực, Bộ Công thương.
Phản biện 3: PGS. TS. Ngô Quốc Tạo, Viện Công nghệ thông tin, Viện HL KH&CN VN.

Luận án được bảo vệ trước Hội đồng đánh giá luận án tiến sĩ cấp Học viện họp tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam vào hồi 09 giờ 00, ngày 09 tháng 08 năm 2023.

Có thể tìm hiểu luận án tại:
1. Thư viện Học viện Khoa học và Công nghệ
2. Thư viện Quốc gia Việt Nam

MỞ ĐẦU

1. Tính cấp thiết của luận án

Botnet là tập hợp các máy tính bị mã độc xâm nhập, được điều khiển và quản trị từ xa thông qua các máy chủ điều khiển [1]. Giải pháp phát hiện Botnet là câu hỏi luôn được các nhà khoa học đặt ra và quan tâm giải quyết.

Có hai hướng tiếp cận chính thường được sử dụng trong phát hiện Botnet, bao gồm [5]:

(1) Hướng tiếp cận dựa trên Honeynet (mạng bẫy).

(2) Hướng tiếp cận dựa trên hệ thống phát hiện xâm nhập, bao gồm:

- Kỹ thuật phát hiện Botnet dựa trên sự bất thường.
- Kỹ thuật phát hiện Botnet dựa trên chữ ký.
- Kỹ thuật phát hiện Botnet dựa trên tên miền.

Trong phạm vi của luận án, NCS tập trung nghiên cứu vào DGA Botnet. Một số kết quả nghiên cứu chuyên sâu về bài toán DGA Botnet đã được công bố như: Các kỹ thuật dựa trên phân tích lưu lượng mạng [6], [7], [8], [9], [10]; Các kỹ thuật sử dụng học máy: [11], [12], [13], [14]; Các kỹ thuật sử dụng học sâu: [15], [16], [17], [18], [19], [20].

Từ các vấn đề trên, NCS đặt ra các câu hỏi nghiên cứu cho luận án như sau: "Nghiên cứu cải tiến kỹ thuật như thế nào để tăng cường khả năng phân loại DGA Botnet dựa trên cách tiếp cận học máy, học sâu?".

2. Mục tiêu nghiên cứu

Đề tài đặt ra mục tiêu chính là nghiên cứu, cải tiến các mô hình học máy, học sâu để nâng cao độ chính xác của giải pháp phân loại DGA Botnet.

3. Đối tượng và phạm vi nghiên cứu

Nghiên cứu tập trung vào các đối tượng như sau:

- Đặc điểm, cơ chế, hành vi của DGA Botnet; kỹ thuật phát hiện, phân loại Botnet dựa trên tên miền.

- Bài toán phân lớp nhị phân và phân lớp đa lớp, tương ứng với phát hiện và phân loại DGA Botnet.

- Các bộ dữ liệu công khai, tin cậy và cập nhật về DGA Botnet cùng quy trình xây dựng bộ dữ liệu mới.

4. Nội dung và phương pháp nghiên cứu

a. Nôi dung nghiên cứu

Một số nội dung chi tiết mà NCS sẽ tập trung nghiên cứu như sau:

- Nghiên cứu đặc điểm, các kỹ thuật phát hiện và phân loại DGA Botnet;

- Nghiên cứu mạng LSTM, cơ chế Attention và các biến thể, trên cơ sở đó cải tiến, đề xuất mô hình học sâu mới để nâng cao hiệu quả phân loại DGA Botnet.

- Nghiên cứu về quy trình, tiêu chí, các bộ dữ liệu về DGA Botnet và áp dụng.

b. Phương pháp nghiên cứu

NCS sử dụng các phương pháp nghiên cứu bao gồm:

- Nghiên cứu lý thuyết;

- Tham khảo ý kiến chuyên gia;

- Nghiên cứu thực nghiệm, đánh giá.

5. Các đóng góp của luận án

Luận án có 02 đóng góp bao gồm:

- *Đóng góp 1*: Đề xuất cải tiến kiến trúc lõi kết hợp BiLSTM với cơ chế Attention và sử dụng trong xây dựng mô hình LA_Bin07 để phát hiện và mô hình LA_Mul07 để phân loại DGA Botnet với độ chính xác được cải thiện.

- *Đóng góp 2*: Hoàn thiện bổ sung quy trình xây dựng tập dữ liệu mẫu và đề xuất bộ dữ liệu UTL_DGA22 được mô tả và gán nhãn, phục vụ phân loại DGA Botnet.

6. Bố cục của luận án

Nội dung luận án được cấu trúc thành 04 chương, cụ thể như sau:

- Chương 1: Cơ sở lý thuyết về DGA Botnet
- Chương 2: Phát hiện DGA Botnet sử dụng NCM và học máy
- Chương 3:Phát hiện và phân loại DGA Botnet sử dụng học sâu.
- Chương 4: Quy trình xây dựng và bộ dữ liệu mới UTL_DGA22 cho bài toán DGA Botnet.

Các kết quả nghiên cứu của luận án được công bố tại 04 bài báo trên tạp chí khoa học chuyên ngành quốc tế thuộc danh mục SCIE/Scopus, 01 báo cáo tại hội thảo khoa học chuyên ngành quốc gia và 01 báo cáo tại hội thảo khoa học chuyên ngành quốc tế uy tín, được liệt kê trong phần "Danh mục các công trình công bố liên quan đến luận án" ở cuối của luận án này.

