MINISTRY OF EDUCATION
AND TRAINING

VIETNAM ACADEMY OF SCIENCE
AND TECHNOLOGY

GRADUATE UNIVERSITY OF SCIENCE AND TECHNOLOGY

Tong Anh Tuan

RESEARCH FOCUSES ON IMPROVING SEVERAL MACHINE LEARNING AND DEEP LEARNING MODELS FOR CLASSIFYING DGA BOTNET

SUMMARY OF DISSERTATION IN INFORMATION SYSTEM

HA NOI - 2023

The dissertation is completed at: Graduate University of Science and Technology, Vietnam Academy Science and Technology

Supervisors:

1. Assoc. Prof. Ph.D. Hoang Viet Long, University of Technology – Logistics of Public Security, Ministry of Public Security.
2. Assoc. Prof. Ph.D. Nguyen Viet Anh, Institute of Information Technology, Vietnam Academy of Science and Technology.

Referee 1: Assoc. Prof. Ph.D. Bui Thu Lam, Academy of Cryptography Techniques, Government Cipher Committee.

Referee 2: Assoc. Prof. Ph.D. Nguyen Ha Nam, Electric Power University, Ministry of Industry and Trade.

Referee 3: Assoc. Prof. Ph.D. Ngo Quoc Tao, Institute of Information Technology, Vietnam Academy of Science and Technology.

The dissertation was examined by Examination Board of Graduate University of Science and Technology, Vietnam Academy of Science and Technology at 9:00 AM, August 9, 2023.

The dissertation can be found at:

1. Graduate University of Science and Technology Library
2. National Library

INTRODUCTION

1. Rationale

A botnet is an assemblage of compromised computers that are infiltrated, controlled, and managed remotely through command and control servers [1]. Detecting botnets has been a perennial concern for researchers, making it a topic of paramount importance.

There are two primary approaches commonly employed in botnet detection, as outlined in the literature [5]:

(1) Honeynet-based approach.

(2) Intrusion Detection System (IDS)-based approach, encompassing:

- Anomaly-based botnet detection techniques.
- Signature-based botnet detection techniques.
- Domain-based botnet detection techniques.

The study will focus on the study of Domain Generation Algorithm (DGA) botnets. Several in-depth research outcomes related to the DGA botnet problem have been published, including network traffic analysis techniques [6], [7], [8], [9], [10], machine learning techniques [11], [12], [13], [14], and deep learning techniques [15], [16], [17], [18], [19], [20].

Based on these issues, our research poses the following research questions for the study: "How can we advance techniques to enhance the classification of DGA botnets using machine learning and deep learning approaches?"

2. Objectives of the study

The main objective of this study is to study and enhance machine learning and deep learning models to improve the accuracy of DGA Botnet classification solutions.

3. Scope of the study

The study focuses on the following subjects:

- Characteristics, mechanisms, and behaviors of DGA Botnets; techniques for detection and classification of Botnets based on domain names.

- Binary and multi-class classification problems corresponding to the detection and classification of DGA Botnets.

- Public, reliable, and up-to-date datasets on DGA Botnets along with the process of constructing new datasets.

4. Content and Methodology of the study

a. Content of the study

Some detailed research content that the study will concentrate on includes:

- Researching the characteristics and techniques for detecting and classifying DGA Botnets.

- Investigating LSTM networks, Attention mechanisms, and their variations, and based on these, proposing new deep learning models to enhance the effectiveness of DGA Botnet classification.

- Researching the process, criteria, and datasets related to DGA Botnets and their application.

b. Methodology

There are methodologies used in the study, include:

- Theoretical research.

- Expert consultation.

- Experimental research and evaluation.

5. Contributions of the study

The thesis makes two primary contributions:

- *Contribution 1*: Proposing improvements to the core architecture by combining BiLSTM with an Attention mechanism and utilizing it to build the LA_Bin07 model for detecting and the LA_Mul07 model for classifying DGA Botnets with enhanced accuracy.
- *Contribution 2*: Enhancing and supplementing the process of constructing sample datasets and proposing the UTL_DGA22 dataset, which is described, labeled, and serves the purpose of DGA Botnet classification.

6. Organization of the study

The study is organizatedinto four chapters as follows:

- Chapter 1: Theoretical Foundations of DGA Botnets.
- Chapter 2: Detecting DGA Botnets using NCM and Machine Learning.
- Chapter 3: Detecting and Classifying DGA Botnets using Deep Learning.
- Chapter 4: The Process of Building the New UTL_DGA22 Dataset for the DGA Botnet problem.

