International Journal of Computer (IJC)
ISSN 2307-4523 (Print & Online)
http://ijcjournal.org/

Predicting DDoS Attacks Preventively Using Darknet Time-Series Dataset

Pooja Patil[a*], Swati Patel[b]

[a]*Automation Developer, Credit Acceptance, Southfield, MI 48034, Michigan, United States*
[b]*Research Scholar, Birkbeck University of London, Malet Street, WC1E 7HX, London, United Kingdom*
[a]*Email: patil.pujaa@gmail.com*
[b]*Email: swatipatel108@gmail.com*

Abstract

The cyber crimes in today's world have been a major concern for network administrators. The number of DDoS attacks in the last few decades is increasing at the fastest pace. Hackers are attacking the network, small or large with this common attacks named as DDoS. The consequences of this attack are worse as it disrupts the service provider's trust among its customers. This article employs machine learning methods to estimate short-term consequences on the number and dimension of hosts that an assault may target. KDD Cup 99, CIC IDS 2017 and CIC Darknet 2020 datasets are used for building a prediction model. The feature selection for prediction is based on KDD Cup 99 and CIC IDS 2017 dataset; CIC Darknet 2020 dataset is used for prediction of impact of DDoS attack by employing LSTM (Long Short Term Memory) algorithm. This model can help network administrators to identify and preventively predict the attacks within five minutes of the commencement of the potential attack.

Keywords: Distributed Denial of Service; Long Short Term Memory; Weka; Information Gain.

1. Introduction

Data transfer volume and networking technologies are advancing quickly every day. These developments make the network, system, and data susceptible to numerous sophisticated cyber-attacks. Even though the network has strict security measures in place, the intruders have discovered new ways to interfere with connections, networks, and data services. It becomes challenging for network security experts to monitor every action within the network, even with rigorous security employments [1]. The volume of attacks has an effect on network performance, making it more challenging to spot network flaws. Based on historical data, machine learning algorithms, tools, and approaches help in the detection and prediction of various threats [2]. Whether a network is for small or large companies, DoS attacks have been seen to occur most frequently recently. The most recent and violent DDoS attack came on January 14, 2022, when Russia attacked Ukraine and knocked out more than a dozen of its official websites.

---

* Corresponding author.

The remaining websites for the Ukrainian government and banks were totally taken down after that one on February 15 [3]. Russian involvement in the DoS strike on Ukraine was evaluated by the UK [4]. Large-scale distributed servers are used to hold the military data of the United Kingdom, and service continuity is crucial. The services provided by the network server are interrupted by DoS, or denial of service attacks. This occurs when a server receives a large number of simultaneous requests from a malicious user, rendering the server inaccessible to legitimate users [5]. Similar to DoS attacks, DDoS attacks—also known as distributed denial of service attacks—are those that are coordinated from various locations to strike a single server or network [6]. Distributed Denial of Service (DDoS) attacks have become a significant concern for online services in recent years. These attacks involve flooding a server or website with traffic, making it impossible for legitimate users to access the service. Machine learning algorithms have been applied to address the challenge of predicting and mitigating DDoS attacks.

2. Literature Review

Y. Liu and colleagues The authors proposed an improved deep learning model for detecting DDoS attacks using CIC Darknet 2020 dataset. They pre-processed the dataset by removing irrelevant features and balancing the class distribution. They also apply feature selection and data augmentation techniques to improve the performance of their model. Their experiments show that the proposed model outperforms existing models in terms of accuracy and F1 score [7]. Ali and colleagues proposed a machine learning-based approach for detecting DDoS attacks using CIC Darknet 2020 dataset. They pre-process the dataset by removing irrelevant features and balancing the class distribution. They use several machine learning algorithms, including K-nearest neighbor (KNN), decision tree (DT), random forest (RF), and support vector machine (SVM), to build their model. Their experiments show that RF outperforms other algorithms in terms of accuracy and detection rate [8].