CHƯƠNG 1. CƠ SỞ LÝ THUYẾT VỀ DGA BOTNET

1.1. Tổng quan chung về Botnet

1.1.1. Khái niệm Botnet

Theo Provos & Holz, Botnet là một "mạng gồm rất nhiều máy tính bị xâm nhập và có thể bị kẻ tấn công điều khiển từ xa".

1.1.2. Các bước phát triển về công nghệ Botnet

1.1.3. Một số đặc điểm của Botnet

Botnet có những đặc trưng riêng về vòng đời hoạt động, phương thức lây nhiễm và các hành vi độc hại.

1.1.4. Phân loại Botnet

Botnet có thể được phân loại theo các tiêu chí như: Giao thức, thiết bị lây nhiễm hoặc kiến trúc.

1.2. Kỹ thuật phát hiện Botnet

Có kỹ thuật chính được sử dụng để phát hiện Botnet:

- (1) Các kỹ thuật dựa trên honeynet.
- (2) Các kỹ thuật dựa trên hệ thống phát hiện xâm nhập:
  + Phát hiện Botnet dựa trên sự bất thường.
  + Phát hiện Botnet dựa trên chữ ký.
  + Phát hiện Botnet dựa trên tên miền.

1.3. Bài toán DGA Botnet

1.3.1. Khái quát về DGA Botnet

DGA Botnet là khái niệm chỉ một dạng Botnet được triển khai theo mô hình Client-Server. Trong đó, các Bot đóng vai trò là Client sẽ liên kết trở lại máy chủ C&C - đóng vai trò là Server - thông qua các tên miền DNS được sinh một cách tự động và được thống nhất trước đó nhằm qua mặt các hệ thống bảo mật.

1.3.2. Bài toán phát hiện DGA Botnet

Là bài toán với mục tiêu phát hiện các tên miền được sinh ra bởi DGA Botnet so với tên miền lành tính, dữ liệu gồm hai nhãn 0 và 1.

1.3.3. Bài toán phân loại DGA Botnet

Là bài toán nhằm mục tiêu xác định họ của DGA Botnet, dữ liệu gồm có n nhãn, tương ứng với n họ DGA Botnet được xem xét.

1.3.4. Phân biệt với bài toán phát hiện URL giả mạo

Bài toán phát hiện DGA Botnet có sự khác biệt so với bài toán phát hiện URL giả mạo.

1.3.5. Bộ dữ liệu đánh giá cho bài toán DGA Botnet

NCS lựa chọn 04 bộ dữ liệu bao gồm: Andrey Abakumov's DGA Repository [35], OSINT DGA feed [36], UMUDGA Dataset [13] và 360NetLab Dataset [37] (Bảng 1.4).

Bảng 1.4. Mô tả về 04 bộ dữ liệu DGA Botnet được sử dụng trong các đánh giá

	Phát hiện DGA Botnet	Phân loại DGA Botnet	Số mẫu lành tính	Số mẫu DGA Botnet	Số họ DGA Botnet
AADR	✓	✓	1.000.000	801.667	08
OSINT	✓	×	1.000.000	495.186	
UMUDGA	✓	✓	1.000.000	500.000	50
360NL	✓	×	1.000.000	1.513.524	

1.3.6. Thông số đánh giá bài toán

NCS đánh giá qua các tham số gồm Accuracy, Precision, Recall và F_1-score.

1.3.7. Ý nghĩa bài toán DGA Botnet

Vận dụng cơ chế hoạt động của DGA Botnet có thể mang lại một giải pháp hiệu quả và mang lại nhiều ưu điểm như không đòi hỏi quá nhiều năng lực thu thập và xử lý của hệ thống; việc phát hiện hoạt động của DGA Botnet có thể diễn kể ra khi chúng đã lây nhiễm vào thiết bị.

1.4. Một số nghiên cứu giải quyết bài toán DGA Botnet

- Hướng tiếp cận sử dụng các kỹ thuật phân tích DNS: Alieyan và cộng sự [6], Kwon và cộng sự [7], Wang và cộng sự [8], Chowdhury và cộng sự [38], Bisio và cộng sự [9], Wang và cộng sự [40], Trung và cộng sự [10].

- Hướng tiếp cận dựa trên học máy: Hiếu và cộng sự [11], Khan và cộng sự [12], Zago và cộng sự [13], Xuân và cộng sự [14], Suryotrisongko và cộng sự [45], Zhao và cộng sự [46], Alauthman và cộng sự [47].

- Hướng tiếp cận dựa trên học sâu: Đức và cộng sự [15], Curtin và cộng sự [16], Qiao và cộng sự [17], Namgung và cộng sự [19], Vinayakumar và cộng sự [20], Liu và cộng sự [51].

1.5. Kết luận Chương 1

Một phần kết quả trình bày tại Chương 1 được công bố tại [CT2] [CT6] trong Danh mục các công trình công bố liên quan đến luận án.

