

An Empirical Assessment of the Ensemble Framework for Feature Selection in DDoS Attacks

Sneha Tantuway¹, Prof. Jayesh Jain², Prof. Saurabh Sharma³, Prof. Saurabh Kapoor⁴

^{1,2,3,4}Global Nature Care Sangathan Group of Institutions, Jabalpur(M.P), India

Abstract— In recent years, DDoS attack detection has frequently used entropy-based characteristics. However, current methods primarily distinguish between attack traffic and regular traffic using entropy-based features. The difficulty with this approach is limiting detectable DDoS attacks. In order to improve the overall and accurate detection of low and high-intensity DDoS attacks, we have proposed new entropy-based functions to aid in the detection of attacks. We have also introduced a new multi-classification system that is based on the proposed set of multiple entropy-based features and classified devices. Using different intensity data sets, we measured and performed analysis. The investigations' findings demonstrate that our approach was more precise and effective than other cutting-edge methods. Our method consistently yields four distinct data sets that include the best or second-best classification for various attack intensities.

Keywords — prediction models, data models, logistics, machine learning, mathematical model, techniques for machine learning and analysis, DDoS, entropy. **Introduction**

I. INTRODUCTION

A coordinated DDoS attack from multiple sources is part of an assault[1]. DDoS attack sources might number in the hundreds or even thousands, and they are frequently distributed over the globe. DDoS attacks can be detected using conventional traffic measurement techniques like MULTOPS[2] and LADS[3]. If the traffic volume exceeds a predetermined threshold, this method is used to identify a DDoS attack.

High intensity attacks and moderate intensity attacks are frequently used to categorise DDoS attacks. The rate of the trans-packet mission is the primary distinction between these attacks. Low-intensity, application-level attacks like Slowloris[5] transmit low-speed packets, but high-intensity assaults like Smurf[4] transmit high-speed packets, which substantially increase traffic volume. Other low intensity assaults aim to degrade services rather than entirely stop them in order to cause economic harm over a longer period of time[6] or low intensities since the attack's distributed nature prevents all of the packets involved from being observed.

In recent years, DDoS researchers have frequently used entropy as a strategy to get around the restrictive application of a volumetric approach[7]. Information uncertainty is measured by entropy. It offers a view of how network communication is dispersed and can reveal patterns of a network attack[8]. The beginning of a DDoS assault may be indicated by a sudden change in entropy levels. However, only a few specific subtypes of entropy-based attack detection are used by the majority of modern[9]-[11] detectors. Entropical characteristics, such as the originating IP address, are the entropy of traffic features at predetermined intervals. According to various entropy-based criteria, a variety of DDoS attacks may limit the accuracy of detection. For entropy-based approaches, selecting the best detection threshold is a challenge[9]. For instance, traffic whose entropy value above the threshold value will be considered to contain attack traffic for a given period of time. To differentiate between DDoS attacks of different intensities, an appropriate detection threshold must be available. Finding a reasonable threshold, however, is challenging so that various DDoS attack types can be recognised and are less error prone

Utilizing machine learning classifiers is one method of solving this issue (ML). ML classifiers, which use past data to search for patterns and create classification models, are a more advanced and complicated method of categorising traffic[12]. To create a categorization model, these classifiers may automatically learn recurrent patterns in traffic and assault traffic. The learning process eliminates the requirement for manually setting the detection settings. The difficulty with the ML classifier, however, is that there isn't a single, all-encompassing classifier that can correctly categorise every form of attack flow. For instance, moderate-intensity and high-intensity DDoS attacks are fundamentally different types of attacks. Therefore, it is impossible to appropriately categorise both attacks using a single ML classifier.

To address concerns regarding the precision and generality of DDoS attacks, we have developed a new technique for DDoS detection called Machine Learning Classifier Determination (E3ML).

