Hybrid Feature Selection for Efficient Detection of DDoS Attacks in IoT

KHADIJEH A WEHBI, LIANG HONG, TULHA AL-SALAH

College of Engineering, Tennessee State University, Nashville, TN 37209, USA

kwehbi@my.tnstate.edu

ABSTRACT: The increasing Distributed Denial of Service (DDoS) attacks on the Internet of Things (IoT) is leading to the need for an efficient detection approach. Although much research has been conducted to detect DDoS attacks on traditional networks, such as machine learning (ML) based approaches that have improved accuracy and confidence, the limited bandwidth and computation resources in IoT networks restrict the application of ML, especially deep learning (DL) based solutions that require extensive input data. In order to appropriately address the security issues in the resources-constrained IoT network, this paper is aimed to reduce the input data dimensions by extracting a subset of the most relevant features from the original features and using this subset to detect DDoS attacks on IoT without degrading the detection performance. A cost-effective model is developed to clean and prepare raw data before dimensionality reduction. A hybrid feature selection that uses Mutual Information (MI), Analysis of Variance (ANOVA), Chi-Squared, L1-based feature selection, and Tree-based feature selection algorithms is designed to identify important data features and reduce the data inputs needed for detection. Simulation results show that detection accuracy is improved with the combination of features chosen by the proposed hybrid feature selection approach. The training time is much less than the combination of each individual feature selection method.

Keywords: Internet of Things (IoT), Deep Learning (DL), Distributed Denial of Service (DDoS) attack, attack detection, feature selection.

1 INTRODUCTION

The evolution of the Internet of Things (IoT) and its applications in the current decade are making many aspects of our lives easier and more convenient, driving more people to depend on this technology and adding more devices to their networks every day. It is predicted that by 2030, there will be 125 billion IoT devices connected to the internet. IoT consists of an interconnection of intelligent objects, i.e., devices, varying from small cameras, watches, and coffee machines to fridges and big cars, which communicate with each other without human interactions and make decisions to enhance human life. However, these IoT benefits can considerably risk security issues and privacy loss [1-3]. Some of the reasons that contribute to IoT security threats are that IoT devices have limited storage capacity, processing capabilities, and computing resources such as; low power, loss of connectivity, and lots of other limitations. In addition to that, the open nature of the wireless medium makes new devices that enter the network to be configured automatically, which means the attackers can gain access to the devices in the network relatively easily and leaves such networks vulnerable to lots of malicious attacks. Once an attack is effective, it is a threat to the safety of human life, where it can cause death and destruction [4]. One of the most common and dangerous attacks on IoT is distributed denial-of-service (DDoS) attacks. DDoS attacks are usually executed using malicious software (malware), like Mirai, which keeps scanning for devices connected on the internet and protected by factory default usernames and passwords, utilizes vulnerabilities of networked devices to turn them into remotely controlled 'Bots', and then use these bots to create a botnet which can be used for various cyber-attacks [5].

Defending IoT from DDoS attacks is not an easy but fundamental task. In recent years, solutions based on Machine Learning (ML) and Deep Learning (DL) have been investigated and proven to achieve high accuracy in detecting DDoS attacks. ML is a type of artificial intelligence that uses several learning algorithms to train the data without intensive programming algorithms and has been used extensively for classification tasks. DL is a machine learning technique. The DL models are trained using a large set of labeled data and neural network architectures containing many layers [6]. DL models are hefty in real-world deployments, especially for IoT with its limitations and resource constraints [7]. Thus, this work aims to determine only the required input data for detecting DDoS attacks in IoT using a hybrid feature selection approach. Feature Selection can significantly improve a learning algorithm's performance [8]. The proposed hybrid feature approach combines the features from multiple features selection methods. This means if any feature selection method skips or ignores essential features, the other feature selection method will complement by selecting that missed important feature. The hybrid feature selection approach can get the optimal features required for DDoS

detection from any provided dataset. Proper Data preparation and feature selection process is the main contribution of this study. Despite having lots of research on feature selection methods, there is no particular research made on the study which method is superior to other methods and if the use of multiple selection methods surpasses the use of individual selection methods. Our work proposes a hybrid feature selection approach and analyzes the used selection methods.