CHƯƠNG 2. PHÁT HIỆN DGA BOTNET SỬ DỤNG NCM VÀ HỌC MÁY

2.1. Phát hiện DGA Botnet sử dụng NCM

2.1.1. Thuật toán NCM

Tập mờ trung lập – Neutrosophic Set được Smarandache đề xuất [52], là một cải tiến của tập mờ truyền thống. Trên không gian X, một tập mờ trung lập A được định nghĩa như sau:

$$A = \{x, \big(T_A(x), I_A(x), F_A(x)\big) : x \in X\}$$

Trong đó, hàm $T_A(x), I_A(x), F_A(x)$ lần lượt thể hiện độ thuộc của phần tử x gồm thuộc về, trung lập và không thuộc về một tập xác định nào đó. Các giá trị $T_A(x), I_A(x), F_A(x) \in [0, 1]$ và thỏa mãn điều kiện:

$$0 \leq T_A(x) + I_A(x) + F_A(x) \leq 3$$

Neutrosophic C-Means - NCM là thuật toán phân cụm mờ trên tập mờ trung lập, được đề xuất bởi Gou và cộng sự [53], được tóm tắt như sau:

Thuật toán: $NCM(X, \varepsilon)$	
Dữ liệu vào	X, ε
Dữ liệu ra	k
Khởi tạo $T^{(0)}, I^{(0)}, F^{(0)}$	
Khởi tạo $C, m, \varepsilon, \delta,\ \omega_1,\ \omega_2,\ \omega_3$	
Lặp:	
	Tính $c_j^{(k)}$
	Tính $\bar{c}_{i_{max}}$
	Cập nhật $T^{(\mathrm{k+1})}$
	Cập nhật $I^{(\mathrm{k+1})}$
	Cập nhật $F^{(\mathrm{k+1})}$
Điều kiện lặp	$\left\|T_{ij}^{(k+1)} - T_{ij}^{(k)}\right\| > \varepsilon$
Gán mỗi dữ liệu vào lớp với giá trị $TM = [T, I, F]$ lớn nhất. $x_i \epsilon\ k^{th}$ nếu $k = argmax\big(TM_{ij}\big)\ với\ j = 1, 2, \ldots., C + 2$	

2.1.2. Áp dụng NCM để phát hiện DGA Botnet

NCS áp dụng thuật toán NCM qua hai bước như sau:

(1) Lựa chọn đặc trưng: NCS đề xuất và lựa chọn một số đặc trưng cơ bản của tên miền. Các đặc trưng được vector hóa phù hợp để làm đầu vào cho thuật toán NCM.

Kết quả lựa chọn đặc trưng trên 04 bộ dữ liệu được liệt kê tại Bảng 2.2.

Bảng 2.2. Các đặc trưng được lựa chọn làm đầu vào cho thuật toán NCM

STT	AADR	360NetLab	OSINT	UMUDGA
1	CIPA	RCC	DNL	RCC
2	HVTLD	VR	ND	VR
3	VR	Entropy	VR	Entropy
4	RCC	ND	RCC	ND
5	ND	DNL	Entropy	DNL
6	Entropy	NR	NR	NR
7	RCN	CD	CD	CD

(2) Phân cụm và gán nhãn: Sử dụng thuật toán NCM để chia các điểm dữ liệu thành ba cụm, sau đó tiến hành gán nhãn đại diện cho các cụm tương ứng là DGA Botnet, Lành tính và Nhiễu.

Lần lượt đánh giá trên các bộ dữ liệu AADR, 360NL, OSINT và UMUDGA. Kết quả được thể hiện tại Bảng 2.3.

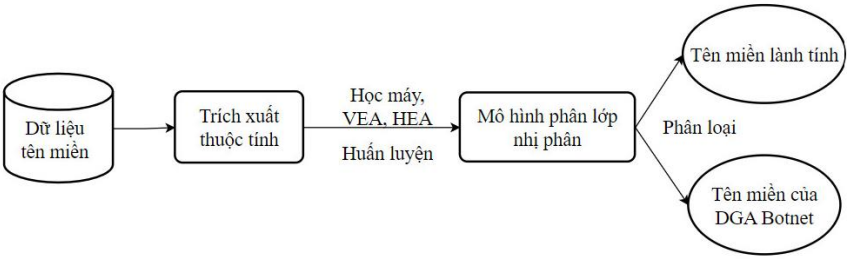
Bảng 2.3. Kết quả phát hiện DGA Botnet của thuật toán NCM trên 04 bộ dữ liệu

	A.Precision	A.Recall	A.F_1-Score
AADR	0,87	0,76	0,79
360NL	0,87	0,81	0,84
OSINT	0,77	0,61	0,54
UMUDGA	0,87	0,81	0,84

2.2. Phát hiện DGA Botnet sử dụng học máy

2.2.1. Mô hình đánh giá thuật toán học máy

Các giai đoạn trong quá trình đánh giá thuật toán học máy được thể hiện ở Hình 2.6.



Hình 2.6. Sơ đồ mô hình huấn luyện, đánh giá

Theo đó, với dữ liệu đầu vào là các tên miền, bao gồm cả lành tính và độc hại đã được gán nhãn, sử dụng n-gram để tách các tên miền và kỹ thuật TF-IDF để biểu diễn các đặc trưng.

- Đối với học máy, NCS sử dụng các thuật toán sau: Support Vector Machines - SVM, Logistic Regression - LR, Naive Bayes - NB, Neural Networks - NN, Decision Trees - DT, Random Forests - RF, k-Nearest Neighbour -k-NN và Adaptive Boosting - AB.

- Đối với mô hình học kết hợp dựa trên bình chọn, NCS đề xuất hai mô hình VEA và HEA.

Đánh giá được thực hiện trên bộ dữ liệu UMUDGA Dataset.

2.2.2. Kết quả phát hiện DGA Botnet của các mô hình học máy

Bảng 2.5 thể hiện kết quả các độ đo Precision, Recall và F_1-score khi sử dụng các thuật toán học máy để phát hiện DGA Botnet.

