MINISTRY OF EDUCATION AND TRAINING

VIETNAM ACADEMY OF SIENCE AND TECHNOLOGY

GRADUATE UNIVERSITY OF SIENCE AND TECHNOLOGY

……..….*……**

TONG ANH TUAN

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

Major: Information System
Major code: 9 48 01 04

SUMMARY OF COMPUTER DOCTORAL THESIS

Ha Noi – 2023

The thesis has been completed at: Graduate University of Science and Technology- Vietnam Academy of Science and Technology

Supervisor 1: **Associate Professor, Dr. Hoang Viet Long.**

Supervisor 2: **Associate Professor, Dr. Nguyen Viet Anh.**

Reviewer 1: …

Reviewer 2: …

Reviewer 3: ….

The thesis shall be defended in front of the Thesis Committee at Vietnam Academy Of Science And Technology - Graduate University Of Science And Technology, at ……. hour……, date…… month…….year 2023

This thesis could be found at:

- The National Library of Vietnam
- The Library of Graduate University of Science and Technology

LIST OF PUBLICATIONS

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.

INTRODUCTION

1. Significance of the thesis

Botnets are a persistent threat on the Internet [1]. They continuously evolve, improve their source code, innovate infection methods, and possess increasingly destructive capabilities. Information systems deployed on the Internet are constantly exposed to the risk of being attacked by botnets, resulting in significant economic, reputational, service-related, and even political consequences.

Some studies have elucidated the dangers and proposed Botnet detection techniques, such as Ghafir et al. [3], Alieyan et al. [7], Kwon et al. with the PsyBoG solution [8], Wang et al. with the DBod solution [9], Bisio et al. [10], Trung et al. with PGS-Graph [11].

In terms of machine learning and deep learning approaches, notable studies include Hieu et al. [12], Khan et al. [13], Xuan et al. [14], Duc et al. [15] with LSTM.MI, Curtin et al. [16], Vinayakumar et al. [17]. Zago et al. introduced a new dataset called UMUDGA Dataset [18], specifically designed for evaluating the DGA botnet problem.

The research results indicate that in the mentioned approaches, traffic analysis using machine learning and deep learning, specifically LSTM networks, achieve high results of 96.3% or above in the DGA botnet detection problem. Another issue is the limited research on the classification or identification of DGA botnet families, which has received less attention or achieved lower accuracy (LSTM achieved 53%, LSTM.MI achieved 49%), and some DGA botnet families have poor recognition capabilities. Lastly, the evaluation on official datasets is still limited.

From the aforementioned issues, the thesis proposes the following research questions:

- *Research Question 1*: For the problem of DGA botnet detection, are new approaches, including the use of fuzzy clustering algorithms and efficient combinations of machine learning models, effective?

- *Research Question 2*: Can the LSTM network be improved to enhance the effectiveness of DGA botnet detection and classification? What specific solutions can be employed, with a focus on classifying DGA botnets?

- *Research Question 3*: What are the limitations posed by current DGA botnet datasets for algorithm testing, research result comparison, and up-to-dateness? Is it possible to construct a new dataset to address these limitations?

The research findings of the thesis can be applied to Botnet prevention modules in traditional security devices such as Firewalls, IDS, as well as advanced solutions like NGFW and UTM.

2. The research aims

The research aims to achieve the following objectives:

- Investigate the characteristics of DGA botnets: This includes exploring the theoretical foundation, techniques, and related research to understand the nature of DGA botnets, providing a solid basis for developing detection and classification algorithms.

- Research and evaluate the effectiveness of two approaches: Fuzzy clustering algorithms and combined machine learning techniques will be studied and assessed to address the problem of DGA botnet detection. The goal is to determine their efficacy and potential for improving accuracy.

- Propose a new deep learning model based on LSTM networks: The focus will be on developing a novel deep learning model inherited from LSTM networks to detect and classify DGA botnets. The primary objective is to significantly enhance the accuracy compared to previous solutions, with particular emphasis on the classification of DGA botnets.

3. The main research contents of the thesis

To address the research questions, the researcher conducted an overview of DGA botnet detection techniques and related studies. A solution is proposed to improve the accuracy of the detection and classification algorithm for DGA botnets. In addition to traditional approaches, the researcher also implemented new approaches such as using fuzzy clustering algorithms and employing ensemble learning techniques. Furthermore, a new dataset on DGA botnets was constructed with enhancements and updates.

Some specific research contents that the researcher will focus on include:

- Investigating the characteristics, detection techniques, and classification of DGA botnets.

- Researching clustering algorithms on Neutrosophic Set, machine learning, and ensemble learning models to apply them to DGA botnet detection.

- Studying LSTM networks and their variations to improve and propose solutions for detecting and classifying DGA botnets, with a focus on the classification problem.

- Researching specialized datasets on DGA botnets, including Botnet DGA Dataset [19], Andrey Abakumov [20], UMUDGA Dataset [21] [18], DGArchive [22], OSINT DGA feed [23], 360NetLab Dataset [24], Johannes Bader [25], and constructing a new dataset.

4. The contributions of the thesis

The contributions of the thesis achieved through the research process are as follows:

- *Contribution 1*: Proposing three solutions for DGA botnet detection and classification, including NCM, LA_Bin07, and LA_Mul07, to improve accuracy compared to previous solutions.

- *Contribution 2*: Proposing a new specialized dataset, UTL_DGA22, for the DGA botnet problem, serving future research in the same direction."

CHAPTER 1. THEORETICAL FOUNDATIONS OF DGA BOTNET

Chapter 1 presents the theoretical foundations of Botnets in general and DGA botnets in particular. The researcher also introduces two problems within the realm of DGA botnets, which are binary classification and multiclass classification, corresponding to the tasks of detection and classification of DGA botnets. These are the issues that the researcher focuses on studying, addressing, and presenting the results in the subsequent chapters of this thesis.

1.1. General Overview of Botnets

1.1.1. Definition of Botnet

According to Provos & Holz, a Botnet is a 'network of compromised computers that can be remotely controlled by an attacker.'

1.1.2. Evolution of Botnet Technology

Initially, Botnets were designed as useful tools operating on the IRC protocol. Over time, new features were introduced, such as remote control capabilities, command and control infrastructure, modular design, espionage capabilities, sophisticated infection and stealth mechanisms, and the current trend of expanding infections to IoT devices.

1.1.3. Characteristics of Botnets

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

1.1.4. Classification of Botnets

Botnets can be classified based on criteria such as the protocol they use, the type of infected devices, or their architectural design."

1.2. Botnet Detection Techniques

There are main techniques used for Botnet detection:

(1) Honeynet-based techniques.

(2) Intrusion detection system-based techniques:

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

1.3. The Problem of DGA botnet

1.3.1. Overview of DGA botnet

DGA botnet refers to a type of Botnet that is deployed in a Client-Server model. In this model, the Bots act as Clients and connect back to the Command and Control (C&C) server through automatically generated and pre-agreed DNS domain names. The domain generation algorithms are designed to generate domain names that can evade security systems."

1.3.2. The Binary Classification Problem in DGA botnet

The binary classification problem aims to detect domain names generated by DGA botnets. The dataset consists of two labels, 0 and 1.

1.3.3. The Multiclass Classification Problem in DGA botnet

The multiclass classification problem aims to detect the family/type of DGA botnets, given the labeled domain names that are identified as DGA botnets. The dataset consists of n labels, corresponding to the n considered DGA botnet families.

1.3.4. Distinguishing from the Fake URL Detection Problem

The detection of DGA botnets differs from the detection of Fake Uniform Resource Locators (URLs) (Table 1.3).

Table 1.3. Comparison between the Fake Website Detection Problem and the DGA botnet Problem

	Input	**Output**	**Binary classification label**	**Multi classification label**	**Domain Generation Algorithm**
Fake Website Detection	URLs / HTML Code	Trang Web giả mạo	0: Normal Website 1: Fake Website	NULL	No
DGA botnet Problem	Domains	C&C Server IP Address	0: Legitimate domain 1: DGA botnet's domain	n labels corresponding to n DGA botnet families	Yes

1.3.5. Evaluation Dataset for the DGA botnet Problem

I selected four datasets that are considered the most suitable for evaluating the algorithms presented in the subsequent chapters. These datasets include Andrey Abakumov's DGA Repository [20], OSINT DGA feed [23], UMUDGA Dataset [18], and 360NetLab Dataset [24] (Table 1.4).

Table 1.4. Description of the four datasets used in the evaluations.

Dataset	**Binary Classification**	**Multi Classification**	**Number of legitimate samples**	**Number of DGA botnet samples**	**Number of DGA botnet families**
AADR	X	X	1,000,000	801,667	08
OSINT	X		1,000,000	495,186	
UMUDGA	X	X	1,000,000	500,000	50
360NetLab	X		1,000,000	1,513,524	

1.3.6. Evaluation Metrics for the Problem

I evaluates the problem using metrics such as Accuracy, Precision, Recall, and F1-score.

1.3.7. Significance of the DGA botnet Problem

Botnets are continuously evolving, with Bots becoming more sophisticated and capable of causing greater damage.

Applying the operational mechanisms of DGA botnets can provide an effective solution with several advantages, such as not requiring excessive data collection and processing capabilities of the system. The detection of DGA botnet activity can be narrated even after they have infected devices.

Solutions to address the binary classification and multiclass classification problems in DGA botnets can be applied to security solutions such as Firewalls, IDS, NGFW, or UTM.

1.4. Some Research Approaches to Solve the DGA botnet Problem

Several approaches have been proposed to detect Botnets in general and DGA botnets in particular, including signature-based detection, anomaly-based detection, and the use of machine learning and deep learning algorithms.

1.4.1. Approaches Using DNS Analysis Techniques

In the study by Kwon et al. [8], they introduced the PsyBoG solution for detecting malicious behavior based on analyzing a large amount of DNS traffic. PsyBoG has advantages such as (1) detecting stealthy Botnets, (2) handling DNS queries at a large scale, and (3) detecting groups of malicious servers.

Wang et al. proposed a Botnet detection scheme based on domain analysis [9], called DBod. This solution relies on analyzing the behavior of DNS queries.

Chowdhury et al. [42] proposed a novel Botnet detection method based on the structural linkage characteristics of nodes in a graph, which was evaluated on the CTU-13 dataset [43]. The proposed method can effectively detect Bots with different behaviors.

In the study by Bisio et al. [10], they presented a report on a DGA botnet detection algorithm based on a Single Network Monitoring. The test dataset consisted of 40 different DGA botnet families, with high accuracy in detecting most DGA botnet families, ranging from 92.67% and above, with only one case achieving 88.85%.

Wang et al. introduced an approach that includes (1) detecting the presence of Botnets and (2) identifying infected nodes [44]. However, the evaluation results were not extensively discussed.

Trung et al. [11] presented an extended research on Botnets in IoT devices. The authors proposed an IoT Botnet detection method based on extracting attributes from the PSI (PGS-Graph) graph. This solution can overcome the issue of multi-architecture in IoT devices and reduce computational complexity. The experimental results showed that the proposed solution achieved an accuracy of 98.7%.

1.4.2. Machine Learning Approaches

Hieu et al. evaluated the effectiveness of supervised machine learning algorithms for DGA botnet detection [12]. Using domain name input data, they built models including Hidden Markov, C4.5 Decision Tree, and Extreme Learning Machine. They also experimented with SVM, Recurrent SVM, CNN combined with LSTM, and Bidirectional LSTM models. The classifiers performed effectively for the binary classification problem but had limitations for multiclass classification.

Khan et al. considered the detection of Peer-to-Peer Botnets [13]. They proposed a multiclass traffic classification method using machine learning models. The average accuracy achieved was 98.7%. The experiments were conducted on the CTU-13 and ISOT datasets.

Zago et al. presented a study on a new dataset for DGA botnet detection. The research team compiled and constructed the UMUDGA Dataset [18][21]. This dataset includes 1,000,000 benign domain names combined with 50 different DGA botnet families.

Xuan et al. [14] proposed improvements to a machine learning model. They suggested using new features and applied them to the Random Forest algorithm. The model achieved a false alert rate below 3.02% and an F1-score of 97.03% in evaluations.

Alauthman et al. [50] proposed a mechanism to reduce complex traffic and integrated it with reinforcement learning. The experimental results showed a detection accuracy of 98.3% and a low false positive rate of 0.012%. The study was evaluated on a synthesis of three datasets, including the ISOT Dataset, P2P Botnet, and Information Security Centre of Excellence Dataset.

1.4.3. Deep Learning Approaches

In the deep learning approach, Duc et al. [15] used Long Short-Term Memory (LSTM) networks to address both DGA botnet problems. The research team proposed a new algorithm called LSTM.MI. The experimental results showed that the proposed algorithm improved the accuracy by at least 7% compared to the traditional LSTM model and achieved high accuracy in binary classification with an F1-score of 98.49%. It also demonstrated the ability to recognize five additional Botnet families.

Curtin et al. used RNN networks to detect and classify DGA botnets [16]. They proposed a new concept called the Smashword score. The experiments showed that the new model had high potential and improved accuracy compared to previous models.

Vinayakumar et al. researched the detection of malicious domain names generated by Botnets or malicious emails and URLs [17]. The evaluation dataset included benign and malicious domain names collected from OpenDNS, Alexa, and OSINT Feeds. The comparison results showed that the CNN-LSTM model was the most effective with an F1-score of 96.3% in binary classification.

1.5. Conclusion of Chapter 1

In Chapter 1, I provided a general overview of Botnets, specifically the DGA botnet problem, and related research. The research scope of the thesis focused on the DGA botnet problem, specifically on the classification problem. Chapter 1 also discussed the differences between the DGA botnet detection problem and the detection of malicious URLs, highlighting the significance of the DGA botnet problem.

In the subsequent chapters, I will present research results, evaluations, and propose deep learning-based solutions to address the binary classification and multiclass classification problems for the detection and classification of DGA botnets.

Some of the research results presented in Chapter 1 have been published in [CT2] [CT6] in the author's list of publications."

CHAPTER 2. EVALUATING THE SOLUTION FOR DGA BOTNET DETECTION USING FUZZY SET THEORY AND MACHINE LEARNING

In Chapter 2, I approaches the binary classification problem of detecting DGA botnets using fuzzy set theory and machine learning. I also proposes two combined machine learning models, VEA and HEA, to improve accuracy compared to single models. The objective of this chapter is to evaluate the effectiveness of the two approaches, based on fuzzy set theory and machine learning, in the binary classification problem.