Our model uses five feature selection methods to scale down the number of features without degrading the system's performance. The main contributions of this research are a cost-effective model towards an efficient detection of DDoS attacks in IoT and its components which include the proper preparation process for any given dataset; data analyzing, normalization, and balancing to reduce data dimensions and reduce memory usage; and a hybrid feature selection approach to identify the most relevant features for detecting DDoS attacks in IoT. The remaining paper is structured as follows: related work in Section II, followed by methodology of the proposed approach in Section III, results in Section IV, and conclusion and future work in Section V.

2 RELATED RESEARCH

Real-time network traffic is tremendously huge, which makes it a critical challenge for any classifier to handle. Therefore, an appropriate feature selection method to select the proper set of features from the raw data is significantly crucial in improving the performance of the DDoS attack detection classifier. Here we are previewing some works that have been done to address the issue of substantial real-time network traffic using feature selection methods: In [9], authors utilize a feature reduction approach for DL-based DDoS detection using the Analysis of Variance (ANOVA), which identifies essential data features and reduces the data inputs needed for detection. In [10], the authors used feature selection methods applied to machine learning models for botnet detection in IoT networks. Their selection method is a combination of filter and wrapper methods. They explain that the filter-based feature selection methods provide a computationally light approach to select the most informative features. The combination of Fisher's score with wrapper methods provided higher accuracy rates in each classification model. In [11], the authors used wrapper-based subset evaluation with a combination of a random forest (RF) classifier to evaluate each of the features first selected by the filter method. The reduced feature selection on both DARPA 1999 and KDD99 datasets was tested using an RF algorithm with ten-fold cross-validation in a supervised environment. Their result shows that the hybrid feature selections outcome was up to expectation.

Although various research has been proposed for feature selection based on a single performance metric such as relevance, accuracy, or redundancy, it might be insufficient as the approaches can be misleading for determining the best features to detect DDoS attacks. Incorrect selection of features may even lead to failed detection of DDoS attacks [12]. Moreover, since datasets are different with various characteristics, a single feature selection method used in existing works may select the optimal features in one dataset but may not be the proper method to choose the optimal features in another dataset.

3 METHODOLOGY

When securing the IoT network and its devices, the reduction of the features is highly needed for the detection of DDoS attacks due to the computation and memory resource constraints in IoT systems. Choosing the feature selection method often requires expert knowledge as it is not easy to determine a good set of features. To address this problem, a hybrid feature selection approach is developed to select the most critical features of any given dataset by using multiple feature selection methods to determine the optimal features. The combination of features selection methods can discover the powers of feature classification and identify the optimal features for DDoS attacks detection while still providing high accuracy rates. In addition to reducing the dimensions of the data, the proposed approach is a cost-effective approach that diminishes the needed memory and CPU time, which will help overcome computation and memory resource constraints while still enhancing the security of IoT networks and their devices.

3.1 DATASET

The dataset used for this work is CICDDoS 2019 [13]. It was provided by the Canadian Institute for Cybersecurity of the UNB, Canada. This dataset contains real network traffic from several days. Datasets have two classes, benign and malicious. The dataset is available for the public in the form of traffic traces in pcap format, plus CSV files containing packet payloads, the labels, and statistical details for each traffic flow.

CICDDoS 2019 [13] is the most recently released dataset that includes different modern reflective DDoS attacks such as LDAP, MSSQL, NetBIOS, PortMap, UDP, SYN, UDP-Lag. This paper uses the name DDoS attacks for all DDoS attack types captured in the traffic.

3.2 PROPOSED WORK

The high-dimensional data decreases the performance of deep learning models in a real-world deployment. Unneeded features not only load the run time of the CPU with extra processing time but also decrease the prediction accuracy of the trained models. When detecting DDoS attacks in IoT, it is necessary to reduce the load coming from data dimensionality via data preparation. The stages of the proposed model are as shown in flowchart Figure 1 and summarized as follows:

1. Raw dataset was downloaded from the Canadian Institute for Cybersecurity of the UNB, Canada website and used as the input for the proposed hybrid approach.
2. Data analysis and pre-processing, all empty spaces were removed, data types were checked and converted to the proper datatype to reduce memory usage, as shown in Figure 2.
3. Data normalization, which included removing the redundant features, and extracting all irrelevant and low variance features to prevent the classifier from being biased towards the majority class, features with infinity values were replaced with zeros.
4. Data balancing, after the data is clean and free from all redundant, irrelevant features, empty and infinity values, it's now unbalanced so that the ratio of data is significantly different, as shown in Figure 3. Through this step, all datasets were balanced, and the percentage of each label in all data was calculated.
   The intermediate class, which has the counts of records between majority and minority classes, was selected. The majority class records were down-sampled equal to middle counts class, and then the minority class counts were up-sampled similar to intermediate class counts. This way, over-fitting issues could be avoided later.

5. In feature selection methods, after data preparation, the data dimensions are reduced, memory storage is saved, and the dataset is ready as input to the hybrid feature selection approach. This approach aims to recognize the optimal features in the network traffic to use to detect DDoS attacks in IoT. Every selection method first finds the relevance between the individual features, then finds the relevancy between the target class and individual features. A scoring function is employed to assign each feature a scoring value (as shown in Tables 1-5) using statistical techniques (dependency between the variables, difference between the means of the groupings of the features, etc.). According to the scores, features are ranked in descending order. A subset of features with scores higher than 0.001 is selected by each method. The aggregated features are then the output of the hybrid feature selection method. It is clear that the proposed approach covers features that demonstrate significant differences between the normal data and the DDoS data in multiple aspects.

   The five different features selection methods employed for this purpose are:

   A. Mutual information (MI) [14]: MI between two random variables captures how much information entropy is obtained about one random variable by observing the other; it measures the information that X and Y share. MI is considered a more robust method as it is predicated on joint probability. It selects the relevant features and explores the degree of relationship between features. For example, if X and Y are independent, knowing X does not give any information about Y and vice versa. The MI between two random variables, X and Y, can be stated formally as follows:

   $$I(X;Y) = H(X) - H(X\,|\mathrm{Y})$$

   Where I (X; Y) is the mutual information for X and Y, H(X) is the entropy for X, and H (X | Y) is the conditional entropy for X given Y. The result has the units of bits (zero to one). The measure is symmetrical, meaning that I (X; Y) = I (Y; X). Moreover, mutual information can also detect non-linear dependencies among variables.
   B. Analysis of Variance (ANOVA) [15]: This function selects the top most relevant features (most significant values), identifies important data features, and reduces the data inputs needed for detection. ANOVA

measures the significance of numerical features depending on the difference between the means of the groupings of these features according to the target class label.

C. Chi-Squared [16, 17]: This method measures dependence between random variables and "weeds out" class-independent features. Chi-Squared is employed to reduce the range of features not relevant in the classification process. This selection method uses the theory of statistics to test the independence of a feature. Below is the equation of chi-square.

$$x^2(t_k, c_j) = \frac{N.(A.D - B.C)^2}{(A+B).(C+D).(A+C).(B+D)}$$

D. Tree-based feature selection [18]: This method takes the value of each feature that is considered necessary. The basic concept of this method is to apply logic like the decision tree algorithm to calculate the impurity using Gini Index

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

where C denotes total classes, and Pi is the probability of picking a data-point with I class.

E. L1-based feature selection [19]: This approach applies penalizing terms to the least square errors in linear regression and assigns zero coefficient to the irrelevant features to discard them from the model. It considers only non-zero coefficient variables to minimize prediction error by tackling overfitting and simplifying the model computationally and complexity stability. L1-regularization is a method that helps in reducing the complexity of the model by adding a penalty term as shown below to the least square errors E(w)

$$p = \alpha \sum_{i=1}^{d} |w|$$

where p is the magnitude of the penalty, α is the control parameter, d is the dimension of the features, and w is the weight of each feature.

$$E(w) = \sum_{j}^{n} (y_j - \sum_{i=0}^{d} w.x_i)^2$$

In this paper, feature selection methods, MI, ANOVA and Chi-Square, L1 based, and Tree-based are implemented in Python library sklearn.

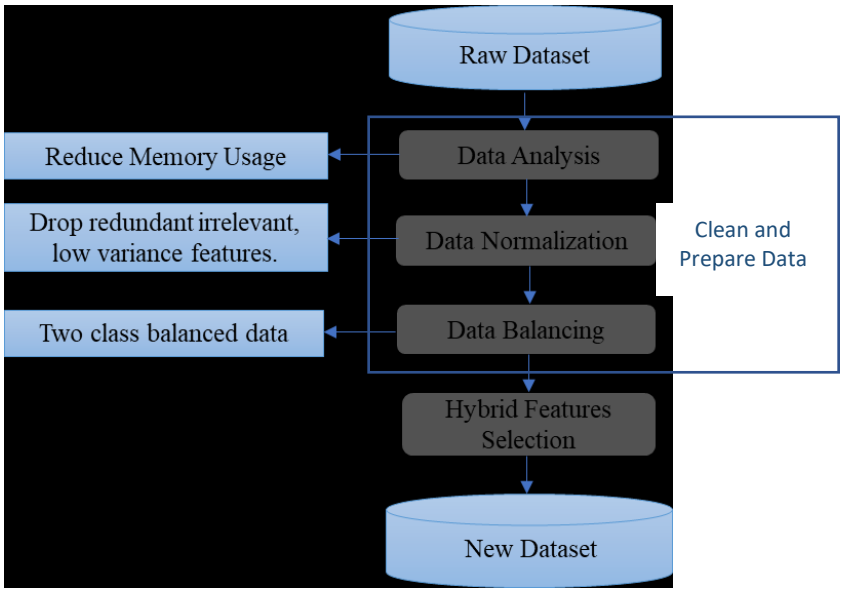

Figure 1: Proposed Model workflow.

After going through all the stages of the proposed model, a new dataset containing only the needed features is identified as the combination of all the features selected by every selection method.

4 RESULTS AND ANALYSIS

The new dataset obtained through the proposed workflow was fed as input to a convolutional neural network (CNN) to detect DDoS attacks. The accuracy was calculated as performance metrics. To compare the performance of detection, the dataset with features selected by each individual selection method is fed to the same CNN model.

Figure 2 shows the comparison of memory usage before and after data analyzing and pre-processing. Almost 50% of memory was saved by using our cost-effective model.

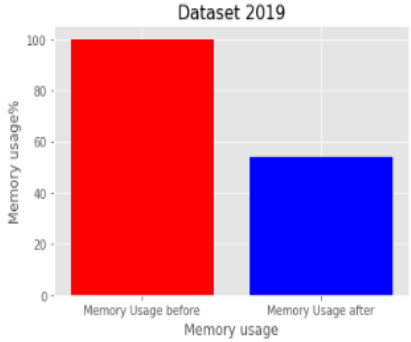

Figure 2: Memory usage comparison

Figure 3 shows the count of data before and after data balancing. It is clear that the unbalanced ratio of data is fixed through data balancing.

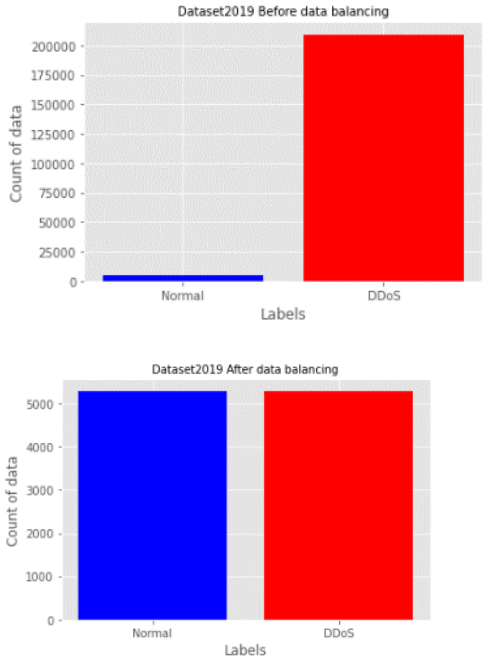

Figure 3: 2019 dataset before and after balancing.

Figure 4 and 5 show the 38 features selected from the original 88 features in the raw dataset. Therefore, the proposed hybrid feature selection approach reduces more than 50% of the data dimensions and identify the most relevant features from the given dataset.

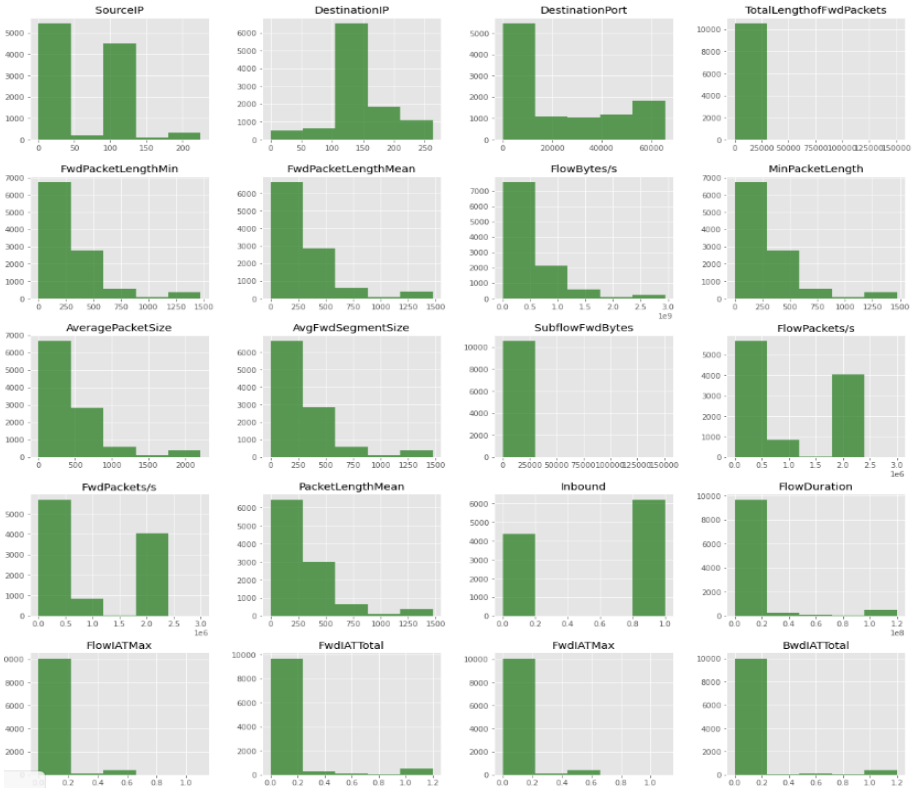

Figure 4: Features selected from the 2019 Dataset and their count in the dataset.

Figure 5: Features selected from the 2019 Dataset and their count in the dataset.

Tables 1- 5 show the selected features by each individual method and their scores. The repeated features are crossed out. Total 38 aggregated features are then fed to the deep learning algorithm to detect DDoS attacks.

Table 1: Selected Features by Mutual Information based method and their scores

	Feature	Feature Scores
0	SourceIP	0.690100
1	DestinationIP	0.688767
2	DestinationPort	0.610635
3	TotalLengthofFwdPackets	0.568095
4	FwdPacketLengthMin	0.516750
5	FwdPacketLengthMean	0.498509
6	FlowBytes/s	0.474439
7	MinPacketLength	0.331974
8	AveragePacketSize	0.309989
9	AvgFwdSegmentSize	0.163954

Table 2: Selected Features by ANOVA based method and their scores

	Feature	Feature Scores
0	~~SourceIP~~	1.000000
1	~~DestinationIP~~	0.578426
2	~~DestinationPort~~	0.112172
3	~~TotalLengthofFwdPackets~~	0.111986
4	~~FwdPacketLengthMin~~	0.085490
5	~~FwdPacketLengthMean~~	0.082934
6	~~FlowBytes/s~~	0.064355
7	~~MinPacketLength~~	0.064355
8	~~AveragePacketSize~~	0.035762
9	~~AvgFwdSegmentSize~~	0.014782

Table 3: Selected Features by Chi-Squared based method and their scores

	Feature	Feature Scores
0	FlowDuration	1.000000
1	~~FlowBytes/s~~	0.012123
2	FlowIATMax	0.011868
3	FwdIATTotal	0.011698
4	FwdIATMax	0.003940
5	BwdIATTotal	0.003707
6	BwdIATMax	0.003700
7	FwdPackets/s	0.003682
8	IdleMean	0.003646
9	IdleMax	0.003619

Table 4: Selected Features by Tree-based method and their scores

	Feature	Feature Scores
0	SourceIP	1.000000
1	~~SourcePort~~	0.999363
2	DestinationIP	0.999045
3	~~TotalLengthofFwdPackets~~	0.998839
4	~~TotalLengthofBwdPackets~~	0.997752
5	FwdPacketLengthMin	0.996622
6	~~FwdHeaderLength~~	0.995931
7	BwdHeaderLength	0.995501
8	PacketLengthVariance	0.992063
9	FwdHeaderLength.1	0.989235
10	SubflowFwdBytes	0.777649
11	SubflowBwdBytes	0.727901
12	Init_Win_bytes_backward	0.633708

Table 6 compares the accuracies of the proposed approach and that of each individual feature selection method. It is clear that the proposed approach outperforms every other feature selection method. Mutual Information and ANOVA selection methods accuracy results are very close, as they both selected ten features with which six of these features were the same from both methods. The chi-Squared accuracy result was also close to the previous two methods. One of the biggest strengths of Chi-Squared is that it is easier to compute than some other feature selection methods. The tree-based feature selection method achieved excellent performance, and was able to classify data with non-linear relationships. Therefore, tree-based predictions tend to be weak, as singular decision tree models, and are vulnerable to overfitting. The tree-based feature selection method is also not very stable, as a minimal change in the input dataset can significantly impact the final results. L1-based scored the lowest in terms of detection accuracy.

Table 5: Selected Features by L1-based method and their scores

	Feature	Feature Scores
0	SourceIP	1.000000
1	SourcePort	0.500379
2	DestinationIP	0.499776
3	DestinationPort	0.424349
4	Protocol	0.363766
5	FwdPacketLengthMin	0.302140
6	FwdPacketLengthMean	0.283747
7	FlowBytes/s	0.276575
8	FlowPackets/s	0.247789
9	FwdPSHFlags	0.117847
10	FwdPackets/s	0.111024
11	MinPacketLength	0.095064
12	PacketLengthMean	0.041335
13	ACKFlagCount	0.024169
14	URGFlagCount	0.021131
15	CWEFlagCount	0.011535
16	Down/UpRatio	0.007477
17	AveragePacketSize	0.003437
18	AvgFwdSegmentSize	0.001536

Table 6: Selection Methods Accuracy

Feature Selection Method	Accuracy
Mutual Information (MI)	0.93037
ANOVA	0.93982
Chi-Squared	0.89729
L1-based	0.60397
Tree-based	0.94310
Hybrid	0.97930

Table 7 compare the training time of the proposed approach and that of each individual feature selection method. Since each individual feature selection method could be executed in parallel in the proposed hybrid feature selection scheme, the training time of the proposed hybrid approach is significantly less than the combination of the training time of all the individual feature selection methods.

Table 7: Training Time

Feature Selection Method	Training Time
Mutual Information (MI)	99.512
ANOVA	99.972
Chi-Squared	155.777
L1-based	129.643
Tree-based	137.212
Hybrid	244.197

Table 8 illustrates the confusion matrix that shows two possible predictive classes (normal or DDoS) for each of the two actual classes. In confusion matrix, true positive (TP) is a decision that correctly classifies a DDoS activity as being DDoS; true negative (TN) is a decision that correctly classifies a normal activity as being normal; false positive (FP) is a decision that wrongly classifies a normal activity as being DDoS, and false negative (FN) is a decision that wrongly classifies a DDoS activity as being normal. Therefore, TN and TP imply the correct decisions made by the detector. In contrast, FP and FN imply the wrong decisions that the detector makes. It is clear that the hybrid method surpasses all five methods with the highest accuracy. An interesting result is that although L1-based method had the lowest in terms of detection accuracy, it showed a good performance in the confusion matrix, where it didn't pick any DDoS attack traffic as normal; its false negative percentage was zero.

Table 8: Confusion Matrix

Feature Selection Method	TP	FP	FN	TN
Mutual Information (MI)	1.0	0.0	.20	.75
ANOVA	.99	.01	.13	.90
Chi-Squared	1.0	.07	.20	.80
L1-based	.30	.70	0.0	1.0
Tree-based	1	0.0	.24	.78
Hybrid	.96	.05	.01	.98

Overall, simulation results show that the proposed method can reduce the input data noticeably, reduce memory usage and CPU time by almost 50%, and identify the most relevant features for any given dataset. All research objectives were achieved successfully. Selected features were tested and evaluated to prove that the proposed model and the hybrid approach are effective, light, and provide higher accuracy to existing research.

5 CONCLUSION AND FUTURE WORK

Detecting DDoS attacks in IoT is an important issue that requires timely analysis. Although feature selection schemes with single method have been used in the detection of DDoS attacks previously, a hybrid feature selection approach is implemented in this paper to deal with the huge amount of data that is usually generated from network traffic, which contains many features. This hybrid feature selection algorithm helps select a subset of features used in a CNN based deep learning model to decrease data dimension without sacrificing prediction accuracy. The proposed hybrid feature selection approach is implemented with five different feature selection methods. Simulation results show that the hybrid method has better accuracy than that of each selection method individually. The selected features are relevant to the detection problem and help in proper prediction.
For future work, the hybrid approach will be tested on more datasets, comparison for all datasets using the proposed hybrid approach will be provided. Improvement on the CNN based deep learning model that implemented in this work will take a place as well.

