iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VIII **Month of publication:** August 2022

DOI: https://doi.org/10.22214/ijraset.2022.46506

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VIII Aug 2022- Available at www.ijraset.com

Comparison of Supervised Learning Algorithms for DDOS Attack Detection

Rithik Sachdev[1], Shreya Mishra[2], Shekhar Sharma[3]
[1] *Nextuple India Pvt Ltd, India,* [2] *Nextuple India Pvt Ltd, India,* [3]*Computer Science Department, Boston University, Massachusetts*

Abstract: In today's world, when ubiquitous computing has become quite prevalent, there has been an upsurge in the number of users on the internet. The Distributed Denial of Service attack is the most widespread attack that disrupts the functioning of websites, servers, and services. In such attacks, the resources are exhausted by overwhelming requests from multiple attackers and thus become unavailable to users. Hence, it is essential to detect these attacks and prevent network security breaches. This work presents a supervised learning-based DDoS detection comparison developed using the CIC-IDS 2017 dataset [7]. Various models have been compared on different performance metrics to analyze efficiency in detecting DDoS attacks.
Keywords: DDoS attack, machine learning, supervised learning, detection, mitigation.

I. INTRODUCTION

With the advent of the internet in the past decades, there has been a significant increase in connected devices through the internet due to the internet of things (IoT) paradigm, thus giving malicious attackers control of our devices. A distributed denial of service attack is an attempt made by an attacker to disrupt the normal flow of traffic over a network or server and overwhelm it with a flood of illegitimate traffic. All the devices over the internet interact with services and applications over the network, thus getting control of these devices will ease the task of bringing down a network, for an attacker. As more and more devices connect to the internet, more launching points for such attacks are created for the attacker. A DDoS attack originating from a botnet, which is a network of numerous compromised devices that send data over the network or the server at the same time, makes the server highly unresponsive.

The DDoS attacks have caused a lot of havoc to some esteemed organizations such as Amazon Web Services (February 2020), Github (February 2018), and BBC (December 2015). These attacks can cause huge financial losses, sudden outages, loss of important data and applications, and service inaccessibility. This makes the detection and mitigation of such attacks important in today's world.

The detection of DDoS attacks before their onset will turn out to be a boon for any organization, helping them to take the correct steps for its mitigation and eradication. DDoS attacks are easy to implement since they do not require significant knowledge from the attacker's end, and a variety of applications and platforms are present for its orchestration.

This study presents a comparison of different supervised learning algorithms on the CIC-IDS 2017 dataset [7] by creating a feature set using five statistical methods, applying the Smart Detection feature selection algorithm [5]
then comparing the different supervised learning algorithms.

II. RELATED WORKS

Intrusion detection in networks is a widely discussed and researched topic among computer scientists. Before this research, there have been significant developments in this field.

H.H. Jazi *et.al.*[1] proposed a technique that uses data sampling to detect HyperText Transfer Protocol (HTTP) based attacks targeted toward servers. The anomaly-based strategy uses the proposed algorithm (CUMSUM), to classify the traffic as DOS attack or benign.

S. Behal *et.al.*[2] proposed D-FACE, which is a system that detects different types of Distributed Denial of Service attacks and Flash events (events when there is a sudden increase in legitimate traffic) by making use of Generalized Information Distance (GID) and Generalized Entropy (GE). Flash events can occur when a lot of clients try to access a resource simultaneously.

C.Wang *et.al.*[3] proposed The SkyShield system, which tries to detect DDoS attacks and mitigate the same at the application layer. For protecting against these attacks, methods like filtering, black-listing, and CAPTCHA are used.

Z.Liu *et.al.*[4] proposed 'Umbrella', which can protect from a large number of Distributed DoS attacks by multilayered defense architecture.