Our primary contribution is the creation of a multi-classifier system with majority voting, an arbiter, and new entropy-based characteristics, which combine common entropy-driven characteristics and new entropy-based characteristics to properly identify high and low-intensity DDoS attacks. We investigated and assessed E3ML using several intensity data sets. Our testing results unmistakably show that, when compared to other models, E3ML performs best and second best.

II. LITERATURE REVIEW

A. Classification of traffic

A classification of traffic can be used to determine between DDoS and typical network traffic attacks. Significant steps to classify traffic are packet titles including source and destination IP addresses, source and destination port numbers, protocol, time to live (TTL) and flag.

The goal of this research is to construct a model of classification that classifies traffic classes that use machinery-learning techniques based on entropy values.

B. Entropical

Entropy has been commonly used in contemporary DDoS assault detection processes[9],[13] and[14], often used to produce important traffic grading features. Entropy is a statistical approach that evaluates the uncertainty of knowledge. Entropy detects the distribution of network traffic modifications using a single value measurement[15]. Sufficient observation of these changes showed anomalies in the network[9],[13],[14].

Recent detection investigation shows higher accuracy of detection than existing entropy-based detection methods[16]. The benefits of this approach are: quick estimation, high susceptibility, low FPR, without traffic or devices and without the usage of the network[9], more than classic volumetric approaches[2][3].

Entropy measurements are typically employed for raw traffic functions including source and destination IP addresses, port and entropy-based protocol numbers. The high entropy value, for example, indicates a considerable fluctuation and the low entropy value means that traffic packets have a smaller origin variation. This is helpful for detecting attacks, since prominent DDoS attacks with many target attack sources are often very different from ordinary traffic at source and destination.

Methods of so-called entropy-based variants can also produce entropical functions.

[9] In these functions, you can measure the difference between two distinct entropy-based properties, such as the difference between the source and the target.

Address IP. IP address. IP address. IP address. IP address. Address IP. The source IP and destination IP address for legitimate traffic usually shows a similar entropy value[9], whereas the majority of DDoS assault traffic has substantial differences between the two entropy-based properties as multiple distributed sources of attack send assault traffic to a certain goal.

C. Multiple Classification System.

The Multiple Classification System (MCS) is the technique for developing a broader and more accurate classification system employing distinct ML classifiers and merging one or more models.

In the past, neural network combinations and decision tree classifiers have shown the ability to generalise and detect accuracy[17]. Robinson et al. [18] analysed and graded many machine classifications based on their success in detecting DDoS attacks. They have also shown that classification methods such as AdaBoost and Random Forest are better combined than classification approaches. Chand et al.[19] created MCS approaches using attack stacking methods and showed the maximum accuracy in Vector Machine Support (SVM) stacking with other ML classifiers, rather than using the SVM classification alone.

However, MCS approaches only focus on the detection of a specific sort of DDoS assault using standard entropy-based calculations. No further entropy functions were considered in any of the existing works. It is desirable to have a broad and reliable method of detection. We are presenting a system for identifying various forms of DDoS attacks with wide knowledge integrated in numerous entropy-based functions and a classification based on MCS.

III. ENTROPY AND MACHINE LEARNING BASED CLASSIFIER SYSTEM FOR DDOS DETECTION

We offer a new classification model that uses entropy based features and MCS approaches to detect various forms of DDoS attacks. Our approach, E3ML, combines the comprehensive knowledge of various entropy features and intensity to differentiate between attack and legitimate traffic in several classification models. We are using both old entropy-based characteristics and adding new entropy-based features to reveal various forms of network DDoS assault.

In three classifications of the machine, we apply various forces to increase generality and precision in the detection of various DDoS attack traffics utilising novel entropy features to classify models.

A. System overview

Our strategy includes building the features displayed in Fig. 1 and detecting attacks as indicated in Fig. 2.