ACKNOWLEDGMENT

This work is supported in part by the US National Science Foundation (NSF) under Grants HRD-1912313 and HRD-1912414.

REFERENCES

[1] Yang, Y., Wu, L., Yin, G., Li, L. and Zhao, H., 2017. A survey on security and privacy issues in Internet-of-Things. IEEE Internet of Things Journal, 4(5), pp.1250-1258.

[2] Sonar, K. and Upadhyay, H., 2014. A survey: DDOS attack on Internet of Things. International Journal of Engineering Research and Development, 10(11), pp.58-63.Alsboui, T., Qin, Y., Hill, R. and Al-Aqrabi, H., 2020. Enabling distributed intelligence for the Internet of Things with IOTA and mobile agents. Computing, 102(6), pp.1345-1363.

[3] Alsboui, T., Qin, Y., Hill, R. and Al-Aqrabi, H., 2020. Enabling distributed intelligence for the Internet of Things with IOTA and mobile agents. Computing, 102(6), pp.1345-1363.

[4] A DDoS Attack Detection and Mitigation With Software-Defined Internet of Things Framework
[5] Wehbi, K., Hong, L., Al-salah, T. and Bhutta, A.A., 2019, April. A survey on machine learning based detection on DDoS attacks for IoT systems. In 2019 SoutheastCon (pp. 1-6). IEEE.
[6] https://www.mathworks.com/products/matlab.html LIMITS: Lightweight machine learning for IoT systems with resource limitations
[7] LIMITS: Lightweight machine learning for IoT systems with resource limitations
[8] https://medium.com/@mehulved1503/feature-selection-and-feature-extraction-in-machine-learning-an-overview-57891c595e96
[9] Sanchez, O.R., Repetto, M., Carrega, A., Bolla, R. and Pajo, J.F., 2021, June. Feature selection evaluation towards a lightweight deep learning ddos detector. In ICC 2021-IEEE International Conference on Communications (pp. 1-6). IEEE.
[10] Guerra-Manzanares, A., Bahsi, H. and Nõmm, S., 2019, October. Hybrid feature selection models for machine learning based botnet detection in IoT networks. In 2019 International Conference on Cyberworlds (CW) (pp. 324-327). IEEE.
[11] Kamarudin, M.H., Maple, C. and Watson, T., 2019. Hybrid feature selection technique for intrusion detection system. International Journal of High Performance Computing and Networking, 13(2), pp.232-240.
[12] Roopak, M., Tian, G.Y. and Chambers, J., 2020. Multi-objective-based feature selection for DDoS attack detection in IoT networks. IET Networks, 9(3), pp.120-127.
[13] The Canadian Institute for Cybersecurity. (2019). Datasets. Accessed: Oct. 31, 2019. [Online]. Available: https://www.unb.ca/cic/datasets/index.html
[14] Mohammadi, S., Desai, V. and Karimipour, H., 2018, October. Multivariate mutual information-based feature selection for cyber intrusion detection. In 2018 IEEE electrical power and energy Conference (EPEC) (pp. 1-6). IEEE.
[15] M. Sheikhan et al., “Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method,” Neural Comput. Appl., vol. 23, no. 1, pp. 215–227, 2013
[16] Classification of Twitter Contents using Chi-Square and K-Nearest Neighbour Algorithm
[17] A. Wibowo Haryanto, E. Kholid Mawardi, and Muljono, “Influence of Word Normalization and Chi-Squared Feature Selection on Support Vector Machine (SVM) Text Classification,” Proc. - 2018 Int. Semin. Appl. Technol. Inf. Commun. Creat. Technol. Hum. Life, iSemantic 2018, pp. 229–233, 2018.
[18] Ratmana, D.O., Shidik, G.F., Fanani, A.Z. and Pramunendar, R.A., 2020, September. Evaluation of Feature Selections on Movie Reviews Sentiment. In 2020 International Seminar on Application for Technology of Information and Communication (iSemantic) (pp. 567-571). IEEE.
[19] Shekar, B.H. and Dagnew, G., 2020. L1-regulated feature selection and classification of microarray cancer data using deep learning. In Proceedings of 3rd international conference on computer vision and image processing (pp. 227-242). Springer, Singapore.