TABLE I
COMPARISON OF RELATED WORKS

Authors	Advantage	Disadvantage
H.H.Jazi et.al.[1] (CUMSUM Algorithm)	Detects HTTP based DDoS attacks.	High Sampling Rate, therefore, cannot be used in automatic mitigation systems.
S. Behal et.al.[2] (DFACE)	Detects Flash Events and DDoS attacks separately.	Incompatible with the latest DDoS attacks.
C.Wang et.al.[3] (SkyShield)	Detects as well as mitigates application level attacks.	Vulnerable to transport and network layer attacks.
Z.Liu et.al.[4] (Umbrella)	Incorporates multilayered defense architecture.	Fails in case of truly massive attacks.

III. PROPOSED WORK

A. *Feature Extraction*

Supervised learning classification requires parameters for training the classifier model. Each sample of the dataset has variables associated with labels or classes that will help to identify the DDoS attacks from the normal traffic. The TCP/IP architecture model consists of several header variables, thus making it important to extract the most important ones. This would further help in increasing the efficiency of the model, and also optimizing computational resources.

The network traffic data from CIC-IDS 2017[7] was obtained, and the raw data consisting of the packet entries were selected. It consists of 84 attributes including, destination and source IP addresses, destination and source ports, packet lengths, protocol, TCP flags, etc. The most commonly used eleven variables were selected for creating the customized dataset. The Source and Destination IP addresses are not necessary for identifying the behavior of the network packet, reducing the number of variables to nine, out of which one is the TCP flag which consists of FIN flag, SYN flag, RST flag, PSH flag, ACK flag, URG flag, CEW flag, and ECE flag, making a total of 16 variables. With these 16 variables, dataset variables were created by applying five statistical methods to a group of 10 packets, thus making a total of 80 variables in the customized dataset.

The statistical methods used are mean, variance, standard deviation, entropy, interquartile range, and median-absolute-deviation for better feature selection results and precision scores.

- Entropy:
$$\text{Entropy}(X) = -\sum_i p(X_i)\log_2 p(X_i)$$
- Mean:
$$\text{Mean} = \sum \frac{X_i}{N}$$
- Standard Deviation:
$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$
- IQR (Interquartile Range):
$$\text{IQR} = Q_3 - Q_1, Q_1 = \left(\frac{n+1}{4}\right)^{\text{th}}\text{term}, Q_3 = \left(\frac{3(n+1)}{4}\right)^{\text{th}}\text{term}$$
- Median Absolute Deviation:
$$\text{MAD} = \text{median}(|X_i - \mu|)$$

B. Oversampling and Undersampling:

The data obtained was balanced using oversampling and undersampling techniques. The sampling techniques help to balance data and provide efficient distribution of minority and majority class instances. For the proposed model, the oversampling technique used is the synthetic minority oversampling technique (SMOTE) [8], and the under-sampling technique used is random under sampling [9].

C. Feature Selection

The balanced data thus obtained from oversampling, and undersampling techniques is fed to the Smart Detection Random Forest Classifier [5], to achieve the attributes of maximum importance required to classify the network traffic.

The precision threshold is set to 0.95, the threshold by class as 0.85, and the threshold of importance as 0.09. The algorithm is executed 1000 times. The precision was 99.95%, the accuracy was 99.94%, and out of 80 variables, 39 of maximum importance were selected for supervised learning algorithms. Finally, supervised machine learning algorithms were applied to train a model. The selected variables are shown in Table II.