Bảng 2.5. Kết quả phát hiện DGA Botnet sử dụng học máy trên bộ dữ liệu UMUDGA

Mô hình học máy	A.Precision	A.Recall	A.F_1-score
LR	0,97	0,97	0,97
NB	0,93	0,89	0,91
DT	0,93	0,95	0,94
NN	0,97	0,97	0,97
SVM	0,97	0,96	0,97
RF	0,74	0,82	0,77
k-NN	0,97	0,66	0,78
AB	0,83	0,85	0,84

Hầu hết các thuật toán học máy đạt được độ chính xác cao. Mô hình LR, NN, SVM cho kết quả tổng thể cao nhất với F_1-score đạt 0,97 và mô hình có kết quả thấp nhất là RF với 0,77.

2.2.3. Kết quả phát hiện DGA Botnet của mô hình học kết hợp

Kết quả của VEA và HEA được thể hiện tại Bảng 2.6.

Bảng 2.6. Kết quả phát hiện DGA Botnet của mô hình VEA và HEA trên bộ dữ liệu UMUDGA

Thuật toán	A.Precision	A.Recall	A.F_1-score
Trung bình các mô hình đơn	0,92	0,88	0,89
Neural Network	0,97	0,97	0,97
Random Forrest	0,74	0,82	0,77
VEA	0,98	0,99	0,98
HEA	0,97	0,97	0,97

Hạn chế của giải pháp NCM và học máy:

- Độ chính xác vẫn có thể tiếp tục được nâng cao;
- Đòi hỏi nhiều thời gian chạy huấn luyện bởi chạy trên CPU;
- Chưa phù hợp với bài toán phân lớp đa lớp.

2.2.4. Thời gian huấn luyện và đánh giá của các mô hình học máy

Mô hình học kết hợp cho cải thiện độ chính xác nhưng thời gian huấn luyện lâu hơn.

2.3. Kết luận Chương 2

Một phần kết quả trình bày tại Chương 2 được công bố tại [CT1] [CT3] trong Danh mục các công trình công bố liên quan đến luận án.

CHƯƠNG 3. PHÁT HIỆN VÀ PHÂN LOẠI DGA BOTNET SỬ DỤNG HỌC SÂU

3.1. Nền tảng kỹ thuật học sâu cho bài toán DGA Botnet

3.1.1. Mạng Recurrent Neural Network

Recurrent Neural Network - RNN hay mạng nơ-ron hồi quy, được thiết kế để huấn luyện với các dữ liệu đầu vào có dạng chuỗi (sequence/ time-series).

3.1.2. Mạng Long-Short Term Memory và biến thể

Mạng LSTM cải tiến so với RNN, biến thể BiLSTM có thể huấn luyện hai chiều.

3.1.3. Cơ chế Attention và biến thể

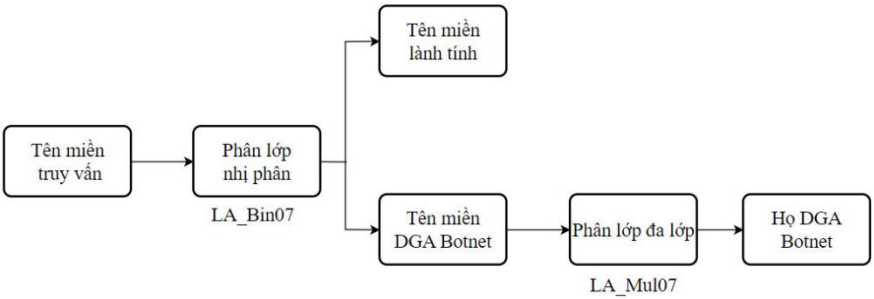
Cơ chế Attention và SelfAttention giúp tăng hiệu quả huấn luyện.

3.1.4. Mạng LSTM tích hợp Attention

Mạng LSTM tích hợp Attention giúp tăng hiệu quả huấn luyện với bài toán DGA Botnet

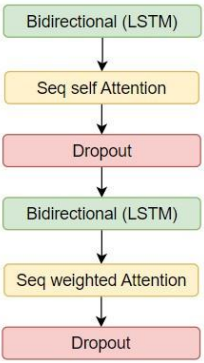
3.2. Đề xuất kiến trúc lõi và hai mô hình học sâu mới

3.2.1. Quy trình thực hiện bài toán DGA Botnet



Hình 3.11. Giải pháp phát hiện và phân loại DGA Botnet với hai mô hình học sâu mới LA_Bin07 và LA_Mul07