The research results from this study have been published in four articles in international specialized scientific journals listed in the SCIE/Scopus category, presented in one report at a national scientific conference, and presented in one report at a reputable international scientific conference. These publications are listed in the "List of Related Publications" at the end of this study.

CHAPTER 1. THEORETICAL FOUNDATIONS OF DGA BOTNETS

1.1. General Overview of Botnets

1.1.1. Botnet Concept

According to Provos & Holz, a Botnet is a "network composed of numerous compromised computers that can be remotely controlled by attackers."

1.1.2. The Evolution of Botnet Technology

1.1.3. Some Characteristics of Botnets

Botnets exhibit specific characteristics related to their lifecycle, infection methods, and malicious behaviors.

1.1.4. Classification of Botnets

Botnets can be classified based on criteria such as protocols, infection vectors, or architecture.

1.2. Botnet Detection Techniques

There are main techniques used for detecting Botnets:

(1) Honeynet-based techniques.

(2) Intrusion Detection System (IDS)-based techniques, including:

- Anomaly-based Botnet detection.
- Signature-based Botnet detection.
- Domain-based Botnet detection.

1.3. The DGA Botnet Problem

1.3.1. Overview of DGA Botnets

DGA Botnets refer to a type of Botnet deployed in a Client-Server model. In this model, the Bots act as Clients and establish connections back to a Command and Control (C&C) server, acting as the Server, through automatically generated and pre-agreed DNS domain names. This technique is used to evade security systems.

1.3.2. Detecting DGA Botnets Problem

This problem aims to detect domain names generated by DGA Botnets in contrast to benign domain names. The data is binary, with two labels, 0 and 1.

1.3.3. Classifying DGA Botnets Problem

The classification problem seeks to determine the family or group of DGA Botnets, where the data can have n labels, corresponding to n considered DGA Botnet families.

1.3.4. Distinguishing from the Fake URL Detection Problem

Detecting DGA Botnets is distinct from the task of detecting fake URLs.

1.3.5. Evaluation Datasets for the DGA Botnet Problem

The researher selected four datasets, including: Andrey Abakumov's DGA Repository [35], OSINT DGA feed [36], UMUDGA Dataset [13], and 360NetLab Dataset [37] (Table 1.4).

Table 1.4. Description of the 4 DGA Botnet datasets used in the evaluations.

	DGA Botnet Detection	DGA Botnet Classification	Number of legitimate samples	Number of DGA Botnet samples	Number of DGA Botnet families
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. Evaluation Metrics for the Problem

The researher evaluates the problem using metrics including Accuracy, Precision, Recall, and F_1-score.

1.3.7. Significance of the DGA Botnet Problem

Leveraging the operation mechanism of DGA Botnets can provide an effective solution with several advantages, such as not requiring excessive data collection and processing capabilities of the system. Detecting the activities of DGA Botnets can be crucial, as it can be deployed even after they have infiltrated devices.

1.4. Some Research Approaches to Addressing the DGA Botnet Problem

Approaches using DNS analysis techniques: Alieyan et al. [6], Kwon et al. [7], Wang et al. [8], Chowdhury et al. [38], Bisio et al. [9], Wang et al. [40], Trung et al. [10].

Machine learning-based approaches: Hieu et al. [11], Khan et al. [12], Zago et al. [13], Xuan et al. [14], Suryotrisongko et al. [45], Zhao et al. [46], Alauthman et al. [47].

Deep learning-based approaches: Duc et al. [15], Curtin et al. [16], Qiao et al. [17], Namgung et al. [19], Vinayakumar et al. [20], Liu et al. [51].

1.5. Conclusion of Chapter 1

Some of the results presented in Chapter 1 have been published in [CT2] [CT6] in the List of Related Publications.