TABLE II
SELECTED VARIABLES

S.No	Variable	Details
1.	psh_flg_mad	MAD of psh flag
2.	ack_flg_entropy	Entropy of ack flag
3.	urg_flg_entropy	Entropy of urg flag
4.	t_fwd_pkt_std	Standard dev. of total fwd. packets
5.	t_fwd_pkt_iqr	IQR of total fwd. packets
6.	t_fwd_pkt_entropy	Entropy of total fwd. packets
7.	psh_flg_std	Standard dev. of psh flag
8.	min_pkt_len_std	Standard dev. of min. packet length
9.	urg_flg_mean	Mean of urg flag
10.	urg_flg_std	Standard dev. of urg flag
11.	ack_flg_mean	Mean of ack flag
12.	b_pkt_per_sec_entropy	Entropy of backward packets/sec
13.	b_pkt_per_sec_mad	MAD of backward packets/sec
14.	b_pkt_per_sec_std	Standard dev. of backward

		packets/sec
15.	dport_mad	MAD of dport
16.	dport_iqr	IQR of dport
17.	dport_entropy	Entropy of dport
18.	max_pkt_len_mad	MAD of max packet length
19.	b_pkt_per_sec_iqr	IQR of backward packets/sec
20.	ip_proto_mean	Mean of IP Protocol
21.	fduration_mad	MAD of flow duration
22.	min_pkt_len_mean	Mean of min packet length
23.	psh_flg_entropy	Entropy of psh flag
24.	dport_std	Standard dev. of dport
25.	psh_flg_mean	Mean of psh flag
26.	sport_mad	MAD of sport
27.	t_fwd_pkt_mean	Mean of total forward packets
28.	b_pkt_per_sec_mean	Mean of backward packets/sec
29.	sport_iqr	IQR of sport
30.	max_pkt_len_iqr	IQR of max packet length
31.	dport_mean	Mean of dport

32.	fduration_iqr	IQR of flow duration
33.	sport_std	Standard dev. of sport
34.	fduration_std	Standard dev. of flow duration
35.	sport_entropy	Entropy of sport
36.	max_pkt_len_entropy	Entropy of max packet length
37.	fduration_mean	Mean of flow duration
38.	max_pkt_len_mean	Mean of max packet length
39.	max_pkt_len_std	Standard deviation of max packet length

D. MLA Selection

The customized dataset was split into two parts comprising 70% training and 30% testing data. Different classifiers, namely, Random Forest, Decision Tree, Adaboost, and Voting Classifier were used, and their results were compared using various evaluation metrics.

The Voting classifier is a combination of a Random Forest classifier, Decision Tree classifier, and Adaboost Classifier. It aggregates the findings of each of its resident classifiers. The accuracy of the voting classifier was evaluated by assigning different weights to these classifiers.

As shown in Table III, having the weights as 2:1:2, respectively, gave the best results.

TABLE III
COMPARISON OF DIFFERENT WEIGHTS IN VOTING CLASSIFIER
(RANDOM FOREST: DECISION TREE: ADABOOST)

Ratio	F1 Score	Accuracy
1:1:1	0.99	0.989415
1:2:2	1.0	0.991126
2:2:1	1.0	0.989517
2:1:2	1.0	0.999528

The results from all of these classifiers are then compared with each other. The final results are shown in Table 4.

IV. RESULTS

A. *Dataset Description*

The CIC-IDS 2017 dataset by ISCX [7] contains normal traffic and common attack packets. This recent dataset includes packets relating to new families of attacks and is also available publicly. The data was captured on July 5, 2017, under controlled conditions and focusing on certain computers. The dataset consists of benign packets as well as packets relating to Portscan, DDoS, Botnet, and Web attacks.

B. *Results and Evaluation Metrics*

In the proposed research, F-measure(F1), Recall (REC), and precision (PREC) metrics are used to evaluate the performance and efficiency of the DDoS Attack detection model. F1 score is the harmonic mean of PREC and REC. Its highest value can be 1[6].
The voting classifier with weights of 2:1:2 (Random Forest: Decision Tree: Adaboost) applied to the data generated by statistical methods gave the best results than other classifiers and similar studies proposed, as shown in Table IV and Table V. However, the proposed method cannot perform live sniffing. The addition of different attack classes can enhance the robustness of the system.