B. Entropical Building Feature

We construct two types of entropy function; typical entropy-based functions that depend on raw entropy

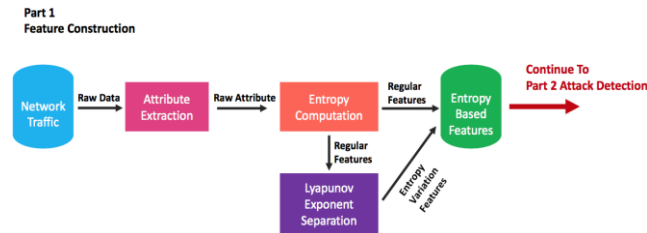


Fig. 1. Construction function

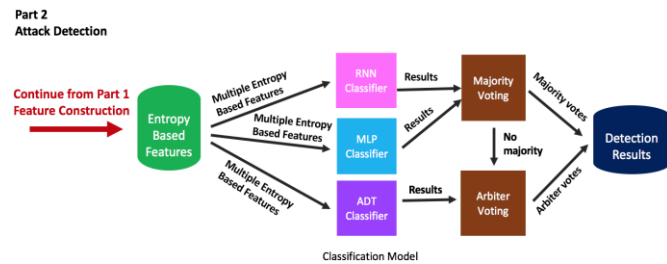


Fig. 2. Detection of Attack

Characteristics and entropy variance based on two conventional entropy characteristics. We have selected two separate features to increase the efficiency of traffic classification. While standard entropy (source and destination IP addresses, address and protocol, source and destination port) features commonly used in the classification of traffic, they are restricted to certain types of DDoS (common DDoS attacks) and cannot operate in other types of DDoS, such as low-intensity DDoS attacks. The variation in IP in Entropy showed promising improvements in continuous, low-intensity attacks[20]. We can boost the overall detection of various sorts of DDoS attacks by providing more useful traffic classification.

The design of the function takes place in three phases.

- *Extraction of raw features:* we extract raw features from every packet header. Features include source IP, target IP, source port address, destination port and protocol commonly used for DDoS detection[21].

We also examine the extraction of other functionality not usually used for entry-based function formatting, such as delta time, packet length, TCP sequence, and TCP window duration (cf: Table I.)

- *Entropy Calculation:* each entropy shall be computed by a specific time interval based on the Shannon entropy, with $W = 60$ seconds in all our past tests. The literature considers the time interval of 60 seconds as an adequate classification and standard practice[9]. The complex time interval change remains a projected target of the investigation. Shannon entropy is used since it is a standard traffic entropy measure. We have also tried different techniques to entropy, such Tsallis[22].

Table I
Regular Entropy-Based Features Generted From Raw Traffic Features

No	Features	Definition
1	Delta time *c*	Time since the previous packet was captured
2	Source IP Address *c*	Source IP address of the packet
3	Destination IP Address *c*	Destination IP address of the packet
4	Source Port Address *c*	Source port number of the packet
5	Destination Port Address *c*	Destination port address of the packet
6	Source MAC Address	Hardware address of the previous network router/host the packet is coming from
7	Destination MAC Address	Hardware address of the next-hop network router/host the packet is going to
8	Source Network Address	Source Network address of the packet
9	Destination Network Address	Destination Network address of the packet
10	Protocol *c*	Type of protocol (HTTP, TELNET, DNS)
11	Packet Length	Size of packets in bits
12	IP DSCP Value	Differentiated Services Code Point value of each packet
13	TCP Sequence Number	TCP sequence number relative to the first segment in a TCP session
14	TCP Window Size	Maximum amount of received data, in bytes, that can be buffered at one time on the receiver
15	TCP Length	Size of TCP Payload

c - commonly used features in traffic classification [21]

Table II
Entropy Variation Features Generated Using A Variation Of Lyapunov Exponent Separation

No	Features	Definition
1	Separation IP	Rate of separation between source and destination IP Address
2	Separation Port *new*	Rate of separation between source and destination Port Address
3	Separation MAC *new*	Rate of separation between source and destination MAC Address
4	Separation Network *new*	Rate of separation between of source and destination Network Address
5	Separation TCP *new*	Rate of separation between TCP window size and destination TCP Length

new - new entropy-based features

Reinyi [23] and produce results similar to the entropy of Shannon. Features that are calculated at this stage are referred to in Table I as normal entropy-based features.

