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is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources).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

List of Figures

List of Tables

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