CHAPTER 2. DETECTING DGA BOTNETS USING NCM AND MACHINE LEARNING

2.1. Detecting DGA Botnets using NCM

2.1.1. The NCM Algorithm

Neutrosophic Set, proposed by Smarandache [52], is an enhancement of the traditional fuzzy set. In the space X, a neutrosophic set A is defined as follows:

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

Where $T_A(x), I_A(x), F_A(x)$ represent the membership degree of an element x in terms of belonging, neutrality, and non-belonging to a certain specified set. The values $T_A(x), I_A(x), F_A(x) \in [0, 1]$ and satisfy the condition:

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

Neutrosophic C-Means (NCM) is a fuzzy clustering algorithm on neutrosophic sets, proposed by Gou et al. [53], and is summarized as follows:

```
Algorithm: $NCM(X, \varepsilon)$
Input       $X, \varepsilon$
Output      $k$
Init $T^{(0)}, I^{(0)}, F^{(0)}$
Init $C, m, \varepsilon, \delta,\ \omega_1,\ \omega_2,\ \omega_3$
Loop:
            Calculate $c_j^{(k)}$
            Calculate $\bar{c}_{i_{max}}$
            Update $T^{(\mathrm{k+1})}$
            Update $I^{(\mathrm{k+1})}$
            Update $F^{(\mathrm{k+1})}$
Condition   $\left|T_{ij}^{(k+1)} - T_{ij}^{(k)}\ \right| > \varepsilon$
Assign each data point to the class with the highest $TM = [T\ , I\ , F]$.
```

$x_i \epsilon\ k^{th}$ if $k = argmax(TM_{ij})$ with $j = 1, 2, \ldots., C + 2$

2.1.2. Applying NCM for DGA Botnet Detection

The NCM algorithm is applied through two steps as follows:

(1) Feature Selection: The researher proposes and selects a set of fundamental domain features. These features are appropriately vectorized to serve as inputs for the NCM algorithm.

The results of feature selection on the four datasets are listed in Table 2.2.

Table 2.2. Selected features used as inputs for the NCM algorithm.

No.	**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) Clustering and Labeling: Using the NCM algorithm to divide the data points into three clusters, then assigning representative labels to the corresponding clusters, which are DGA Botnet, Benign, and Noise.

Subsequently, evaluations were conducted on the AADR, 360NL, OSINT, and UMUDGA datasets. The results are presented in Table 2.3.

Table 2.3. Detection Results of the NCM Algorithm for DGA Botnet on 4 datasets

	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. Detecting DGA Botnets Using Machine Learning

2.2.1. Evaluation Model for Machine Learning Algorithms

The stages in the process of evaluating machine learning algorithms are depicted in Figure 2.6.

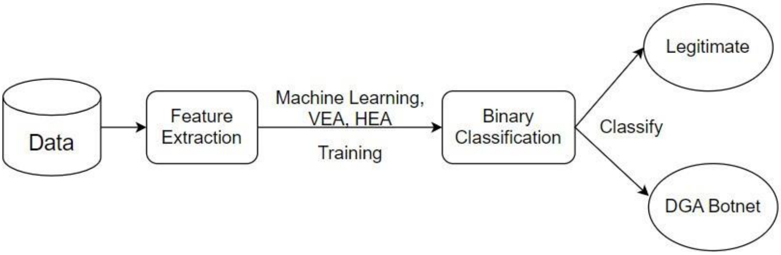


Figure 2.6. Training and Evaluation Model Diagram

In this diagram, with input data comprising domain names, both benign and malicious, labeled accordingly, N-grams are used to split the domain names, and TF-IDF is employed to represent the features.

For machine learning, The researher utilizes the following algorithms: Support Vector Machines (SVM), Logistic Regression (LR), Naive Bayes (NB), Neural Networks (NN), Decision Trees (DT), Random Forests (RF), k-Nearest Neighbor (k-NN), and Adaptive Boosting (AB).

For ensemble learning models based on voting, The researher proposes two models, VEA and HEA.

Evaluation is performed on the UMUDGA Dataset.

2.2.2. Detection Results of Machine Learning Models for DGA Botnet

Table 2.5 presents the results of Precision, Recall, and F1-score when using machine learning algorithms for DGA Botnet detection.

Table 2.5. Detection Results of DGA Botnet Using Machine Learning on the UMUDGA Dataset.

Model	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

Most machine learning algorithms achieve high accuracy. The LR, NN, and SVM models yield the highest overall results with an F1-score of 0.97, while the RF model has the lowest result at 0.77.

2.2.3. Detection Results of Ensemble Learning Models

The results of VEA and HEA are shown in Table 2.6.

Table 2.6. Detection Results of DGA Botnet for VEA and HEA Models on the UMUDGA Dataset.

Algorithm	A.Precision	A.Recall	A.F$_1$-score
The average of the individual models	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

Limitations of the NCM and Machine Learning Solutions:

- Accuracy can still be further improved.