- *Build entropy variation characteristics:* We create entropy variation features using the method given by Ma et al.[9] by combining two normal entropy-based features obtained by pre-violent stages. Ma et al. recommended to compute the rate of separation between the source and the destination IP address by the Lyapunov Exponent Separation method. They used the separation rate to highlight the disparities between two separate characteristics. This IP separation capability can successfully separate DDoS attacks from a single type of DDoS attack. We have introduced new entropy modification characteristics that can be used for detecting different kinds of DDoS attacks, such as Lyapunov separation technique, separation mac, separation network and separation TCP, as described in Table II.

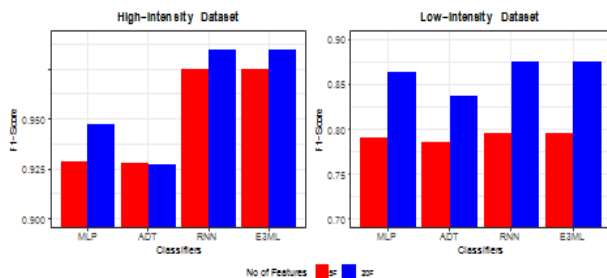


Fig. 3.

Fig. 3. Effects of conventional 5-fold (5F) vs 20 F1 scoring (20F) compared with traditional 5-fold data set. This applies to all four classifiers, with the exception of the high-intensity ADT classifier. However, the difference in F-measure performance in both sets of characteristics can be minimal as it is only 0.06 percent. In other classifications, 20 features have a 5-fold output of at least 1%. We therefore employ all 20 features as input for our Attack Detection classifier model.

A. Attack detection

Attack detection (Fig. 2) involves (1) the categorization of traffic through many ML classifiers (RNN, Multilayer Perceptron, ADT) and (2) the use of a simple majority vote method based on the classification results to determine the occurrence of DDoS. The vote takes place between two standard ML classifications (i.e. MLP and RNN).

Each classifier provides votes on the basis of its categorization result (attack traffic or normal traffic). If voting is equal for attack and normal, we propose a third classification, i.e. ADT for arbitrators. We only compare the ADT detection result to the RNN detection results because our preliminary test suggests that RNN is more accurate. We compare the outcomes of ADT and RNN detection with simple majority voting to decide the ultimate outcome by comparing RNN and MLP in the former phase. However, we identify traffic as an assault if ADT voting and RNN voting are different.

IV. PERFORMANCE EVALUATION

We validated our technique by using a relatively fresh ISCX 2012 (ISCX'12)[24] and 1998 DARPA Intrusion Detection (DARPA'98) data set [25]. ISCX'12 is a newer dataset containing more recent attacks, such as high intensity attacks generated with IRC botnets and low intensity attacks made by the Slowloris tool. ISCX'12 is a new dataset (Low-Intensity). The dataset DARPA'98 features more typical attacks, such as SMURF, Neptune and Land that have a mixing intensity due to many forms of attacks in a single dataset. In our assessment we use data from Week 1 (Mix Intensity 1) and Week 2 (Mix Intensity 2). The dataset is still used to evaluate DDoS detection and defence approaches[26] for comparison in the literature, however the DARPA'98 has various problems[27]. In the literature, these datasets were widely utilised to evaluate detection of DDoS and protection approaches[9], [28].

V. CONCLUSION & FUTURE WORK

In this study we proposed a DDoS detection approach using several entropy-based capabilities and E3ML classifiers. E3ML comprises of function builds which provide two types of entropy functions (i.e. conventional entropy-based and entropy) and an attack detector to recognise traffic in the network in an attack or normal using our detection algorithm. E3ML is a voting method which compares the results of MLP and RNN classification as arbitrators with ADT.

Results from performance evaluations demonstrated that our technology can efficiently detect DDoS attacks with varied intensities via data sets. Our approach surpasses other previous approaches (EMD-Li and ESDA) and has shown that data sets with various types of DDoS attacks can consistently produce high-precision results. We have shown that a combination of features and ML approaches on entropy produces promising results for the identification of DDoS.