Table IV
Comparison of different methods

Technique Used	Comparison Metrics		
	PREC	No. of Features	F1 Score
Voting Classifier	0.999528	39	1.0
AdaBoost	0.999210	39	1.0
Random Forest	0.998692	39	1.0
DTC	0.992563	39	1.0

TABLE V
Comparison with research approaches of related works

Work	Dataset	PREC
H.H.Jazi et. al.[1]	CIC-DoS	NA
M.Aamir and S.M.A. Zaidi [7]	CICIDS2017	0.8210
F. S. de Lima Filho et. al.[5]	CIC-DoS	0.9990
F. S. de Lima Filho et. al.[5]	CICIDS2017	0.9992
Voting Classifier in the proposed approach	CICIDS2017	0.9995

V. CONCLUSION

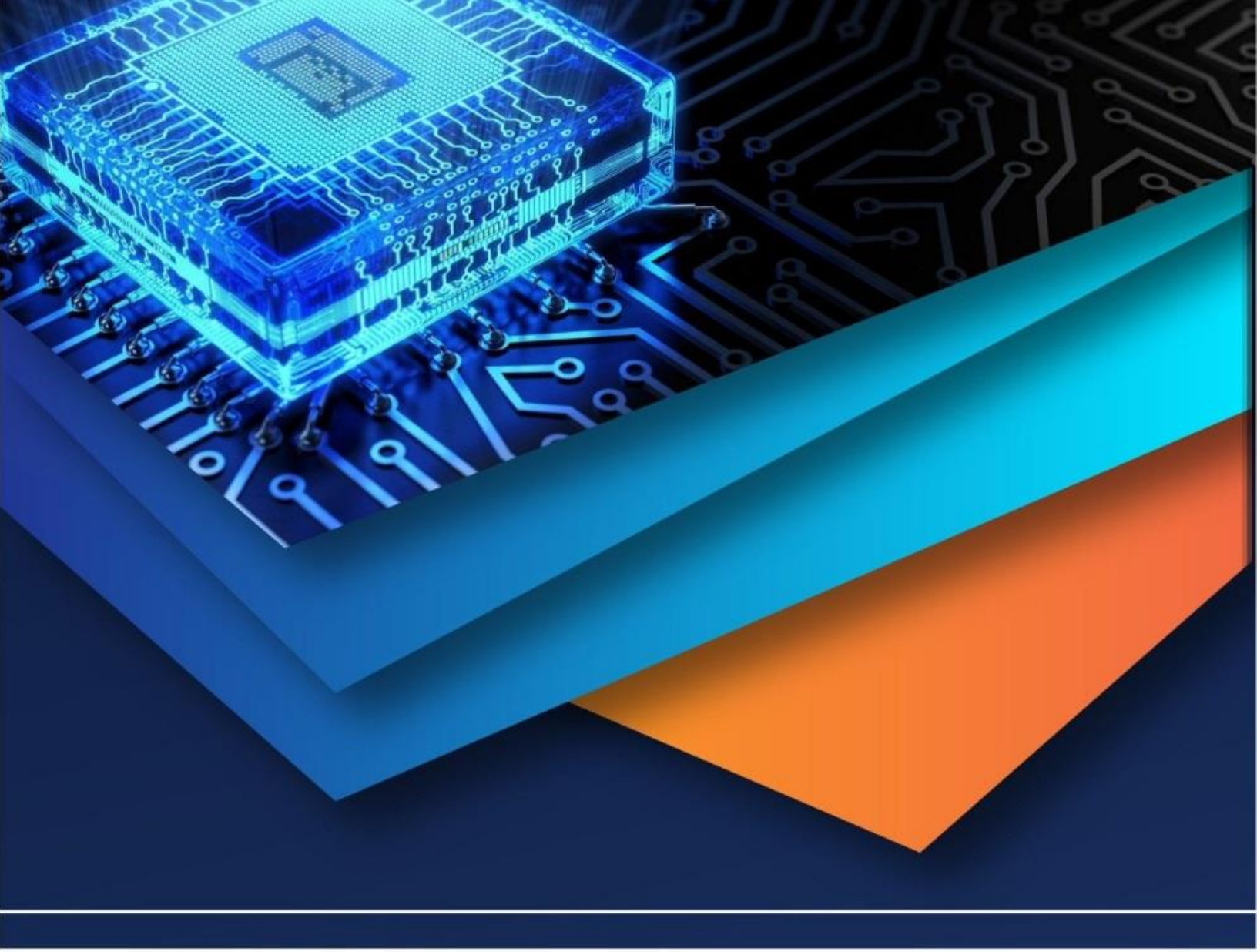
The work presents a comparison between different supervised learning algorithms to detect DDoS attacks. A customized dataset was created by applying various statistical methods to the benchmark CICIDS 2017 dataset. Different classifiers, namely, Random Forest, Decision Tree, Adaboost, and Voting Classifier, were compared based on performance metrics. The performance metrics included precision, recall, and F1 score. Based on the experiments conducted, the voting classifier with weights of 2:1:2 (Random Forest: Decision Tree: Adaboost) gave the best results than other classifiers. The results of this work can improve the detection of these malicious DDOS attacks using the proposed classifier.