- Requires significant training time when run on a CPU.

- Not suitable for multi-class classification problems.

2.2.4. Training and Evaluation Time of Machine Learning Models

Ensemble-based learning models improve accuracy but have longer training times.

2.3. Conclusion of Chapter 2

Some of the results presented in Chapter 2 have been published in [CT1] [CT3] in the List of Related Publications.

CHAPTER 3. DETECTING AND CLASSIFYING DGA BOTNETS USING DEEP LEARNING

3.1. Technical Deep Learning Foundation for DGA Botnet Problem

3.1.1. Recurrent Neural Network

Recurrent Neural Network **-** RNN is designed for training on input data in the form of sequences or time-series data.

3.1.2. Long-Short Term Memory Network and Variants

LSTM networks are an improvement over RNN, and variants like BiLSTM can be trained bidirectionally.

3.1.3. Attention Mechanism and Variants

Attention mechanisms, including Self-Attention, enhance training efficiency.

3.1.4. LSTM Networks Integrated with Attention

LSTM networks integrated with attention mechanisms enhance training efficiency for the DGA Botnet problem.

3.2. Proposal of Core Architecture and Two New Deep Learning Models

3.2.1. Execution Process for the DGA Botnet Problem

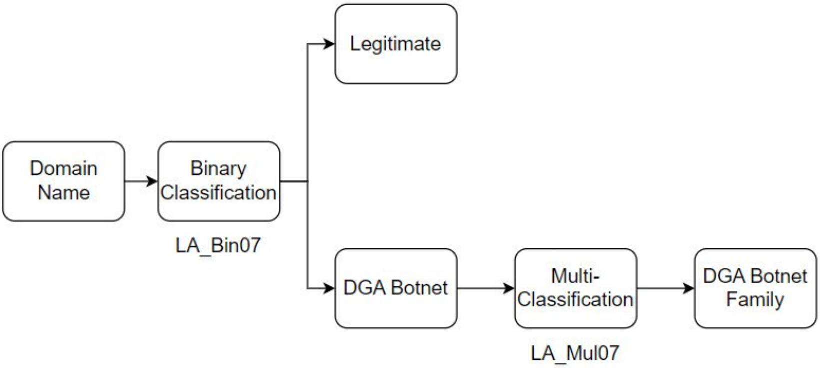


Figure 3.11. DGA Botnet Detection and Classification Solution with Two New Deep Learning Models, LA_Bin07 and LA_Mul07

3.2.2. Proposal of the Core Architecture for the Deep Learning Model.

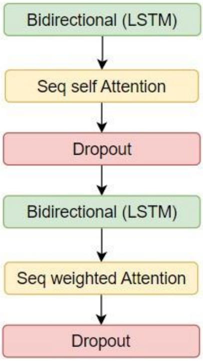


Figure 3.12. Proposed Core Architecture of BiLSTM_SelfA_Double

3.2.3. Data Preprocessing

Consisting of 2 steps: Encoding and Word Embedding.

3.2.4. LA_Bin07 Model for DGA Botnet Detection

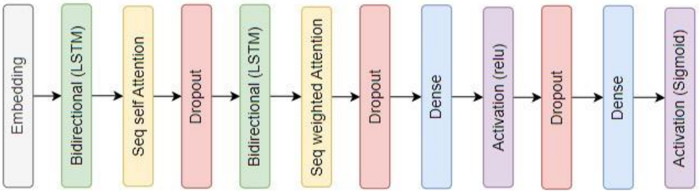


Figure 3.13. Architecture of the LA_Bin07 Model.

3.2.5. LA_Mul07 Model for DGA Botnet Classification

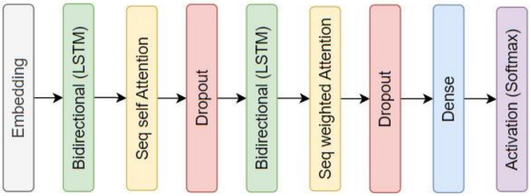


Figure 3.14. Proposed Structure of the LA_Mul07 Model.

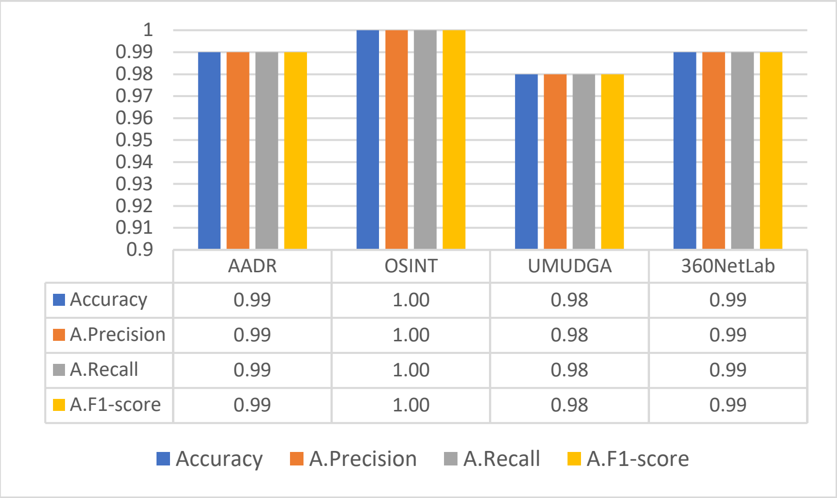
3.3. Evaluation of the Two Proposed Deep Learning Models

3.3.1. Evaluation Dataset and Environment

Table 3.4. Symbols for Evaluating
the Two Proposed Deep Learning Models.

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

3.3.2. Evaluation of the LA_Bin07 Model for DGA Botnet Detection



	AADR	OSINT	UMUDGA	360NetLab
🟦 Accuracy	0.99	1.00	0.98	0.99
🟧 A.Precision	0.99	1.00	0.98	0.99
◼ A.Recall	0.99	1.00	0.98	0.99
🟨 A.F1-score	0.99	1.00	0.98	0.99

Figure 3.15. Results of Evaluations B1, B2, B3, and B4.

3.3.3. Evaluation of the LA_Mul07 Model for DGA Botnet Classification

The LA_Mul07 model achieves high accuracy in classifying DGA Botnet families, even when there are multiple families to classify. Specifically, it achieves 1.00 on the AADR dataset and 0.86 on the UMUDGA dataset.

3.4. Evaluation with Related Studies

3.4.1. Evaluation of the Two Proposed Models on the UMUDGA Dataset

The LA_Bin07 model outperforms the SVM model by a significant margin in terms of detection accuracy. It also performs nearly as well as the AB, NN, RF, DT, and kNN models.

The LA_Mul07 model achieves much higher accuracy in classification compared to other machine learning models.

3.4.2. Evaluation of the Two Proposed Models Against Other Deep Learning Architectures

The researcher also evaluates the proposed models against other deep learning architectures based on CNN and LSTM, including Basic CNN, Basic LSTM, Bi-LSTM, and CNN-LSTM. The results show that the LA_Bin07 and LA_Mul07 models perform the best among the tested models.

3.4.3. Evaluation of the LA_Mul07 Classification Model Against Related Models

In this section, The researcher uses the LA_Mul07 model to evaluate it against the model by Qiao and Namgung, which follows a similar LSTM-based approach with attention. The evaluations are performed on the same dataset as described and published by those authors.

The experimental results show that the LA_Mul07 model improves the F_1-score by 3% compared to Qiao and colleagues' LSTM_AM model, improves accuracy by 1.03% compared to the BiLSTM_Attention model, and by 0.38% compared to the CNN-BiLSTM_Ensemble model by Namgung and colleagues.

The LA_Mul07 model also demonstrates consistent detection performance across different DGA Botnet families (as indicated by Precision and Recall for each label) compared to the models by Qiao and Namgung.

3.5. Conclusion of Chapter 3

Some of the results presented in Chapter 3 have been published in [CT4] in the list of related works related to the thesis.

CHAPTER 4. THE PROCESS OF BUILDING THE NEW UTL_DGA22 DATASET FOR THE DGA BOTNET PROBLEM.

4.1. Problem Statement on DGA Botnet Datasets

4.1.1. Overview of the Issue

The proposed solutions in previous research are often evaluated on datasets collected by research groups at different times, with varying sample sizes, and have limited public availability, making them less suitable for comparative studies.

4.1.2. General Botnet Datasets

There are several datasets related to botnets in general, such as CTU-13, UGR16, DreLAB, UNSW-NB15, ISCX-Bot-2014. None of these datasets are specifically designed for evaluating DGA Botnet detection as they lack domain names from DGA Botnet families and corresponding labels.

4.1.3. DGA Botnet Datasets

There are also datasets specifically focused on DGA Botnets, including Andrey Abakumov's DGA Repository, Johannes Bader's Domain Generation Algorithms Repository, Alexa Top 1 Million Domains, Botnet DGA Dataset, UMUDGA Dataset, DGArchive by Fraunhofer FKIE, OSINT DGA feed, 360NetLab Dataset, and The Majestic Million.

4.1.4. Research Problem Statement

There are differences in the structure and purpose between datasets for general botnets and those for DGA Botnets. The researcher categorizes and details these differences in Table 4.4.

Table 4.4. Evaluation of Characteristics of Dataset Groups for Botnets

Dataset	Group	Botnet Detect	DGA Botnet Detect	Attack Detect	Network Traffic	Format		
						PCAP	SCV	TXT
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. Criteria for Constructing DGA Botnet Datasets

The researcher proposes a set of 6 fundamental criteria for a DGA Botnet dataset, which includes:

(1) Binary Labels

(2) Multi-class Labels

(3) Domain Name

(4) Feature Extraction

(5) Public Availability

(6) Documentation

Table 4.5 provides an overview of the advantages and limitations of existing specialized DGA Botnet datasets.

Table 4.5. Summary of the Advantages and Limitations of Existing DGA Botnet Datasets and the Proposed UTL_DGA22 Dataset

Dataset	Detection	Classification	Domain Name	Feature Extraction	Public	Document
AADR	✓	✓	✓	✗	✓	N/A
JBR	✓	✓	✓	✗	✓	✓
AT1D	✓	✗	✓	✗	✓	N/A
BDD	✓	✗	✗	✓	✓	✓
UMU	✓	✓	✓	✓	✓	✓
DFF	✓	✓	✓	✗	✓	✓
OSINT	✓	✗	✓	✗	✓	N/A
360NL	✓	✓	✓	✗	✓	N/A
TMM	✓	✗	✓	✗	✓	N/A
UTL	✓	✓	✓	✓	✓	✓

The proposed UTL_DGA22 dataset will fully meet the requirements mentioned above and include additional new data and extracted attributes.

4.2. Proposed UTL_DGA22 Dataset

4.2.1. Dataset Construction Process

The researcher proposes a dataset construction process consisting of 7 steps, with corresponding results summarized in Figure 4.1.

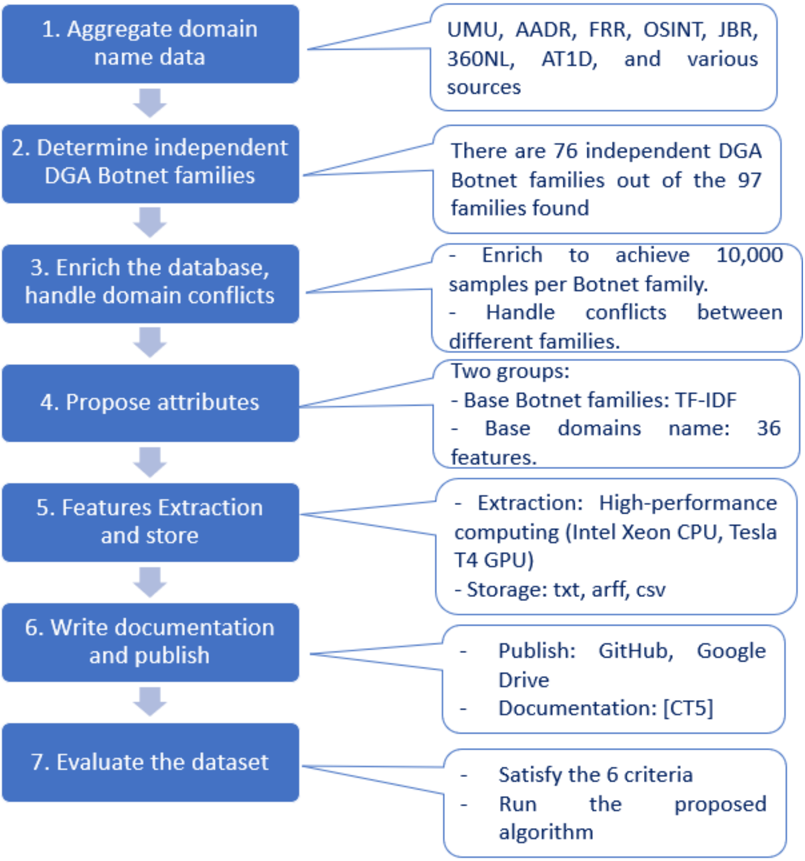

1. Aggregate domain name data
UMU, AADR, FRR, OSINT, JBR, 360NL, AT1D, and various sources
2. Determine independent DGA Botnet families
There are 76 independent DGA Botnet families out of the 97 families found
3. Enrich the database, handle domain conflicts
- Enrich to achieve 10,000 samples per Botnet family.
- Handle conflicts between different families.
4. Propose attributes
Two groups:
- Base Botnet families: TF-IDF
- Base domains name: 36 features.
5. Features Extraction and store
- Extraction: High-performance computing (Intel Xeon CPU, Tesla T4 GPU)
- Storage: txt, arff, csv
6. Write documentation and publish
- Publish: GitHub, Google Drive
- Documentation: [CT5]
7. Evaluate the dataset
- Satisfy the 6 criteria
- Run the proposed algorithm

Figure 4.1. The 7-step process for constructing the DGA Botnet dataset and a summary of the results achieved at each step.

4.2.2. List of DGA Botnet Families in the UTL_DGA22 Dataset

The UTL_DGA22 dataset comprises 76 distinct DGA Botnet families, corresponding to 76 labels (Table 4.6).

Table 4.6. List of 76 DGA Botnet families in the UTL_DGA22 dataset.

No.	Name *(Other name)*	No.	Name *(Other name)*
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. Description of Proposed Attributes

The researcher proposes 02 attribute groups, including domain-based attributes (BaseFeatures) and family name-based attributes (TF-IDF).

4.2.4. Dataset Storage Structure

The DGA_UTL22 dataset is structured into two corresponding directories, DGA_Botnets_Domains and DGA_Botnets_Features_Extraction.

4.2.5. Evaluation Against Zago et al.'s Criteria

Zago et al. proposed 09 criteria for a DGA Botnet dataset [44], including: Def 2.1. SYNT, Def 2.2. GNRL, Def 2.3. RPST, Def 2.4. BLNC, Def 2.5. EXTS, Def 2.6. VRFB, Def 2.7. PROR, Def 2.8. MLRD, Def 2.9. LABL.

The UTL_DGA22 dataset fully meets all 09 criteria.

4.3. Testing Several Algorithms on the Proposed Dataset

4.3.1. Testing the Proposed Attributes

Both attribute sets, BaseFeatures and TF-IDF, are shown to be suitable as inputs for machine learning algorithms to solve the problem of detecting and classifying DGA Botnets.

4.3.2. Testing Several Algorithms

In this section, The researcher conducted experiments on the new UTL_DGA22 dataset using several algorithms presented in Chapters 2 and 3, including the NCM algorithm, machine learning, LA_Bin07 and LA_Mul07.

The results show that (1) the UTL_DGA22 dataset is entirely suitable for evaluating the DGA Botnet problem, and (2) the proposed algorithms still achieve high accuracy when evaluated on the new dataset.

4.4. Conclusion of Chapter 4

A part of the results presented in Chapter 4 was published in [CT5] in the list of related works for the thesis.

CONCLUSION AND RECOMMENDATIONS

The thesis "*Research focuses on improving several machine learning and deep learning models for classifying DGA Botnet*" has been completed at the Graduate University of Sciences and Technology, Viet Nam Academy of Sciences and Technology, with two main contributions:

1. Proposing an improved core architecture combining BiLSTM with the Attention mechanism and using it to build the LA_Bin07 model for detecting and the LA_Mul07 model for classifying DGA Botnets with improved accuracy.

2. Completing the sample dataset construction process and proposing the dedicated dataset UTL_DGA22, which is described and labeled to serve DGA Botnet classification.

Answering the initial research question: The BiLSTM_SelfA_Double core architecture has shown improvements over previous architectures, as demonstrated by the increased accuracy of the LA_Mul07 model in DGA Botnet classification.

In addition to the achieved results, there are several directions for future development, including:

- Applying the TCN network to propose a new deep learning architecture achieving higher accuracy in classification.

- Developing a training mechanism specifically for DGA Botnet families with high similarity or those that are successive versions of each other.

The proposed solution can serve as a module for detecting and classifying DGA Botnets, which can be integrated into network security solutions such as Firewalls or Unified Threat Managements.

LIST OF THE PUBLICATIONS RELATED TO THE DISSERTATION

[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.