REFERENCES

- [1] Neural network. https://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html.
- [2] Peer to peer lending and alternative investing. <https://www.lendingclub.com/>.
- [3] (2017). Variable selection with elastic net. <https://www.r-bloggers.com/variable-selection-with-elastic-net/>.
- [4] Addo, P. M., Guegan, D., and Hassani, B. (2018). Credit risk analysis using machine and deep learning models. *Risks*, 6(2):38.
- [5] Antonakis, A. and Sfakianakis, M. (2009). Assessing naive bayes as a method for screening credit applicants. *Journal of applied Statistics*, 36(5):537–545.
- [6] Attigeri, G. V., Pai, M., and Pai, R. M. (2017). Credit risk assessment using machine learning algorithms. *Advanced Science Letters*, 23(4):3649–3653.
- [7] Batista, G. E. and Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence*, 17(5-6):519–533.
- [8] Bekhet, H. A. and Eletter, S. F. K. (2014). Credit risk assessment model for jordanian commercial banks: neural scoring approach. *Review of Development Finance, Universiti Tenaga Nasional (UNITEN)*, 43000 Kajang, Selangor, Malaysia, 4(1):20–28.
- [9] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.
- [10] Chen, T., He, T., Benesty, M., et al. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, pages 1–4.
- [11] DAVIS, R. H., Edelman, D., and Gammernan, A. (1992). Machine-learning algorithms for credit-card applications. *IMA Journal of Management Mathematics*, 4(1):43–51.
- [12] Duan, K.-B. and Keerthi, S. S. (2005). Which is the best multiclass svm method? an empirical study. In *International workshop on multiple classifier systems*, Nanyang Technological University, Nanyang Avenue, Singapore, pages 278–285. Springer.
- [13] Faggella, D. (2017). The rise of neural networks and deep learning in our everyday lives - a conversation with yoshua bengio -.
- [14] Faggella, D. (2018). What is machine learning? - an informed definition. <https://www.techemergence.com/what-is-machine-learning/>.
- [15] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, Stanford University, USA, pages 1189–1232.
- [16] Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, Stanford University, Stanford, CA 94305, USA, 38(4):367–378.
- [17] Hamid, A. J. and Ahmed, T. M. (2016). Developing prediction model of loan risk in banks using data mining. *Machine Learning and Applications: An International Journal*, University Khartoum, Sudan, 3(1):1–9.
- [18] Hsu, C.-W., Chang, C.-C., Lin, C.-J., et al. (2003). A practical guide to support vector classification. National Taiwan University, Taipei 106, Taiwan.
- [19] Islam, S. (2017). Bad loans cripple the banking sector.
- [20] Khandani, A. E., Kim, A. J., and Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787.
- [21] Khashman, A. (2010). Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes. *Expert Systems with Applications*, Lefkosa, Mersin 10, Turkey, 37(9):6233–6239.
- [22] Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, Stanford University Stanford, CA., volume 14, pages 1137–1145. Montreal, Canada.
- [23] Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.
- [24] Lopes, R. G., Carvalho, R. N., Ladeira, M., and Carvalho, R. S. (2016). Predicting recovery of credit operations on a brazilian bank. In *Machine Learning and Applications (ICMLA)*, 2016 15th IEEE International Conference on, pages 780–784. IEEE.
- [25] Mackiewicz, A. and Ratajczak, W. (1993). Principal components analysis (pca). *Computers and Geosciences*, Department of Mathematics, Technical University of Poznan´, Piotrowo 3a, Poznan´ Poland, 19:303–342.
- [26] Mowla, G. (2018). Default loans plague banking sector.
- [27] Nova, A. (2018). More than 1 million people default on their student loans each year. [28] of India, P. T. (2018). 9 million loan defaulters blacklisted in china; 27 billion dollar frozen.