VI. ACKNOWLEDGMENT

We would like to thank the authors of the reviewed papers for their insightful contributions.
The codebase for the entire research, analysis and interpretation is available on https://github.com/rithiksachdev/DDoS-Attack-Detection

REFERENCES

[1] H.H. Jazi, H. Gonzalez, N. Stakhanova, A. A. Ghorbani (2017) Detecting HTTP-based application layer DoS attacks on web servers in the presence of sampling. Computer Networks, 121:25–36, https://doi.org/10.1016/j.comnet.2017.03.018
[2] S. Behal, K. Kumar, M. Sachdeva (2018) D-face: an anomaly-based distributed approach for early detection of DDoS attacks and flash events. Journal of Network and computer applications, 111: 49–63, https://doi.org/10.1016/j.jnca.2018.03.024
[3] C. Wang, T. T. N. Miu, X. Luo, J. Wang (2018) SkyShield: a sketch-based defense system against application-layer DDoS attacks, IEEE Transactions on Information Forensics and security, 13(3):559–573, https://doi.org/10.1109/TIFS.2017.2758754
[4] Z. Liu, Y. Cao, M. Zhu, W. Ge (2019) Umbrella: enabling ISPs to offer readily deployable and privacy-preserving DDoS prevention services, IEEE Transactions on Information Forensics and Security, 14(4):1098–1108, https://doi.org/10.1109/TIFS.2018.2870828
[5] Francisco Sales de Lima Filho, Frederico A. F. Silveira, Agostinho de Medeiros Brito Junior, Genoveva Vargas-Solar, Luiz F. Silveira (2019) Smart Detection: An Online Approach for DoS/DDoS Attack Detection Using Machine Learning, Security and Communication Networks, vol. 2019, Article ID 1574749, https://doi.org/10.1155/2019/1574749
[6] Abigail Koay, Aaron Chen, Ian Welch, Winston K. G. Seah (2018) A new multi classifier system using entropy-based features in DDoS attack detection, International Conference on Information Networking (ICOIN), https://doi.org/10.1109/ICOIN.2018.8343104
[7] Iman Sharafaldin, Arash Habibi Lashkari, Ali A. Ghorbani (2018) Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization, 4th International Conference on Information Systems Security and Privacy (ICISSP), Portugal, January 2018 https://doi.org/10.5220/0006639801080116
[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer,"SMOTE: Synthetic Minority Over-sampling Technique", arXiv:1106.1813, https://doi.org/10.1613/jair.953
[9] Guillaume Lemaître, Fernando Nogueira, Christos K. Aridas (2017) Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning, Journal of Machine Learning Research 18(17):1−5, https://jmlr.org/papers/v18/16-365



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY