



Early Detection of DDoS attacks and Enhancing Feature Selection using Network Traffic Analysis with Machine Learning Techniques

D R A ISURU KARUNARATHNA
(Reg. No.: MS22044518)

A THESIS
SUBMITTED TO
SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE IN INFORMATION TECHNOLOGY
(CYBER SECURITY)

December 2024

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

.....

Mr. Amila Senarathna (Supervisor)

Approved for MSc. Research Project:

MSc. Programme Co-ordinator, SLIIT

Approved for MSc:

Head of Graduate Studies, FoC, SLIIT

DECLARATION

This is to certify that the work is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources). The work has not previously been submitted in any form to the Sri Lanka Institute of Information Technology or to any other institution for assessment for any other purpose.

Sign: 

D R A Isuru Karunarathna

Date: ...03/11/2024.....

ABSTRACT

Early Detection of DDoS attacks and Enhancing Feature Selection using Network Traffic Analysis with Machine Learning Techniques

D.R.A Isuru Karunarathna

MSc. in Information Technology

Supervisor: Mr. Amila Senarathna

December 2024

Distributed Denial-of-Service (DDoS) attacks are a very serious and developing menace to many providers of online services. Web services have become more important because of new technology, making them appealing targets. DDoS means Distributed Denial of Service. This is a way to attack where a lot of 'zombie' computers work together to send so many requests to a system that it can't respond anymore. Such attacks interfere with normal functioning and as a consequence the services providers may end up losing money and suffering from tarnished reputations.

For the contemporary DDoS menace, researchers have come up with solutions that can detect and prevent the attack. A most hopeful solution in this regard is the combination of Machine Learning (ML) methods with Intrusion Detection Systems (IDS). IDS is capable of detecting DDoS attacks by comparing them through the application of the ML algorithms with normal patterns that are characteristic of network traffic. In the last decade, IDS enhanced with ML evolved significantly even if just in the last years a distributed architecture is consolidating its position which is able to protect from individual attacks by dividing the task among multiple IDS.

This research employed the CICIDS2017 dataset which is standard for any cybersecurity research in developing and evaluating the DDoS detection models by feature enhancing. Data normalization has been performed as the initial stage to rank the data values for better comparability. Using both passive and active ML-based feature selection approaches, only the most selective traffic features were isolated. Passive feature selection is specially used for controlling incoming traffic, whereas the active feature selection mainly focuses on the identification of features in real time. Two testing sets were also developed for comparing the ML classification models of choice, as well as the best hyperparameter s for each model. In particular, Random Forest algorithm was examined by its scalability and by the ability to classify the DDoS attacks accurately.

Many classification models in the ML process were built and tested, and the hyperparameters were adjusted in accordance with the result. On the same, the Random Forest algorithm was tested based on its performance on big data and success rate towards DDoS detection.

The use of ML has several advantages such as high efficiency in recognizing DDoS attacks, perspectives to update the method if new kinds of attacks appear, and real-time work with the enormous amount of network data. When these systems are implemented within distributed architectures, they improve scalability and reliability to accommodate large scale deployment in the services environment. Passive and active feature selection also ensures that a lot of the data processing load is removed without a negative impact on the detection rate. Thus, this experiment identifies that the Random Forest algorithm model yields the highest detection accuracy with the mean detection accuracy of 97.5% for DDoS attacks. This result is essential to understand how ML techniques, specifically the Random Forest model, can accurately identify malicious traffic from the legitimate one. Such high accuracy proves that the applicability of ML-based DDoS detection systems can improve the security of application layer as a strong protection against future cyber threats.

Keywords: Botnet detection, DDOS, DDOS behavior, Machine learning algorithms, CICIDS2017

ACKNOWLEDGEMENT

I would like to convey my heartfelt gratitude to SLIIT for their invaluable guidance and consistent supervision throughout this journey. Their support, along with the provision of essential project information, has been crucial, and I deeply appreciate their continued assistance as I work towards completing the project.

Additionally, I am sincerely thankful to everyone who has offered their cooperation, encouragement, and support along the way. A special note of appreciation goes to my project supervisor, Mr. Amila Senarathna, whose expertise, time, and attention have been instrumental in guiding me.

Furthermore, I would like to convey my heartfelt thanks to my colleagues and all those who generously gave me their skills and volunteered their time to assist me. Your contributions have been vital to the progress and success of this project

TABLE OF CONTENTS

DECLARATION	II
ABSTRACT.....	III
ACKNOWLEDGEMENT	V
TABLE OF CONTENTS.....	VI
List of Figures	IX
List of Tables	X
Chapter 1 Introduction.....	1
1.1 Introduction.....	1
1.2 Motivation.....	3
1.3 Problem Statement.....	3
1.4 Contribution	5
1.5 Goal and Objectives	5
1.5.1 Goals.....	5
1.5.2 Objectives	6
Chapter 2 Background and Literature Review	7
2.1 Background.....	7
2.1.1 Information Security	7
2.1.2 Information Security Process.....	7
2.1.2.1 Prevention.....	8
2.1.2.2 Detection	9
2.1.2.3 Response.....	9
2.2 Type of Attacks in CICIDS2017 Dataset	9
2.2.1 DoS HULK.....	10
2.2.2 DoS GoldenEye.....	10
2.2.3 DoS GoldenEye.....	11
2.2.4 DoS Slowloris	11
2.2.5 Botnet.....	12
2.2.6 DDoS.....	13
2.3 DDoS Attack Classification.....	14
2.3.1 Volume-Base DDoS	15
2.3.1.1 UDP Flood	15
2.3.1.2 ICMP Flood.....	15
2.3.1.3 ICMP Flood.....	16
2.3.2 Protocol-Based DDoS Attacks	16
2.3.2.1 SYN Flood	16

2.3.2.2	Fragmented Packet	16
2.3.2.3	Ping of Death.....	17
2.3.2.4	Smurf Attack	17
2.3.3	Application Layer DDoS Attacks	17
2.3.3.1	HTTP Flood	17
2.3.3.2	Slowloris.....	17
2.3.3.3	Zero-day DDoS	18
2.3.4	Mitigation Of DDoS Attacks	18
2.3.4.1	Signature-Base Detection	18
2.3.4.2	Anomaly-Based Detection	18
2.4	Machine Learning (ML)	19
2.4.1	Supervised Learning	19
2.4.2	Unsupervised Learning.....	20
2.4.3	Semi-Supervised Learning	20
2.4.5	Reinforcement Learning.....	20
2.4.6	Machine Learning Algorithms.....	21
2.4.6.1	Naïve Bayes.....	21
2.4.6.2	Decision Trees	23
2.4.6.3	Random Forest.....	25
2.4.6.4	Support Vector Machines (SVM).....	28
2.4.6.5	MLP (Multi-Layer Perceptron)	30
2.4.6.6	K Nearest Neighbour (KNN)	33
2.4.7	Feature Selection Techniques Using Machine Learning.....	36
2.4.7.1	Feature Selection from Filter Method	36
2.4.7.2	Feature Selection from Wrapper Method	37
2.4.7.3	Feature Selection from Embedded Method	38
2.5	Related Works	39
Chapter 3	Methodology	44
3.1	Model	44
3.2	Dataset	45
3.2.1	K Dataset composition	45
3.3	Implementation	46
3.3.1	Data Processing.....	46
3.3.2	Data Cleansing	46
3.3.3	Features and Labels.....	47
3.3.4	Attack's Count.....	48

3.3.5 Impact on the Application	48
3.3.6 Limitations	48
3.3.7 Train and Test Data	48
3.3.8 Selection of Features.....	49
3.3.8.1 DDoS Attack Vs Features	49
3.4 Tools and Methods	51
3.4.1 Software Platform	51
3.5 ML Algorithms Implementation	52
3.6 Performance Evaluation Metrix.....	53
3.6.1 Accuracy	53
3.6.2 Precision, Recall, and F1-Score	54
3.6.3 Cross Validation	55
3.6.4 Confusion Matrix.....	55
Chapter 4 Results and Discussion	56
4.1 DDoS Top 10 Features performance.....	56
4.2 DDoS Feature Selection According to Machine Learning Algorithms.....	57
4.3 Experiment Dataset Result with Important Feature.....	57
4.4 Evaluation and Discussion	59
Chapter 5 Conclusion and Future Works.....	61
5.1 Conclusion	61
5.2 Future Works	61
Chapter 6 References.....	63
Chapter 7 Appendix.....	68
Appendix 1: CICIDS2017 Dataset Features and Explanations	68

List of Figures

Figure 1.The Growing Number of Internet Users Over the Years (1990-2024)	1
Figure 2.Botnet architecture	13
Figure 3. DoS and DDoS structure.	14
Figure 4. Type of DDoS attacks	15
Figure 5. Decision tree structure	23
Figure 6. Diagram explains the working of the Random Forest algorithm	26
Figure 7. Support Vector Machines (SVM) classification	29
Figure 8. Basic structure of the MLP	31
Figure 9. KNN Algorithm working visualization.....	33
Figure 10 Filter model.....	37
Figure 11 Wrapper Model.....	37
Figure 12 Embedded methods	38
Figure 13. Methodology model.....	44
Figure 14 DDoS attacks top 10 features and importance	50
Figure 15. A confusion matrix	55
Figure 16 DDoS Top 10 Feature Performance	56

List of Tables

Table 1.Features in CICIDS2017 dataset	47
Table 2. CICIDS2017 DDOS attacks type's.....	48
Table 3. DDoS attack Top 10 feature and importance	50
Table 4. DDoS Top 10 Important Feature list and Percentage.....	52
Table 5. Selected feature according to the threshold value.....	53
Table 6 DDoS Top 10 Features Performance	56
Table 7 Important Feature Performance	57
Table 8 Features according to Machine Learning Algorithm	57
Table 9. Result of the dataset	58
Table 10 Comparison of the performance of two studies	59

Abbreviations

DoS	: Denial of Service
DDoS	: Distributed Denial of service
TCP	: Transmission Control Protocol
ICMP	: Internet Control Message Protocol
UDP	: User Datagram Protocol
HTTP	: Hypertext Transfer Protocol
IDS	: Intrusion detection System
ML	: Machine Learning
MLP	: Multi-Layer Perceptron
DT	: Decision Tree
RF	: Random Forest
NB	: Naïve Bayes
SVM	: Support Vector Machines
KNN	: K-Nearest Neighbors
FN	: False negatives
FP	: False positive
TP	: True positives
TN	: True negatives