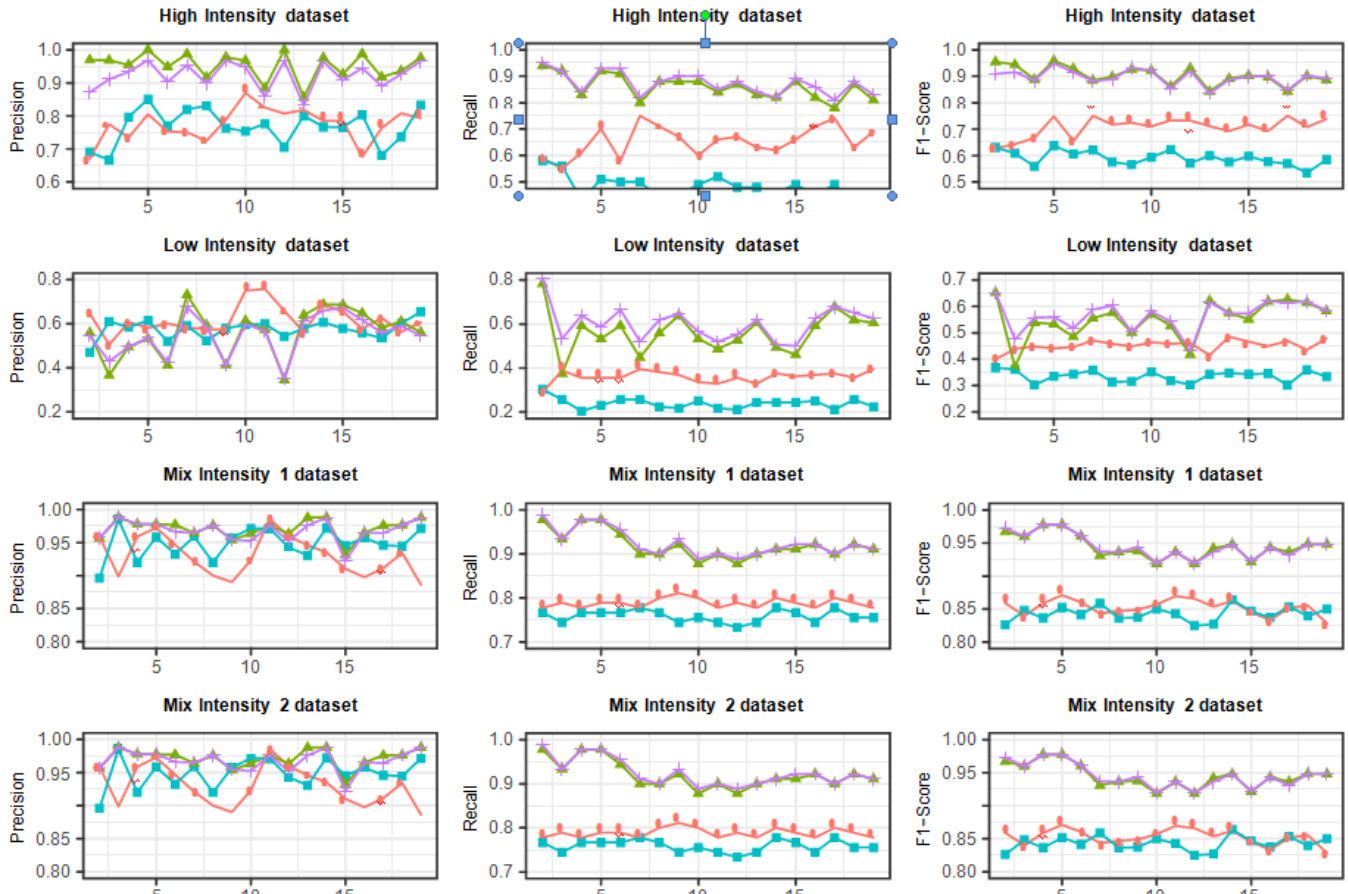


Fig. 4. Performance Evaluation