2.1. Detecting DGA botnets based on fuzzy set theory

2.1.1. Fundamentals of fuzzy clustering algorithm

The Fuzzy C-Means (FCM) algorithm, proposed by Bezdek, is a fuzzy clustering algorithm that has been widely applied in various fields.

The Neutrosophic C-Means (NCM) clustering algorithm on Neutrosophic Set was proposed by Gou et al. [57]. Given X as a non-empty set with an element denoted as x ∈ X, a neutrosophic set A defined on the space X is characterized by three functions:

- The membership function $T_A(\mathrm{x})$ measures the degree of membership indicating the likelihood of event x occurring.

- The indeterminacy function $I_A(\mathrm{x})$ measures the degree of neutrality, indicating no opinion on whether event x will occur or not.

- The non-membership function $F_A(\mathrm{x})$ measures the degree of non-membership, believing that event x will not occur.

The objective function:

$$J_{NCM}(T,I,F,c) = \sum_{i=1}^{N}\sum_{j=1}^{C}(\omega_1 T_{ij})^m \|x_i - c_j\|^2 + \sum_{i=1}^{N}(\omega_2 I_i)^m \|x_i - \bar{c}_{i_{max}}\|^2 + \delta^2 \sum_{i=1}^{N}(\omega_3 F_i)^m \quad (2.1)$$

The Neutrosophic C-Means (NCM) Clustering algorithm can be summarized as follows:

Algorithm: $NCM(X, \varepsilon)$
Input X, ε
Output k
$init(T^{(0)}, I^{(0)}, F^{(0)})$
$init(C, m, \varepsilon, \delta,\ \omega_1,\ \omega_2,\ \omega_3)$
do:
$calculate(c_i^{(k)})$:
$calculate(\bar{c}_{i_{max}})$
$update(T^{(\mathrm{k+1})})$
$update(I^{(\mathrm{k+1})})$
$update(F^{(\mathrm{k+1})})$
while $\left|T_{ii}^{(k+1)} - T_{ii}^{(k)}\right| > \varepsilon$
$TM = [T, I, F]$ *với* $x(i) \epsilon\ k^{th}$
$k = argmax(TM_{ij})$ *với* $j = 1, 2, \ldots, C+2$
return k

2.1.2. DGA botnet Detection Algorithm with NCM

The stages of the NCM fuzzy clustering-based solution for DGA botnet detection are illustrated in Figure 2.1.

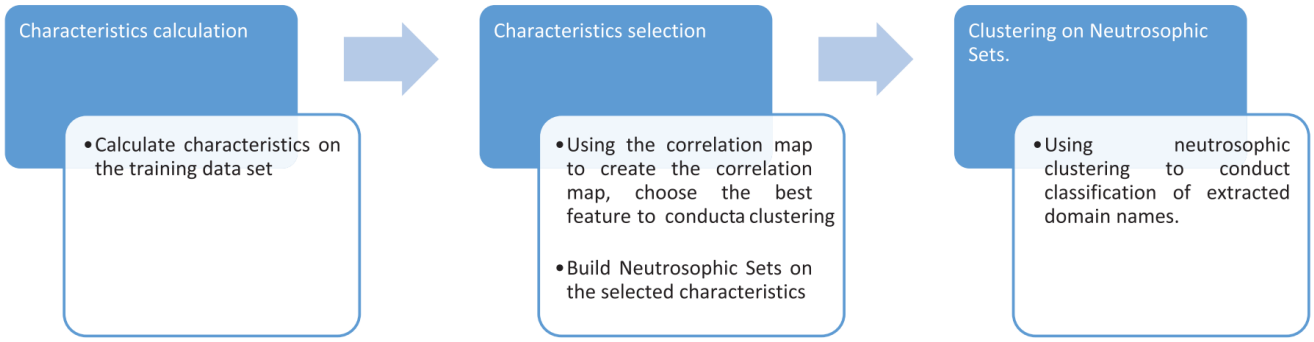


Figure 2.1. Model applying the NCM algorithm for DGA botnet detection.

I propose features based on domain names and utilizes the Pearson correlation coefficient to select the most influential features. The results of feature selection are provided in Table 2.4.

Table 2.4. Features with the highest impact in the datasets.

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

2.1.3. Evaluation and Discussion

The evaluation was performed on the AADR, 360NetLab, OSINT, and UMUDGA datasets. The results are presented in Table 2.5.

Table 2.5. Binary classification results of the NCM algorithm.

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

In general, the NCM algorithm performs well in classifying benign and malicious domain names, achieving the best overall performance on the 360NetLab and UMUDGA datasets with an F1-score of 0.84. The lowest performance is observed on the OSINT dataset, with an F1-score of only 0.54.

A comparison between the proposed NCM algorithm and other fuzzy-based algorithms such as Sahin, K-means, FCM, SVM, TSVM, and FSVM was conducted. The results of this comparison are presented in Table 2.6.

Table 2.6. Comparison of NCM with other similar algorithms."

Bảng 2.6. So sánh NCM với một số thuật toán tương tự

Phương pháp	Precision	Recall	F_1-score
*NCM**	*0,85*	*0,75*	*0,75*

Sahin	0,48	0,80	0,60
K-means	0,74	0,86	0,79
FCM	0,76	0,83	0,79
SVM	0,82	0,72	0,77
TSVM	0,80	0,81	0,81
FSVM	0,83	0,73	0,78

** Average performance across the 4 datasets*

The model utilizing the NCM algorithm demonstrates a good balance between runtime and accuracy.

2.2. Machine Learning-based DGA botnet Detection

2.2.1. Model for evaluating machine learning algorithms

The stages involved in evaluating machine learning algorithms for binary classification are illustrated in Figure 2.8.

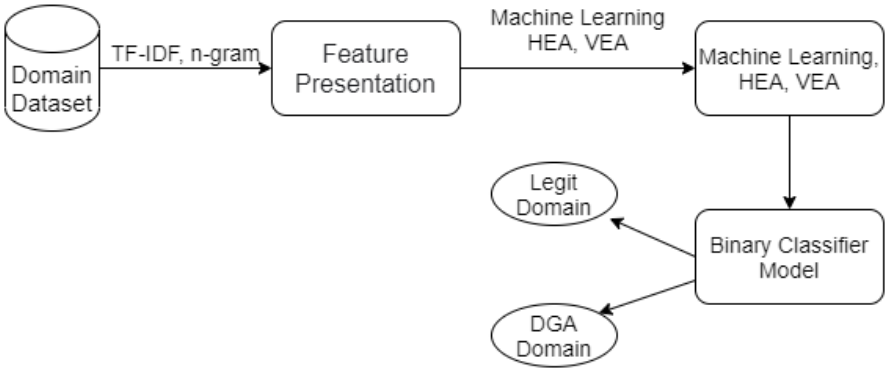


Figure 2.8. Diagram of the training and evaluation model

I conducted experiments with several machine learning algorithms, including Support Vector Machines (SVM), Logistic Regression (LR), Naive Bayes (NB), Neural Networks (NN), Decision Trees (DT), Random Forests (RF), k-Nearest Neighbour (kNN), and Adaptive Boosting (AB). The evaluation was performed on the UMUDGA Dataset.

2.2.2. Evaluation results and discussion

Table 2.8 presents the Precision, Recall, and F1-score for the binary classification problem of DGA botnet detection.