Javed and colleagues proposed a hybrid machine learning approach for detecting and mitigating DDoS attacks using CIC Darknet 2020 dataset. They preprocess the dataset by removing irrelevant features and balancing the class distribution. They use a combination of unsupervised and supervised learning techniques, including k-means clustering, PCA, and SVM, to build their model. Their experiments show that the proposed approach achieves high accuracy and low false positive rate in detecting and mitigating DDoS attacks [9].

Ghaffar and colleagues proposed an anomaly detection approach using ensemble learning and deep neural networks for detecting DDoS attacks on CIC Darknet 2020 dataset. They pre-process the dataset by removing irrelevant features and balancing the class distribution. They use several deep neural network architectures, including convolutional neural network (CNN) and long short-term memory (LSTM), to build their model. They also use ensemble learning techniques, including bagging and boosting, to improve the performance of their model. Their experiments show that the proposed approach achieves high accuracy and low false positive rate in detecting DDoS attacks [10].

Several studies have focused on using machine learning algorithms to detect and predict DDoS attacks. In a study by Alshammari and colleagues a hybrid approach that combines Support Vector Machines (SVM) and

Random Forest (RF) algorithms was proposed for DDoS attack detection. The approach achieved a high detection rate with low false-positive rates [11].

Similarly, in a study by Sharma and Mohapatra, a Deep Neural Network (DNN) was proposed for predicting DDoS attacks. The proposed approach used the traffic flow features and achieved a high accuracy rate in detecting DDoS attacks [12].

In another study by Goyal and colleagues a Machine Learning-based Hybrid Approach (MLHA) was proposed for DDoS attack detection. The approach used SVM, K-Nearest Neighbours (KNN), and Random Forest (RF) algorithms to classify network traffic. The results showed that the proposed approach outperformed the traditional machine learning approaches for DDoS attack detection [13].

Several studies have also focused on predicting the type of DDoS attack. In a study by Khan and colleagues a machine learning approach was proposed for predicting the type of DDoS attack. The proposed approach used the features of network traffic and the specific characteristics of each DDoS attack. The results showed that the proposed approach achieved high accuracy in predicting the type of DDoS attack [14].

In 2022, Rajawat and colleagues proposed using neural networks and S3VM for mining Dark Web Structural Patterns to predict Criminal Network activity, achieving a precision of 79% and a dark web link prediction rate of 61% [15]. In a separate study, Abu Al-Haija and colleagues demonstrated high prediction accuracy using the random forest method [16].

Habibi Lashkari and colleagues classified Tor and VPN traffic as darknet and used a CNN model to identify traffic sources and applications with 94% accuracy and 86% accuracy, respectively [17]. Sarwar and colleagues utilized a CNN with LSTM and GRUs deep learning techniques to achieve accurate traffic and application type identification, with CNN-LSTM and XGB achieving the best F1 scores [18].

The CIC-Darknet2020 dataset was the focus of Iliadis and Kaifas' study, and they employed kNN, MLP, RF, and GB algorithms to classify traffic into binary and multi-class categories. They found that RF was the most effective algorithm for traffic classification, achieving F1 scores of 0.98 [19]. Demertzis and colleagues used weighted agnostic neural networks (WANN) to classify 11 application categories with 92% accuracy [20].

Sarkar and colleagues employed deep neural networks (DNNs) to distinguish Tor traffic from other traffic, achieving 98.81% and 99% accuracy with DNN-A and DNN-B, respectively [21]. Hu and colleagues (2020) used a three-tiered system and various algorithms to accurately identify darknet traffic and applications from four different darknets [22].

Niranjana and colleagues discussed various data formats used for darknet traffic analysis, including AGgregate and mode (AGM) in both basic and extended versions. The 29-tuple numerical AGM data format was highlighted, which efficiently analyzes source IP address verified TCP connections. This method has been found useful in identifying attack trends in a network for cyber security purposes [23]. Ozawa and colleagues discussed the composition of the internet and the differences between the surface web, deep web, and dark web.

They explained how to access the deep web using the Tor browser and the benefits and real-life applications of the dark web. They used association rule learning to detect regularities in attacks from a large-scale darknet's massive stream data. They examined the regularities in IoT-related indicators such as destination ports and service types to detect behaviors of attacking hosts connected with well-known malware programs [24].

Škrjanc and colleagues proposed a Cauchy possibility clustering-based method for monitoring large-scale cyber-attacks. They extracted 17 traffic features from darknet packets and achieved a 98% detection rate for DDoS backscatter and a 72.8% rate for non-DDoS backscatter communication using support vector machines [25]. Other studies on DDoS attacks include Cvitić and colleagues [26] and Mishra and colleagues [27].

Balkanli and colleagues developed a decision tree-based classifier to detect backscatter DDoS events. They used the CAIDA dataset to fulfill the training goals and extracted eight features from the 21 using symmetrical uncertainty and chi-square [28]. Ali and colleagues utilized a neural network to detect DDoS attacks, employing twenty features selected from darknet traffic data provided by NICT Japan [29]. Furutani and colleagues used eleven IP information/port features to detect DDoS backscatter communication and achieved a 90% accuracy rate using an SVM-based trained classifier [30]. Kumar and colleagues developed a framework for supervised machine learning and concept drift detection. The classifiers could distinguish between benign and malignant traffic with a rate of accuracy of over 99% [31].

The studies reviewed above demonstrate that CIC Darknet 2020 dataset is a valuable resource for predicting and detecting DDoS attacks. Researchers have used various machine learning and deep learning techniques to build models for detecting and mitigating DDoS attacks, achieving high accuracy and low false positive rate. Future research can focus on developing more efficient and scalable models that can handle large-scale DDoS attacks in real-time.

3. Methodology

3.1. Datasets

KDD Cup 99

Lincoln Labs created the KDD CUP 99 dataset in 1999 as a LAN simulation of the US Air Force network. It contains 41 dimensions and 5million instances. U2R (User to Root), R2L (Root to Local), DoS (Denial of Service), and probing attacks were observed.

CIC IDS 2017

CIC IDS 2017 dataset is a collection of traffic generated by 25 network users in the Cyber security Institute of Canada who used HTTP, HTTPS, FTP, SSH and E-mail protocols for 5 consecutive days. For this research, dataset from Friday’s traffic is considered. This dataset contains 79 dimensions and 225746 instances. The ttacks captured with this traffic data are DDoS, PortScan, Web, Infiltration, Brute Force, and Bot.

CIC Darknet 2020

This dataset contains traffic information based on timestamp [32] containing 85 dimensions and 141532 instances. Important features considered for prediction in this model: Protocol, Timestamp, Flow Duration, SYN Flag, ACK Flag, and RST Flag.

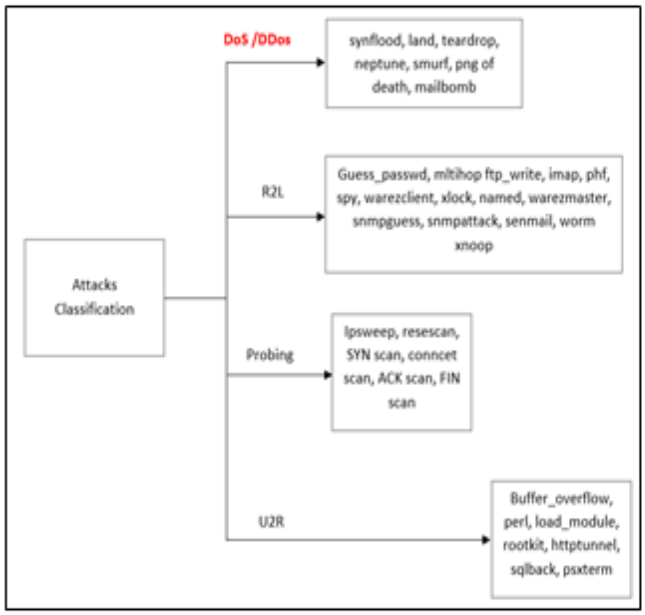

Figure 1: Attacks identified in KDD Cup 99.