3.2.2. Đề xuất kiến trúc lõi của mô hình học sâu

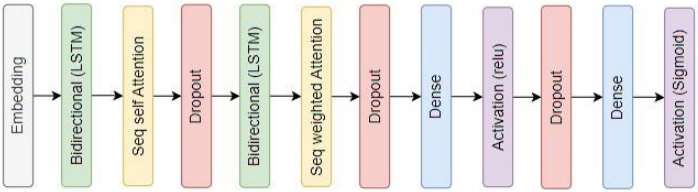


Hình 3.12. Kiến trúc lõi BiLSTM_SelfA_Double đề xuất

3.2.3. Xử lý dữ liệu đầu vào

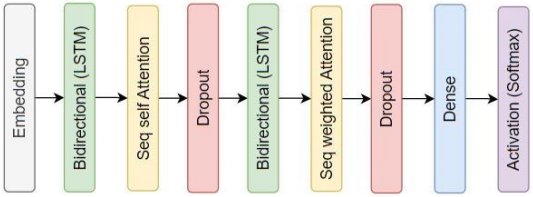
Gồm 2 bước Encoding và Word Embedding

3.2.4. Mô hình LA_Bin07 cho phát hiện DGA Botnet



Hình 3.13. Kiến trúc của mô hình LA_Bin07

3.2.5. Mô hình LA_Mul07 cho phân loại DGA Botnet



Hình 3.14. Cấu trúc đề xuất của mô hình LA_Mul07

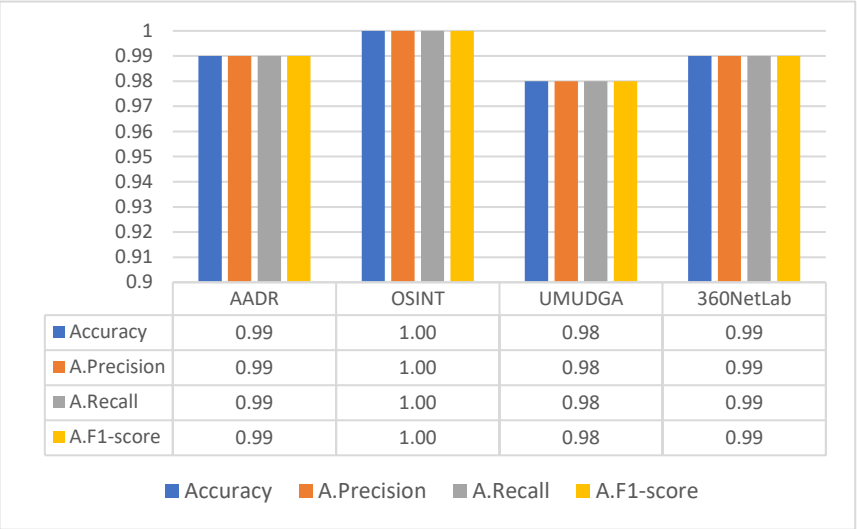
3.3. Đánh giá hai mô hình học sâu đề xuất

3.3.1. Bộ dữ liệu và môi trường đánh giá

Bảng 3.4. Ký hiệu các đánh giá hai mô hình học sâu đề xuất

	LA_Bin07	**LA_Mul07**
AADR	Đánh giá B1	Đánh giá M1
OSINT	Đánh giá B2	
UMUDGA	Đánh giá B3	Đánh giá M2
360NetLab	Đánh giá B4	

3.3.2. Đánh giá mô hình LA_Bin07 cho bài toán phát hiện DGA Botnet



Hình 3.15. Kết quả các đánh giá B1, B2, B3 và B4

3.3.3. Đánh giá mô hình LA_Mul07 cho bài toán phân loại DGA Botnet

Mô hình LA_Mul07 có độ chính xác cao trong phân lớp các họ DGA Botnet, kể cả trong trường hợp số lượng họ DGA Botnet cần phân lớp là nhiều, cụ thể đạt 1,00 trên bộ dữ liệu AADR và 0,86 trên bộ dữ liệu UMUDGA.

3.4. Đánh giá với các nghiên cứu liên quan

3.4.1. Đánh giá hai mô hình đề xuất trên bộ dữ liệu UMUDGA

Mô hình LA_Bin07 có kết quả phát hiện chính xác tốt hơn rất nhiều so với mô hình SVM. Đồng thời, cho kết quả gần như tương đương với các mô hình AB, NN, RF, DT hay kNN.

Mô hình LA_Mul07 có kết quả phân loại cho Accuracy cao hơn nhiều so với các mô hình học máy còn lại.

3.4.2. Đánh giá hai mô hình với một số kiến trúc học sâu khác

NCS cũng đồng thời đánh giá với một số kiến trúc học sâu khác mà NCS xây dựng trên cơ sở CNN và LSTM bao gồm: Basic CNN, Basic LSTM, Bi-LSTM và CNN-LSTM. Kết quả cho thấy, mô hình LA_Bin07 và LA_Mul07 đạt kết quả tốt nhất trong các mô hình thử nghiệm.

3.4.3. Đánh giá mô hình phân loại LA_Mul07 với một số mô hình liên quan

Trong phần này, NCS sử dụng mô hình LA_Mul07 để đánh giá với mô hình của Qiao và Namgung với cùng hướng tiếp cận LSTM kết hợp Attention. Các đánh giá được thực hiện trên cùng bộ dữ liệu mà các tác giả mô tả và công bố.

Kết quả thực nghiệm cho thấy, mô hình LA_Mul07 có A.F1-score được cải thiện 3% so với mô hình LSTM_AM của Qiao và cộng sự và cải thiện Accuracy là 1,03% so với mô hình BiLSTM_Attention và 0,38% so với mô hình CNN-BiLSTM_Ensemble của Namgung và cộng sự.

Mô hình LA_Mul07 cũng có ưu điểm là có khả năng nhận diện đúng các họ DGA Botnet (thông qua chỉ số Precision, Recall của từng nhãn) đồng đều hơn so với mô hình của Qiao và Namgung.

3.5. Kết luận Chương 3

Một phần kết quả trình bày tại Chương 3 được công bố tại [CT4] trong Danh mục các công trình công bố liên quan đến luận án.

CHƯƠNG 4. QUY TRÌNH XÂY DỰNG VÀ BỘ DỮ LIỆU MỚI UTL_DGA22 CHO BÀI TOÁN DGA BOTNET

4.1. Đặt vấn đề bộ dữ liệu DGA Botnet

4.1.1. Khái quát vấn đề

Giải pháp đề xuất trong các nghiên cứu trước đó thường được đánh giá trên những bộ dữ liệu do nhóm nghiên cứu thu thập vào những thời điểm khác nhau, số lượng mẫu không đồng đều, tính công bố rộng rãi không cao và thường không thuận tiện cho việc đối sánh.

4.1.2. Bộ dữ liệu về Botnet nói chung

Một số bộ dữ liệu về Botnet nói chung như CTU-13, UGR16, DreLAB, UNSW-NB15, ISCX-Bot-2014. Cả 5 bộ dữ liệu ở trên đều không được thiết kế để đánh giá chuyên biệt cho bài toán DGA Botnet bởi chúng thiếu tên miền của các họ DGA Botnet và nhãn tương ứng.

4.1.3. Bộ dữ liệu về DGA Botnet

Một số bộ dữ liệu về DGA Botnet như: Andrey Abakumov's DGA Repository, Johannes Bader's Domain Generation Algorithms Repository, Alexa Top 1 Milion Domains, Botnet DGA Dataset, UMUDGA Dataset, DGArchive by Fraunhofer FKIE, OSINT DGA feed, 360NetLab Dataset, The Majestic Million.

4.1.4. Đặt vấn đề nghiên cứu

Có sự khác nhau về cấu trúc và mục đích giữa các bộ dữ liệu cho Botnet nói chung và bộ dữ liệu cho DGA Botnet nói riêng. NCS phân nhóm và thể hiện chi tiết tại Bảng 4.4.

Bảng 4.4. Đánh giá về đặc điểm các nhóm bộ dữ liệu cho Botnet

Bộ dữ liệu	Nhóm	Phát hiện Botnet	Phát hiện	Phát hiện tấn	Lưu lượng mạng	Định dạng		
						PCAP	SCV	TXT

			DGA Botnet	**công lưu**				
CTU	Botnet	✓	×	×	✓	✓	✓	×
UGR	Botnet/IDS	✓	×	✓	✓	✓	✓	×
DLAB	Botnet	✓	×	×	✓	✓	✓	×
UNSW	Botnet/IDS	✓	×	✓	✓	✓	✓	×
ISCX	Botnet/IDS	✓	×	✓	✓	✓	✓	×
AADR	DGA Botnet	✓	✓	×	×	×	✓	✓
JBR	DGA Botnet	✓	✓	×	×	×	✓	✓
AT1D	DGA Botnet	✓	✓	×	×	×	✓	✓
BDD	DGA Botnet	✓	✓	×	×	×	✓	✓
UMU	DGA Botnet	✓	✓	×	×	×	✓	✓
DFF	DGA Botnet	✓	✓	×	×	×	✓	✓
OSINT	DGA Botnet	✓	✓	×	×	×	✓	✓
360NL	DGA Botnet	✓	✓	×	×	×	✓	✓
TMM	DGA Botnet	✓	✓	×	×	×	✓	✓
UTL	*DGA Botnet*	✓	✓	×	×	×	✓	✓

4.1.5. Tiêu chí xây dựng bộ dữ liệu DGA Botnet

NCS đề xuất nhóm gồm 06 tiêu chí cơ bản đối với một bộ dữ liệu về DGA Botnet, bao gồm:

(1) Nhãn nhị phân;

(2) Nhãn đa lớp;

(3) Tên miền gốc;

(4) Trích xuất thuộc tính;

(5) Công khai;

(6) Tài liệu.

Bảng 4.5 khái quát các ưu điểm và hạn chế của những bộ dữ liệu về DGA Botnet chuyên dụng.

Bảng 4.5. Khái quát ưu điểm và hạn chế của các bộ dữ liệu DGA Botnet hiện có và bộ dữ liệu UTL_DGA22 đề xuất

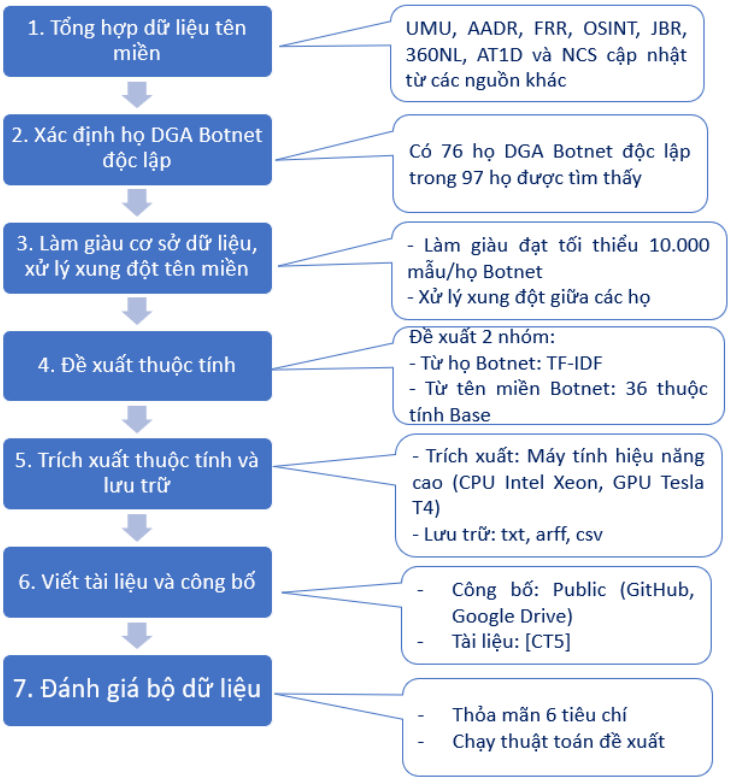
Bộ dữ liệu	**Phân lớp nhị phân**	**Phân lớp đa lớp**	**Tên miền gốc**	**Trích xuất thuộc tính**	**Công khai**	**Tài liệu**
AADR	✓	✓	✓	×	✓	N/A
JBR	✓	✓	✓	×	✓	✓
AT1D	✓	×	✓	×	✓	N/A
BDD	✓	×	×	✓	✓	✓
UMU	✓	✓	✓	✓	✓	✓
DFF	✓	✓	✓	×	✓	✓
OSINT	✓	×	✓	×	✓	N/A
360NL	✓	✓	✓	×	✓	N/A
TMM	✓	×	✓	×	✓	N/A
UTL	✓	✓	✓	✓	✓	✓

Bộ dữ liệu UTL_DGA22 đề xuất sẽ đáp ứng đầy đủ các yêu cầu trên và cập nhật thêm các dữ liệu mới, thuộc tính mới đã trích xuất.

4.2. Bộ dữ liệu UTL_DGA22 đề xuất

4.2.1. Quy trình xây dựng bộ dữ liệu

NCS đề xuất quy trình xây dựng bộ dữ liệu gồm 07 bước và kết quả tương ứng được tóm tắt tại 0:



Hình 4.1. Quy trình 07 bước xây dựng bộ dữ liệu DGA Botnet và tóm tắt kết quả đạt được theo từng bước

4.2.2. Danh sách các họ DGA Botnet trong bộ dữ liệu UTL_DGA22

Bộ dữ liệu UTL_DGA22 bao gồm 76 họ DGA Botnet riêng biệt tương ứng với 76 nhãn (Bảng 4.6)

Bảng 4.6. Danh sách 76 họ DGA Botnet trong bộ dữ liệu UTL_DGA22

STT	Tên (*Tên gọi khác*)	STT	Tên (*Tên gọi khác*)
1	banjori (*MultiBanker 2 / BankPatch / BackPatcher*)	39	qsnatch
2	bazarbackdoor (*BazarLoader / Team9Backdoor*)	40	ramnit

3	bazarbackdoor_v2 (*BazarLoader / Team9Backdoor*)	41	ranbyus_v1
4	bazarbackdoor_v3 (*BazarLoader / Team9Backdoor*)	42	ranbyus_v2
5	chinad	43	reconyc
6	corebot	44	shiotob (*Urlzone / Bebloh*)
7	dircrypt	45	simda (*Shiz*)
8	dnschanger (*Alureon*)	46	sisron (*Tomb / win32_agent.wrq / trojan.scar*)
9	fobber_v1 (*Tinba_v3*)	47	suppobox_1
10	fobber_v2 (*Tinba_v3*)	48	suppobox_2
11	gozi_rfc4343	49	suppobox_3
12	gozi_nasa	50	symmi
13	gozi_luther	51	tempedreve
14	gozi_gpl	52	tinba [91] (*Tinybanker*)
15	kraken_v1 (*Oderoor / Bobax*)	53	vawtrak_v1
16	kraken_v2 (*Oderoor / Bobax*)	54	vawtrak_v2
17	locky	55	vawtrak_v3
18	monerodownloader	56	zloader
19	murofet_v1	57	cryptolocker
20	murofet_v2	58	rovnix
21	murofet_v3	59	matsnu
22	mydoom (*Novarg / Mimail.r / Shimgapi*)	60	ramdo
23	necurs	61	bigviktor
24	newgoz (*Gameover Zeus / Peer-to-Peer Zeus*)	62	ccleaner
25	nymaim	63	enviserv
26	nymaim2	64	vidro
27	padcrypt	65	dyre
28	pitou	66	beautiful baby
29	pizza	67	bamital
30	proslikefan	68	emotet
31	pushdo	69	infy
32	pykspa_improved_useful	70	murofetweekly
33	pykspa_improved_noise	71	oderoor

34	pykspa_precursor	72	pandabanker
35	qadars	73	sphinx
36	qakbot	74	szribi
37	virus	75	tinynuke
38	wd	76	torpig

4.2.3. Mô tả về các thuộc tính đề xuất

NCS đề xuất 02 nhóm thuộc tính gồm nhóm thuộc tính dựa trên tên miền (BaseFeatures) và nhóm thuộc tính dựa trên họ tên miền (TF-IDF),

4.2.4. Cấu trúc lưu trữ của bộ dữ liệu

Bộ dữ liệu DGA_UTL22 được cấu trúc gồm hai phần tương ứng với hai thư mục, gồm DGA_Botnets_Domains và DGA_Botnets_Features_Extraction.

4.2.5. Đánh giá với tiêu chí của Zago và cộng sự

Zago và cộng sự đề xuất 09 tiêu chí cho một bộ dữ liệu về DGA Botnet [44], bao gồm: Tính tổng hợp (Def 2.1. SYNT), Tính phổ biến (Def 2.2. GNRL), Tính đại diện (Def 2.3. RPST), Tính cân bằng (Def 2.4. BLNC), Tính mở rộng (Def 2.5. EXTS), Tính xác minh (Def 2.6. VRFB), Tính định hướng riêng tư (Def 2.7. PROR), Tính sẵn sàng cho học máy (Def 2.8. MLRD), Tính gán nhãn (Def 2.9. LABL).

Bộ dữ liệu UTL_DGA22 đáp ứng đầy đủ 09 tiêu chí trên.

4.3. Thử nghiệm một số thuật toán trên bộ dữ liệu đề xuất

4.3.1. Thử nghiệm áp dụng thuộc tính đề xuất

Cả hai bộ thuộc tính BaseFeatures và TF-IDF đều chứng tỏ sự phù hợp khi làm đầu vào cho các thuật toán học máy để giải quyết bài toán phát hiện và phân loại DGA Botnet.

4.3.2. Thử nghiệm áp dụng một số thuật toán

Trong phần này, NCS tiến hành chạy một số thuật toán đã trình bày tại Chương 2, Chương 3 trên bộ dữ liệu mới UTL_DGA22, bao gồm: Thuật toán NCM, học máy, mô hình học sâu LA_Bin07 và LA_Mul07.

Các kết quả trên cho thấy, (1) Bộ dữ liệu UTL_DGA22 hoàn toàn phù hợp để đánh giá bài toán DGA Botnet và (2) Các thuật toán đề xuất vẫn có độ chính xác cao khi đánh giá trên bộ dữ liệu mới.

4.4. Kết luận Chương 4

Một phần kết quả trình bày tại Chương 4 được công bố tại [CT5] trong Danh mục các công trình công bố liên quan đến luận án.

KẾT LUẬN VÀ KIẾN NGHỊ

Luận án "*Nghiên cứu cải tiến một số mô hình học máy và học sâu áp dụng cho bài toán phân loại DGA Botnet*" được hoàn thành tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, với hai đóng góp bao gồm:

1. Đề xuất cải tiến kiến trúc lõi kết hợp BiLSTM với cơ chế Attention và sử dụng trong xây dựng mô hình LA_Bin07 để phát hiện và mô hình LA_Mul07 để phân loại DGA Botnet với độ chính xác được cải thiện.

2. Hoàn thiện bổ sung quy trình xây dựng tập dữ liệu mẫu và đề xuất bộ dữ liệu chuyên dùng UTL_DGA22 được mô tả và gán nhãn, phục vụ phân loại DGA Botnet.

Trả lời câu hỏi nghiên cứu đặt ra ban đầu: Kiến trúc lõi BiLSTM_SelfA_Double đã cải tiến được so với các kiến trúc trước đó, thể hiện qua độ chính xác được nâng cao của mô hình LA_Mul07 trong bài toán phân loại DGA Botnet.

Bên cạnh những kết quả đạt được, NCS dự kiến một số hướng phát triển trong thời gian tới, cụ thể như sau:

- Áp dụng mạng TCN để đề xuất kiến trúc học sâu mới đạt độ chính xác cao hơn trong phân loại.

- Xây dựng cơ chế huấn luyện dành riêng cho các họ DGA Botnet có sự tương đồng cao hoặc là các phiên bản kế tiếp của nhau.

Giải pháp đề xuất có đóng vai trò như một module phát hiện và phân loại DGA Botnet, có thể được tích hợp vào các giải pháp đảm bảo an ninh mạng như Tường lửa hoặc Giải pháp an ninh hợp nhất.

DANH MỤC CÁC BÀI BÁO ĐÃ XUẤT BẢN LIÊN QUAN ĐẾN LUẬN ÁN

[CT1] Can, N.V., Tu, D. N., **Tuan, T. A.**, Long, H. V., Son, L. H., & Son, N. T. K. (2020). A new method to classify malicious domain name using Neutrosophic sets in DGA Botnet detection. *Journal of Intelligent & Fuzzy Systems*, *38*(4), 4223-4236. ***(ISI Q2, IF = 1.737)***

[CT2] **Tuan, T. A.**, Long, H. V., Son, L. H., Kumar, R., Priyadarshini, I., & Son, N. T. K. (2020). Performance evaluation of Botnet DDoS attack detection using machine learning. *Evolutionary Intelligence*, *13*(2), 283-294. ***(SCOPUS, ESCI Q2)***

[CT3] **Tuan, T. A.**, Anh, N. V., & Long, H. V. (2021, December). Assessment of Machine Learning Models in Detecting DGA Botnet in Characteristics by TF-IDF. In *2021 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)* (pp. 1-5). IEEE. ***(SCOPUS)***

[CT4] **Tuan, T. A.**, Long, H. V., & Taniar, D. (2022). On Detecting and Classifying DGA Botnets and their Families. *Computers & Security*, *113*, 102549. ***(ISI Q1, IF = 5.105)***

[CT5] **Tuan, T. A.,** Anh, N. V., Luong, T. T., & Long, H. V. (2023). UTL_DGA22-a dataset for DGA botnet detection and classification. *Computer Networks*, 221, 109508. ***(ISI Q1, IF = 5.493)***

[CT6] **Tống Anh Tuấn**, Nguyễn Ngọc Cương, Nguyễn Việt Anh, Hoàng Việt Long. (2022). Đề xuất ứng dụng giải pháp phân lớp nhị phân trong bài toán DGA Botnet cho phát hiện địa chỉ IP độc hại. Hội thảo Quốc gia lần thứ XXV *"Một số vấn đề chọn lọc của Công nghệ thông tin và Truyền thông*" (VNICT 2022), trang 55-60.