Table 2.8. Detection results of DGA botnet using machine learning on the UMUDGA Dataset.

Method	**Precision**	**Recall**	**F_1-score**
Logistic Regression	0,92	0,97	0,97
Naive Bayes	0,98	0,84	0,90
Decision Tree	0,93	0,95	0,94
Neural Network	0,97	0,97	0,97
Support Vector Machine	0,97	0,97	0,97
Random Forrest	0,74	0,82	0,77
K-Nearest Neighbor	0,97	0,66	0,78
Adaptive Boosting	0,83	0,85	0,84

Most of the machine learning algorithms achieved high accuracy in the binary classification problem, including LR, NN, and SVM. The NN model yielded the highest overall result with an F1-score of 0.97, while the RF model had the lowest result with 0.78.

2.2.3. Ensemble Machine Learning Models

I propose two models, VEA and HEA, which are stronger classifiers that operate based on ensemble learning mechanisms. The results of the binary classification problem using the VEA and HEA models are presented in Table 2.9.

Table 2.9. Detection results of DGA botnet using VEA and HEA models on the UMUDGA Dataset

Algorithm	Precision	Recall	F_1-score
Average	0,92	0,88	0,89
Neural Network (the strongest)	0,97	0,97	0,97
Random Forest (the weakest)"	0,74	0,82	0,77
VEA	0,98	0,99	0,98
HEA	0,97	0,97	0,97

Both the VEA and HEA models have improved accuracy compared to the average values of individual models.

Limitations of the NCM and machine learning solutions:

- Accuracy can still be further improved.
- Training time can be demanding as it runs on CPUs.
- Not suitable for multiclass classification problems.

2.3. Conclusion of Chapter 2

In Chapter 2, I presented the research findings and evaluation of the approaches using fuzzy set theory and machine learning for DGA botnet detection.

In the approach using fuzzy set theory, I proposed the application of the Neutrosophic C-Means (NCM) clustering algorithm for DGA botnet detection. I made adjustments to the original algorithm to fit the DGA botnet problem. Evaluation on four standard datasets showed that the NCM algorithm achieved similar accuracy to other fuzzy-based approaches. Additionally, it significantly improved precision compared to these classifiers. The computational time of the model achieved a balance between processing time and accuracy. The NCM model also identified noisy or neutral elements to provide additional testing suggestions. Although it had fast execution speed, the limitation of the NCM model was that its accuracy was considerably lower than machine learning algorithms.

In the approach using machine learning, the evaluation on four standard datasets showed that the machine learning models achieved significantly higher accuracy, with the Neural Network algorithm having the highest F1-score of 0.97. This accuracy was further improved when using the ensemble models based on voting mechanisms, namely VEA and HEA, proposed by the author.

Both approaches have their advantages and limitations in terms of time or accuracy. Furthermore, their application to multiclass classification is still limited. These aspects can be addressed by deep learning models. I proposed and presented these aspects in Chapter 3 of the dissertation.

A portion of the research findings presented in Chapter 2 has been published in [CT1] [CT3] in the author's publication list.

CHAPTER 3. DETECTION AND CLASSIFICATION SOLUTION FOR DGA BOTNET USING DEEP LEARNING TECHNIQUES

Based on previous research and evaluations, I propose a deep learning solution to improve the accuracy of detecting and classifying DGA botnet. I introduce two new deep learning models, LA_Bin07 and LA_Mul07, which are enhancements of the traditional LSTM network. The evaluations show that the proposed models significantly improve accuracy compared to previous studies, particularly in DGA botnet classification.

3.1. Deep Learning Technical Foundation

3.1.1. Recurrent Neural Network (RNN)

Recurrent Neural Network (RNN) is designed to train with sequential input data.

3.1.2. Long-Short Term Memory (LSTM)

LSTM is a type of RNN that can capture long-term dependencies in the data. It was first proposed by Hochreiter in 1996 and has been continuously improved since then.

3.1.3. Attention Mechanism

The attention mechanism consists of three components: Query, Key, and Value. It is proposed to help the model focus on specific features of the data during training.

3.2. Two New Deep Learning Models for DGA botnet Detection and Classification

The solution utilizes two models, LA_Bin07 and LA_Mul07, as shown in Figure 3.8, for DGA botnet detection and classification.

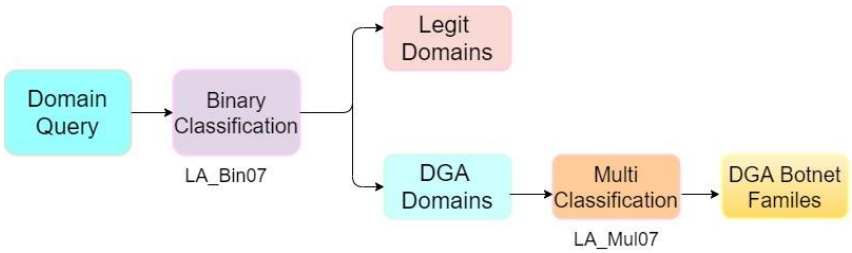


Figure 3.8. Solution for DGA botnet detection and classification using two new deep learning models, LA_Bin07 and LA_Mul07

3.2.1. LA_Bin07 Model for DGA botnet Detection

The LA_Bin07 model is designed in a Sequence-to-Sequence format, with its architecture shown in Figure 3.9:"

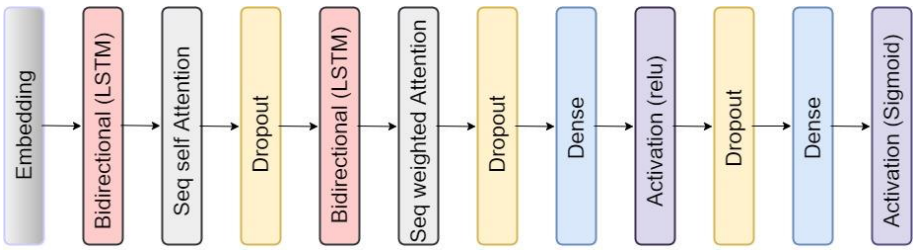


Figure 3.9. Architecture of the LA_Bin07 model.

3.2.2. Architecture of the LA_Mul07 Model for DGA botnet Classification

The layer structure of the LA_Mul07 model is shown in Figure 3.10:

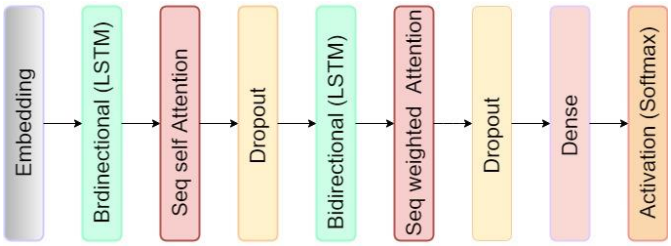


Figure 3.10. Architecture of the LA_Mul07 model.

3.2.3. Improvements over Traditional LSTM

The addition of Attention helps the model prioritize learning important parameters over others, resulting in improved accuracy compared to the original LSTM. Additionally, the order, weights, and sizes of the layers are optimized to achieve the highest effectiveness.