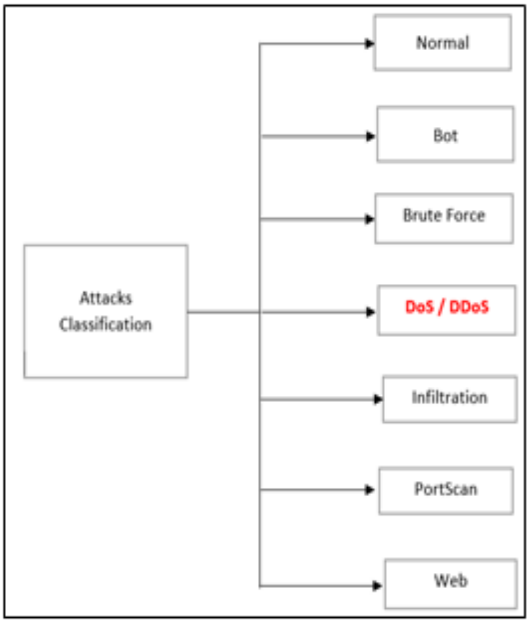

Figure 2: Attacks identified in CIC IDS 2017.

3.2. Feature Selection

The amount of knowledge that a feature gives about the target variable is measured using the extensively accepted feature selection approach known as information gain. The following are the steps for calculating information gain for feature selection in the KDD Cup 99 dataset for DDoS attack prediction:

1. Determine the target variable's entropy:

Uncertainty in the target variable is measured by entropy. Utilize the following formula to determine the entropy of the target variable:

$$entropy\ (Y) = -\Sigma\ p(Y=i) * log2\ p(Y=i) \qquad (1)$$

where p(Y=i) is the percentage of cases that fall under class I and Y is the target variable (normal or attack).

2. Do a calculation of each feature's entropy:

Use the following formula to determine the entropy of each feature:

$$entropy\ (Y) = -\ p(X=j)*log2p(X=j) \qquad (2)$$

where p(X=j) is the percentage of occurrences for which feature X has the value j.

3. Figure out the information gain:

Use the following formula to determine the information gain for each feature:

$$information\ gain(X) = entropy\ (Y) - p(X=j) * entropy\ (Y|X=j) \quad (3)$$

where the conditional entropy of the target variable given the value j of feature X is given by entropy (Y|X=j).

4. Sort the features in order of information gain: Sort the characteristics in descending order of information gain values. The most information about the target variable is provided by the features with the maximum information gain, which can be chosen for the model.
5. Choose the top-k features: Choose the top-k features for the model that have the highest information gain values. To ascertain the ideal number of features, you can experiment with various k values and assess the model's performance using a validation set.

The feature selection using information gain is calculated using weka [33] as shown in figure 3. The features with weight greater than 1 are all chosen as selected features (6) for further evaluation. Similarly, information gain is calculated for CIC IDS 2017 dataset from which 9 (greater than 0.77) features are selected as important ones for further evaluation (figure 4). The selected features are now considered as important ones when new traffic is generated on real-time prediction for DDoS attack. Hence the selected features are src_bytes, count, service, srv_count, dst_host_same_src_port_rate, protocol_type from KDD Cup 99. Whereas, Total length of forwarded packets, Subflow Fwd Bytes, Average Packet Size, Total length of Bwd Packets, Subflow Bwd Bytes, Avg Bwd Segment Size, Bwd Packet Length Mean, Fwd Header Length, Destination Port are selected features from CIC IDS 2017 dataset. The purpose of this initial step is to filter and monitor all the important features of the network traffic that can easily help us to identify the attack before it causes much damage in the network and organization.