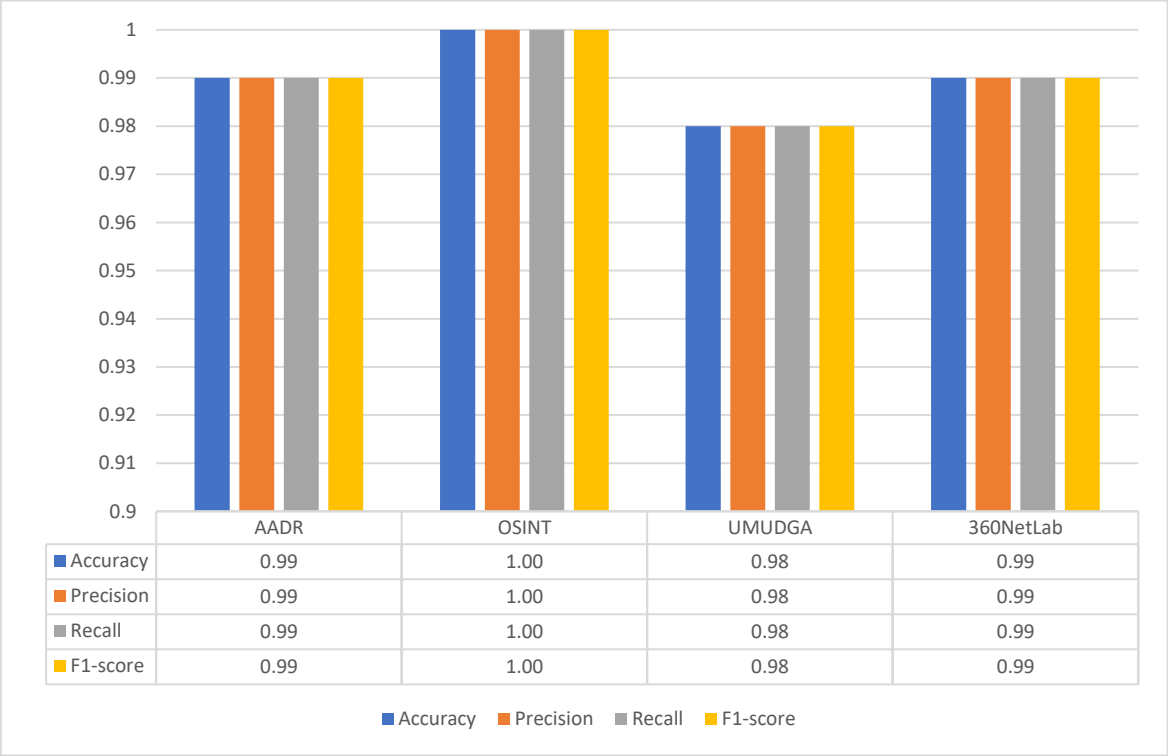
3.3. Evaluation of the Proposed Deep Learning Models

3.3.1. Dataset and Evaluation Environment

I utilized all four datasets for the Binary Classification task and two datasets, including Andrey Abakymov's DGA Repository and UMUDGA Dataset, for the Multi-Classification task.

3.3.2. Evaluation of the LA_Bin07 Model for DGA botnet Detection

The evaluation results of the LA_Bin07 model using the Accuracy, Precision, Recall, and F1-Score parameters are provided in Figure 3.13.



	AADR	OSINT	UMUDGA	360NetLab
Accuracy	0.99	1.00	0.98	0.99
Precision	0.99	1.00	0.98	0.99
Recall	0.99	1.00	0.98	0.99
F1-score	0.99	1.00	0.98	0.99

Figure 3.13. Evaluation results of the LA_Bin07 model for the Binary Classification task on the four standard datasets

The LA_Bin07 model achieves very high accuracy, with an Accuracy score of 0.98 or higher on all four evaluated datasets. Notably, the LA_Bin07 model predicts with perfect accuracy (1.00) on the OSINT dataset.

3.3.3. Evaluation of the LA_Mul07 Model for DGA botnet Classification

The LA_Mul07 model is evaluated on two datasets, AADR and UMUDGA, as they are labeled with different DGA botnet families.

The evaluation results on the AADR dataset are presented in Table 3.3.

Table 3.3. Evaluation results of the LA_Mul07 model on the AADR dataset.

STT	DGA botnet	Precision	Recall	F_1-Score
1	cryptolocker	1,00	0,98	0,99
2	zeus	1,00	1,00	1,00
3	pushdo	1,00	1,00	1,00
4	rovnix	1,00	1,00	1,00
5	tinba	0,98	1,00	0,99
6	conficker	1,00	1,00	1,00
7	matsnu	1,00	1,00	1,00
8	ramdo	1,00	1,00	1,00
Avg Accuracy		***1,00***		

The evaluation results on the UMUDGA dataset are provided in Table 3.4.

Table 3.4. Evaluation results of the LA_Mul07 model on the UMUDGA dataset

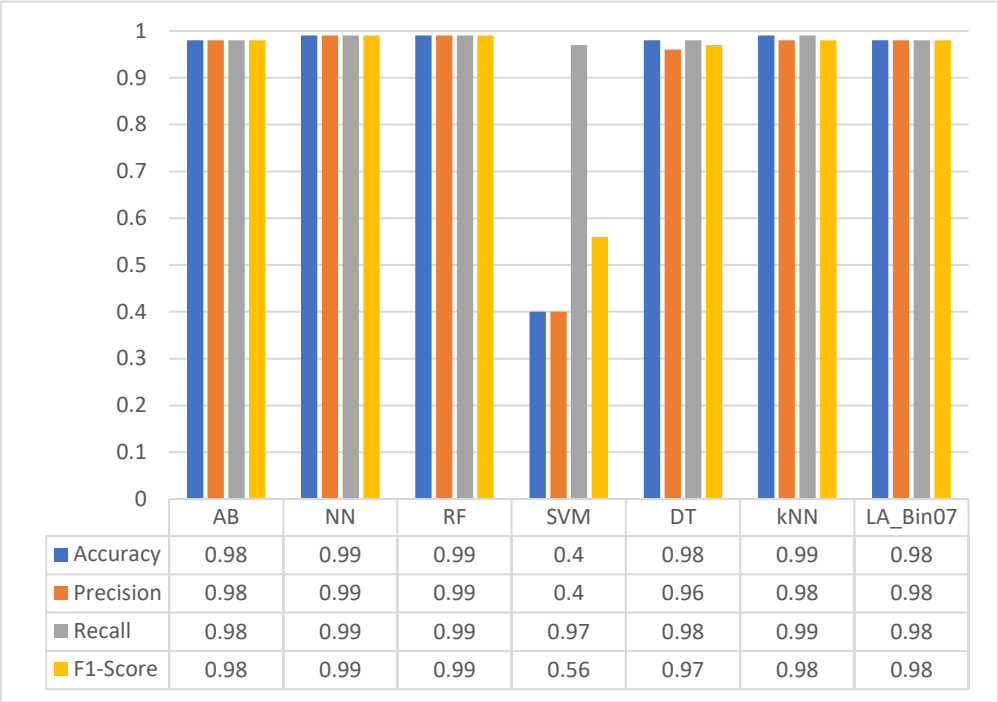
STT	DGA botnet	Pre	Re	F_1	STT	DGA botnet	Pre	Re	F_1
1	alureon	0,45	0,92	0,60	26.	pizd	0,97	0,86	0,91
2	banjori	0,99	1,00	1,00	27.	proslikefan	0,82	0,65	0,73
3	bedep	0,96	0,47	0,63	28.	pushdo	0,99	0,99	0,99
4	ccleaner	1,00	1,00	1,00	29.	pykspa	0,39	0,57	0,47
5	china	1,00	0,99	1,00	30,	pykspa_noise	0,35	0,16	0,22
6	corebot	1,00	1,00	1,00	31.	qadars	0,99	0,99	0,99
7	cryptoloker	0,70	0,66	0,68	32.	qakbot	0,84	0,55	0,67
8	dircrypt	0,52	0,42	0,47	33.	ramdo	1,00	1,00	1,00
9	dyre	1,00	1,00	1,00	34.	ramnit	0,44	0,66	0,52
10	fobber_v1	0,88	1,00	0,93	35.	ranbyus_v1	0,76	0,98	0,86
11	fobber_v2	0,48	0,08	0,14	36.	ranbyus_v2	0,76	0,88	0,82
12	gozi_gpl	0,96	0,99	0,98	37	rovnix	0,97	0,94	0,95
13	gozi_luther	0,97	0,95	0,96	38	shiotob	1,00	0,90	0,95
14	gozi_nase	0,89	0,97	0,93	39	simda	1,00	1,00	1,00
15	gozi_rfc4343	0,91	0,98	0,90	40	aaron	1,00	1,00	1,00
16	kraken_v1	0,72	0,96	0,83	41	suppobox_1	0,87	0,97	0,92
17	kraken_v2	0,82	0,41	0,55	42	suppobox_2	0,98	1,00	0,99
18	locky	0,84	0,62	0,71	43	suppobox_3	0,99	1,00	1,00
19	matsnu	0,98	0,94	0,96	44	symmi	1,00	1,00	1,00
20	murofet_v1	0,99	1,00	1,00	45	tempedreve	0,58	0,86	0,69
21	murofet_v2	0,94	0,96	0,95	46	tinba	0,77	0,97	0,86
22	murofet_v3	1,00	1,00	1,00	47	vawtrak_v1	1,00	1,00	1,00
23	necurs	0,99	0,80	0,89	48	vawtrak_v2	0,99	1.00	1.00
24	maim	0,95	0,94	0,95	49	vawtrak_v3	1.00	1.00	1.00
25	padcrypt	1.00	1.00	1.00	50	zeus_newgoz	1.00	1.00	1.00
Avg Accuracy					***0,86***				