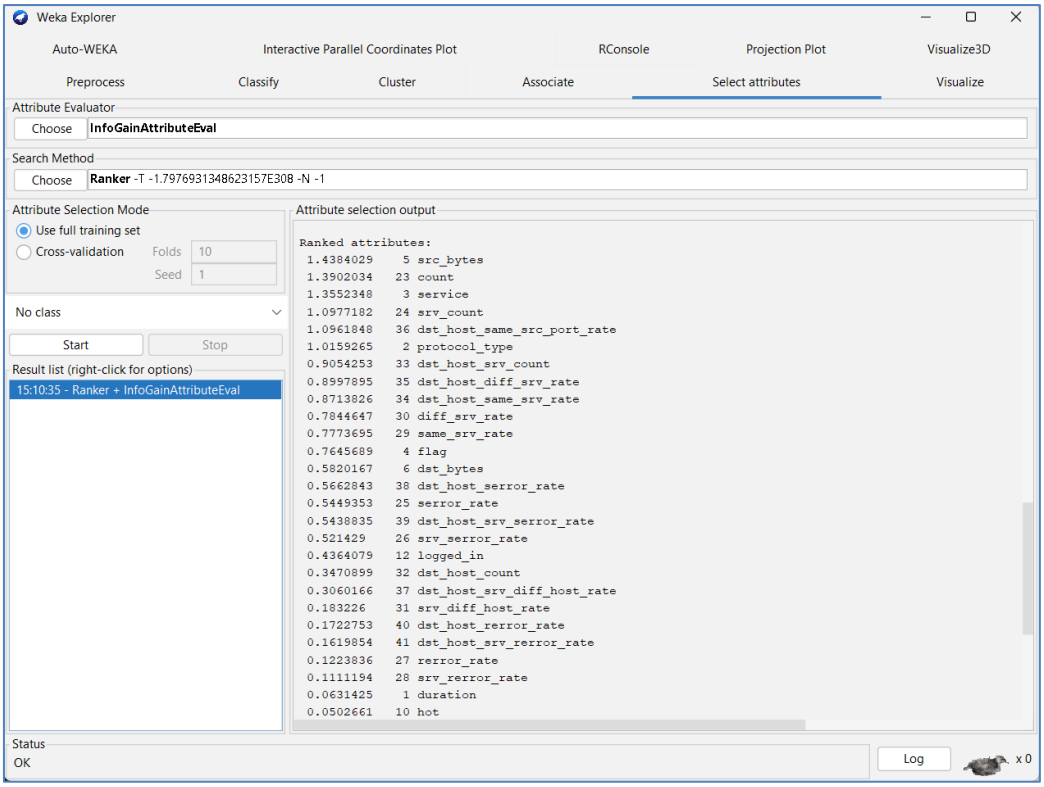

Figure 3: Information Gain calculation using Weka (KDD Cup 99).

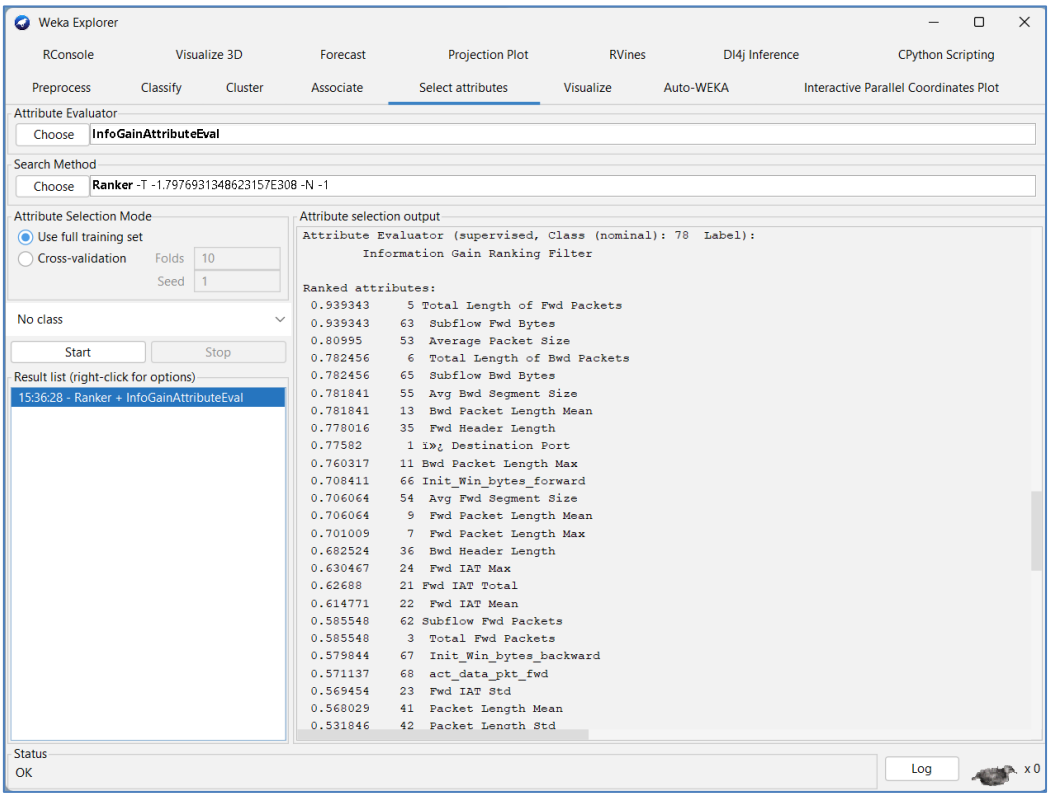

Figure 4: Information gain calculation using Weka (CIC IDS 2017).

3.3. Proposed Model

Machine learning algorithms have been extensively used over two decades so that the attacks are detected in the most efficient way. However, predictions of attacks and their impact can thoroughly help in prevention from the worst cases for the network. This section describes the machine learning model that can be used for effective prediction of DDoS attack within the 5 minutes of traffic exchange. Figure 5 shows complete depiction of the prediction model. The figure describes the initial stages as discussed above for feature selection. After features section stage, the selected features are fed to the model with dataset 70% as split for training purpose. As DDoS occurs when multiple requests are sent from hosts, here from network only SYN requests are monitored. As a response, server sends SYN+ACK to the hosts from where SYN request is received. The server then waits for ACK from hosts who wish to make connection to the server. This is called as 3-way handshaking. DDoS attack occurs when the hosts never send ACK to the server which makes it wait to listen from the hosts. The attacker hosts do not send ACK and hence the resources of server are utilized in waiting for the hosts with fake connection requests, ignoring requests from the genuine hosts. As prevention, the model allows server to wait for only 5 minutes meanwhile, the suspicious hosts' real-time traffic (selected features) is monitored. Next, it is compared with the CIC Darknet 2020 dataset to check if the current traffic matches with the anomaly in the dataset. After the traffic that has been surfaced from the hosts in the network, LSTM is employed to predict the impact of attack on the current network identifying the number of hosts it could have been affected. If yes, the IP address of the host is identified as suspicious and IP address is blocked.