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

3.4. Evaluation with related studies

3.4.1. General evaluation on the UMUDGA dataset

For the binary classification task, the comparative results are shown in Figure 3.21.

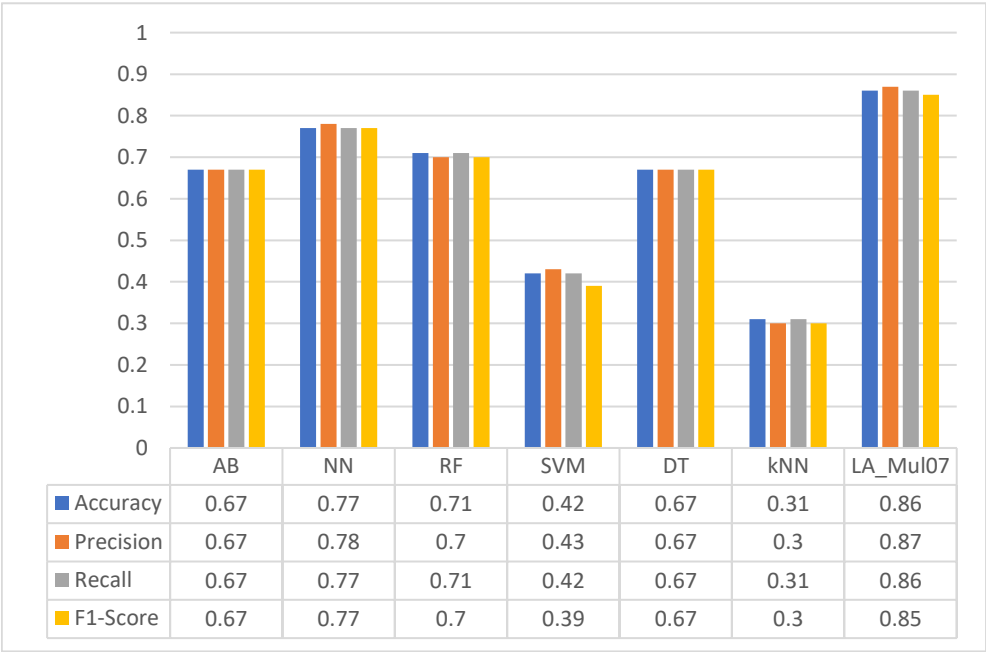


	AB	NN	RF	SVM	DT	kNN	LA_Bin07
Accuracy	0.98	0.99	0.99	0.4	0.98	0.99	0.98
Precision	0.98	0.99	0.99	0.4	0.96	0.98	0.98
Recall	0.98	0.99	0.99	0.97	0.98	0.99	0.98
F1-Score	0.98	0.99	0.99	0.56	0.97	0.98	0.98

Figure 3.21. Comparison of the LA_Bin07 classifier with machine learning algorithms on the UMUDGA dataset

The results show that the LA_Bin07 model achieves significantly higher accuracy than the SVM model. Additionally, it performs comparably to the AB, NN, RF, DT, and kNN models.

For the multi-class classification task, the comparative results are summarized and presented in Figure 3.22.



	AB	NN	RF	SVM	DT	kNN	LA_Mul07
Accuracy	0.67	0.77	0.71	0.42	0.67	0.31	0.86
Precision	0.67	0.78	0.7	0.43	0.67	0.3	0.87
Recall	0.67	0.77	0.71	0.42	0.67	0.31	0.86
F1-Score	0.67	0.77	0.7	0.39	0.67	0.3	0.85

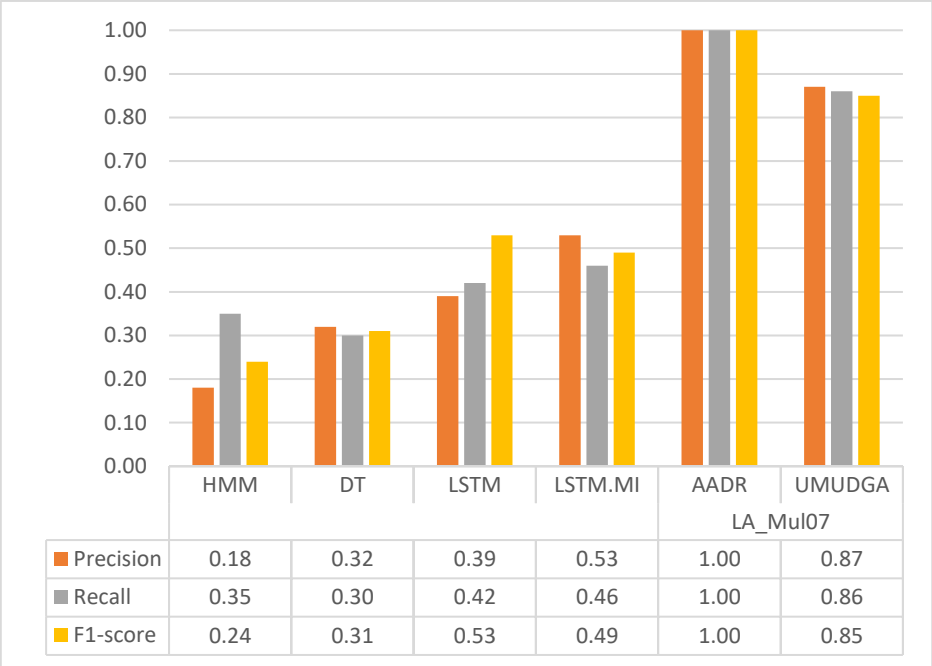
Figure 3.22. Comparison of the LA_Mul07 classifier with machine learning algorithms on the UMUDGA dataset

For the multi-class classification task, the LA_Mul07 model achieves significantly higher accuracy than the other machine learning models.

3.4.2. Evaluation with other deep learning models

I also evaluated several 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 achieve the best performance among the tested models.

3.4.3. Evaluation with other relevant studies in multi-class classification



	HMM	DT	LSTM	LSTM.MI	AADR	UMUDGA
Precision	0.18	0.32	0.39	0.53	1.00	0.87
Recall	0.35	0.30	0.42	0.46	1.00	0.86
F1-score	0.24	0.31	0.53	0.49	1.00	0.85

Figure 3.1. Figure 3.23. Comparison of the LA_Mul07 model with other models

The LA_Mul07 model also outperforms previously proposed deep learning models.

3.5. Conclusion of Chapter 3

In Chapter 3, I presented the research results on improving deep learning models for addressing the DGA botnet problem. I proposed two new deep learning models, LA_Bin07 and LA_Mul07, to respectively tackle the tasks of DGA botnet detection and classification. These models are built upon the traditional LSTM network with improvements. The two models were thoroughly evaluated on four datasets, including Andrey Abakumov's DGA Repository, OSINT DGA feed, UMUDGA Dataset, and 360NetLab Dataset.

The evaluations showed that the LA_Bin07 model achieved high accuracy, ranging from 0.98 on the UMUDGA dataset to a perfect score of 1.00 on the OSINT dataset. The LA_Mul07 model demonstrated high capability in classifying DGA botnet families, significantly improving over previous models with accuracies of 1.00 and 0.86 on the AADR and UMUDGA datasets, respectively.

Solving the DGA botnet problem carries significant implications for network security, especially in the area of DGA botnet classification. Firstly, this approach can quickly provide detection and classification alerts for DGA botnets when integrated into Firewall/IDS devices. Secondly, this solution requires fewer computational resources compared to traditional packet analysis approaches and takes advantage of GPU computing capabilities. Thirdly, the solution

can be extended to address malicious code, malware, and spyware with similar domain query mechanisms. Finally, the malicious domain detection module can be integrated into advanced and modern security solutions such as next-generation firewalls and unified security solutions.

Part of the research results presented in Chapter 3 has been published in [CT4] in the author's publication list.

CHAPTER 4. SPECIALIZED UTL_DGA22 DATASET FOR DGA BOTNET PROBLEM

In Chapter 4, I propose a new specialized dataset for the DGA botnet problem called UTL_DGA22. This dataset builds upon the results of previous datasets while incorporating new features and improvements. These include the addition of new DGA botnet families, data normalization and labeling, proposed and pre-selected new attributes, and detailed documentation. I also evaluates the proposed solutions from Chapter 2 and Chapter 3, including NCM, VEA, HEA, LA_Mul07, and LA_Bin07, on the new dataset, which yields good results. The UTL_DGA22 dataset is expected to serve as a reliable, publicly available, objective, and comprehensive foundation for researchers to experiment, compare, and evaluate their own solutions in the future.

4.1. DGA botnet Dataset Problem Statement

4.1.1. Overview of the Problem

The proposed solutions in previous research studies are often evaluated on datasets collected by research teams at different time points, with uneven sample sizes, limited public availability, and lack of convenience for comparison.

4.1.2. General Botnet Datasets

There are several general botnet datasets such as CTU-13 [86], UGR16 [87], DreLAB [88], UNSW-NB15 [89], ISCX-Bot-2014 [40]. However, these five datasets were not specifically designed to evaluate the DGA botnet problem as they lack domain names of DGA botnet families and corresponding labels.

4.1.3. Specialized DGA botnet Dataset

Table 4.3 summarizes the key characteristics of the current popular DGA botnet datasets mentioned above.

Table 4.3. Key characteristics of the current popular DGA botnet datasets

No.	Dataset	Symbols	Number of legitimate domains	Number of DGA botnet domains	Number of DGA botnet Families	Format	Publish
1	Andrey Abakumov's DGA Repository [20]	AADR	1.000.000	801.667	08	txt	2016
2	Johannes Bader's Domain Generation Algorithms Repository [25]	JBR	Null	N/A	48	txt	2018
3	Alexa Top 1 milion domains [47]	AT1D	1.000.000	0	0	csv	2019
4	Botnet DGA Dataset [92]	BDD	1.000.000	1.803.333	10	csv	2020
5	UMUDGA Dataset [18]	UMU	1.000.000	over 30.000.000	50	arff, csv, txt	2020
6	DGArchive by Fraunhofer FKIE [22]	DFF	Null	N/A	86	csv	2020
7	OSINT DGA feed [23]	OSINT	1.000.000	495.186	0	txt	2021

8	360NetLab Dataset [24]	360NL	0	Not fixed	Not fixed	txt	2021
9	The Majestic Million [96]	TMM	1.000.000	0	0	csv	2021

4.1.4. 4.1.4. Research Problem

There are differences in structure and purpose between datasets for Botnet in general and datasets for DGA botnet in particular. I categorizes and presents the details in Table 4.4.

Table 4.4. Evaluation of characteristics of dataset groups for Botnet

Dataset	**Group**	**Botnet Detection**	**DGA botnet Detection**	**Network Attack Detection**	**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*	✓	✓	×	×	×	✓	✓

Table 4.5 summarizes the advantages and limitations of existing DGA botnet datasets, as well as the proposed UTL_DGA22 dataset.

Table 4.5. Overview of advantages and limitations of existing DGA botnet datasets and the proposed UTL_DGA22 dataset

Dataset	**Binary Classification**	**Multi Classification**	**Origin domain**	**Features Extraction**	**Public**	**Document**
AADR	✓	✓	✓	×	✓	N/A
JBR	✓	✓	✓	×	✓	✓
AT1D	✓	×	✓	×	✓	N/A
BDD	✓	×	×	✓	✓	✓
UMU	✓	✓	✓	✓	✓	✓
DFF	✓	✓	✓	×	✓	✓
OSINT	✓	×	✓	×	✓*	N/A
360NL	✓	✓	✓	×	✓	N/A
TMM	✓	×	✓	×	✓	N/A
UTL	✓	✓	✓	✓	✓	✓

4.2. The Proposed UTL_DGA22 Dataset

The UTL_DGA22 dataset is introduced, which fulfills all the aforementioned requirements and incorporates new data and extracted features.

4.2.1. Construction of the Dataset

The UTL_DGA22 dataset is constructed through the following steps:

- Gathering benign domain data and DGA botnet domain data from reliable sources.
- Aggregating and evaluating the independent DGA botnet families.
- Enriching the domain database.
- Proposing two groups of new attributes, including Base-Features and TF-IDF Features, and feature extraction.

4.2.2. Proposed Attributes

I propose two groups of attributes: domain-based attributes (Base) and subdomain-based attributes (TF-IDF).

4.2.3. Dataset Storage Structure

The UTL_DGA22 dataset is structured into two corresponding directories: DGA_Botnets_Domains and DGA_Botnets_Features_Extraction.

4.3. DGA botnet Families in the UTL_DGA22 Dataset

The UTL_DGA22 dataset comprises 76 distinct DGA botnet families, listed in Table 4.9.

4.4. Evaluation of the Proposed Attributes

I utilize the proposed attributes as inputs for basic machine learning models for binary and multiclass classification tasks. Both the Base and TF-IDF attribute groups demonstrate their suitability as inputs for machine learning algorithms in addressing binary and multiclass classification problems.

4.5. Evaluation of the Proposed Solutions on the UTL_DGA22 Dataset

I perform evaluations of the proposed solutions using the new UTL_DGA22 dataset, including the NCM algorithm, machine learning methods, VEA, HEA, LA_Bin07, and LA_Mul07. The results are presented in Tables 4.11, 4.13, and 4.14.

Table 4.11. Evaluation Results of the NCM algorithm on the UTL_DGA22 dataset

Label	Precision	Recall	F_1-Score
0	0,66	0,93	0,77
1	0,91	0,48	0,62
Avg*	*0,79*	*0,70*	*0,70*

Table 4.13. Evaluation Results of the Proposed Machine Learning Algorithms on the UTL_DGA22 Dataset

Method	Acc.	Pre.	Re.	F_1.	Traning Time	Testing Time
Logistic Regression	0,96	0,97	0,95	0,96	75,59	0,04
Naive Bayes	0,90	0,91	0,86	0,86	1,34	0,14
Decision Tree	0,90	0,90	0,88	0,89	16.004,87	0,71
Neural Network	0,97	0,97	0,96	0,96	4.777,21	0,96

Support Vector Machine	0,96	0,97	0,94	0,96	11,04	0,03
Random Forrest	0,60	0,99	0,07	0,14	62,57	4,06
K-Nearest Neighbor	0,80	0,97	0,56	0,71	0,47	19.462,80
Adaptive Boosting	0,84	0,84	0,79	0,81	2.917,49	12,22
VEA	0,97	0,97	0,96	0,96	16.803,67	18,45
HEA	0,97	0,97	0,95	0,96	7.425,03	10,10

Table 4.14. Evaluation Results of the LA_Bin07 Model on the UTL_DGA22 Dataset

Label	**Precision**	**Recall**	**F_1-Score**
Lành tính	0,98	0,98	0,98
DGA botnet	0,98	0,97	0,97
Accuracy	*0,98*		

The LA_Mul07 model achieves an accuracy of 0.86 when evaluated on the UTL_DGA22 dataset.

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

4.6. Conclusion of Chapter 4

In Chapter 4, I presented research outcomes regarding datasets for binary and multiclass classification in the DGA botnet problem. Specifically, the limitations of existing datasets were discussed in detail, followed by the proposal of a new dataset, UTL_DGA22, built upon previous achievements and incorporating enhancements, new data, and attributes.

The main contributions include:

- The proposal of the DGA_UTL22 dataset, consisting of 76 popular DGA botnet families, with each family containing 20,000 domain samples. The domains are presented in their original form and stored in CSV, ARFF, and TXT file formats.

- The proposal of two attribute groups, including 36 attributes in the Base-Features group for domains and TF-IDF attributes for DGA botnet families. These attributes are extracted, stored in CSV and ARFF formats, and shared along with the dataset. Evaluations using these attributes with machine learning algorithms yielded positive results.

- I also conducted evaluations on the new UTL_DGA22 dataset using the proposed algorithms in the thesis, including NCM, machine learning methods, VEA, HEA, LA_Bin07, and LA_Mul07. The results demonstrated the suitability of the proposed algorithms and their comparable accuracy to evaluations on previous datasets.

I expect UTL_DGA22 to serve as a reliable common foundation widely used to facilitate performance evaluations of new proposed solutions in a convenient, objective, and easily comparable manner.

A portion of the research outcomes presented in Chapter 4 has been published in [CT5] in the author's publication list."

CONCLUSION

After a period of study and research at the Graduate University of Sciences and Technology, Vietnam Academy of Science and Technology, the research on " Research focuses on improving several machine learning and deep learning models for classifying DGA botnet" has been successfully completed, achieving the following main results:

- Presenting a theoretical foundation of DGA botnet, the problem of DGA botnet detection and classification. Evaluating the effectiveness of two approaches using the fuzzy clustering algorithm NCM, machine learning with two combined models VEA and HEA in DGA botnet detection.

- Proposing two new deep learning models, LA_Bin07 and LA_Mul07, for DGA botnet detection and classification. The evaluations have shown that these two new models achieve high accuracy in classification, especially in the task of classifying DGA botnet families.

- Proposing the UTL_DGA22 dataset specifically designed for evaluating the DGA botnet problem. This dataset can be considered as the most comprehensive dataset with complete source code and documentation on DGA botnet as of the present time."

The research results have answered the research questions and provided practical significance in the detection and classification of DGA botnets.

- For research question 1: In the task of DGA botnet detection, the evaluations have shown that applying the fuzzy clustering algorithm NCM for DGA botnet detection is faster but less accurate compared to machine learning models. On the other hand, machine learning models achieve high accuracy, with the highest Precision reaching 0.97, but they require more training time. The proposed deep learning model, LA_Bin07, has improved the classification accuracy, ranging from 0.98 to 1.00 when evaluated on different datasets.

- For research question 2: The traditional LSTM network can be improved to enhance the effectiveness of DGA botnet detection and classification. The proposed LA_Bin07 and LA_Mul07 deep learning models have addressed this issue. In particular, the LA_Mul07 model has significantly improved the accuracy of DGA botnet classification compared to previous solutions. The model achieved the lowest accuracy of 0.86 and the highest accuracy of 1.00 when evaluated on three dedicated datasets. The deep learning approach also has the advantage of faster training time by leveraging the computational power of GPUs.

- For research question 3: The previous DGA botnet datasets have limitations in their applicability to specific tasks, accuracy, sample size, sample balance, lack of new DGA botnet families, documentation, and proposed attributes. I proposed the new UTL_DGA22 dataset specifically designed for evaluating the DGA botnet problem. This dataset inherits all the advantages of previous datasets while also adding additional data samples, attributes, and comprehensive descriptions. Since its publication in early 2023, I has received 6 access requests for research purposes from international scientists. It is expected that UTL_DGA22 will become a widely used standard dataset for researchers and cybersecurity experts in the future for their evaluation tasks.

My research results have been published in 4 papers in reputable specialized scientific journals listed in the ISI/Scopus database and presented in the author's "Publication List" and the attached "Appendix".

In addition to the achieved results, I has identified several directions for future development to further improve the models. Specifically, the improvement of the LA_Mul07 model to increase the accuracy beyond the current level of 0.86 on datasets with more than 50 labels. Building dedicated detection mechanisms for DGA botnet families that exhibit high similarity or are variations of each other.

In terms of application, I plan to inherit the LA_Bin07 and LA_Mul07 models and integrate them into a module for detecting and classifying DGA botnets based on domain names. This module can be integrated into advanced and modern network security solutions such as next-generation firewalls or unified security solutions."