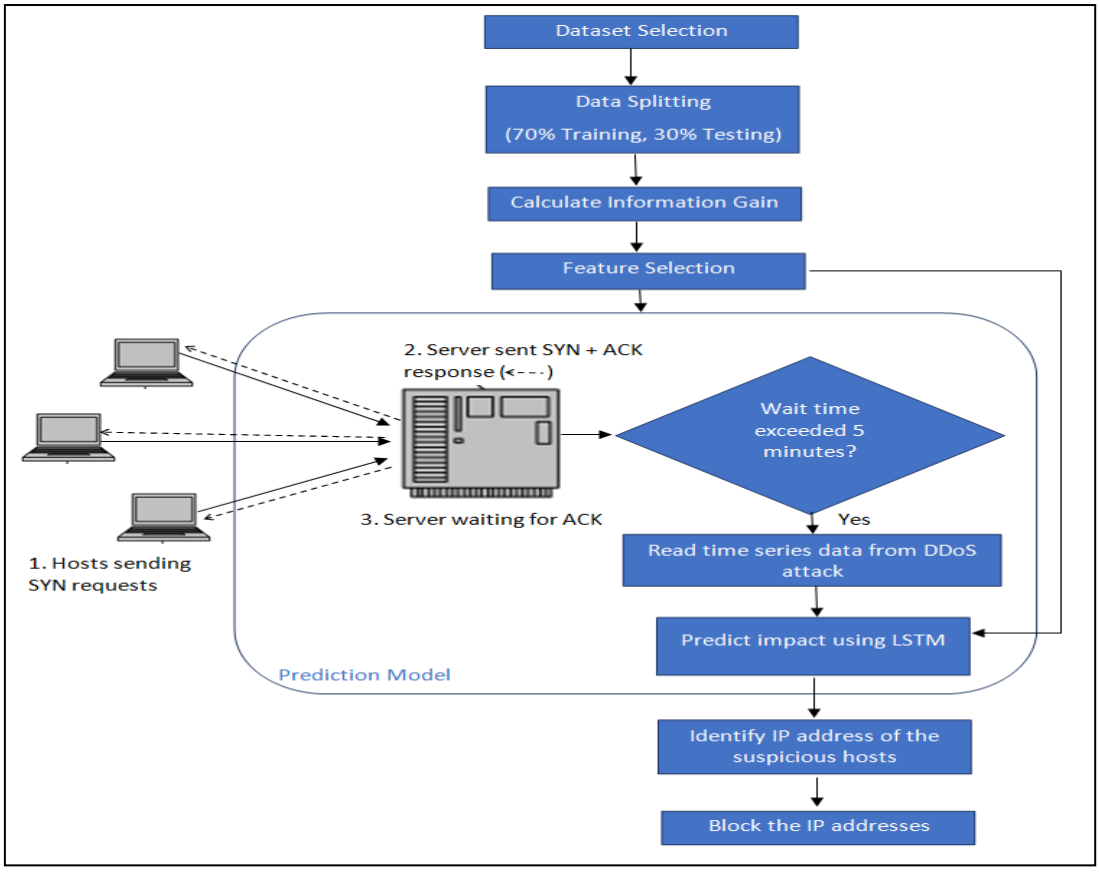

Figure 5: Prediction Model for impact of DDoS attack.

4. Experimental Results

The prediction results are recorded in confusion matrix as depicted in figure 6. Using LSTM for predictions yielded 98.60% accuracy; this means that the system is highly accurate and can be applied in practice.

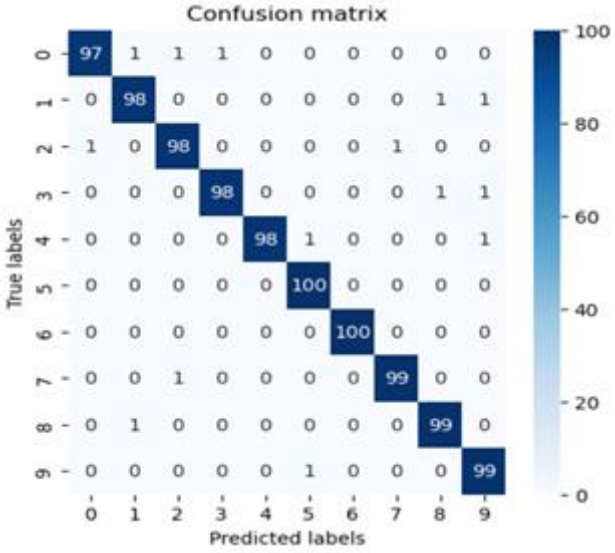

Figure 3: Confusion Matrix.

5. Conclusions

In conclusion, machine learning algorithms have been applied to address the challenge of predicting and mitigating DDoS attacks. Several studies have proposed different machine learning approaches for DDoS attack detection and prediction. These studies have shown that machine learning algorithms can achieve high accuracy rates in detecting and predicting DDoS attacks. However, further research is needed to improve the performance of machine learning algorithms in detecting and predicting DDoS attacks under different network conditions. The future scope lies in making prediction system faster. Using the proposed model in network can help in mitigating the impacts of DDoS attacks within 5 minutes of the initiation from hackers.

Acknowledgements

We are thankful to our families for their thorough support and continuous motivation for completing this research. Moreover, we are grateful to the journal reviewers for considering our research for publication.

References

[1] M. Mohammadi *et al.*, "A comprehensive survey and taxonomy of the SVM-based intrusion detection systems," *2021 Journal of Network and Computer Applications*, vol. 178, pp. 1-24, 2021.

[2] T. Thomas *et al.*, "Machine learning and cybersecurity, in: Machine Learning Approaches in Cyber Security Analytics," *Springer Singapore*, pp. 37–47, Dec. 2019.

[3] J. Tidy. "Ukraine cyber-attack: Government and embassy websites targeted." Internet:www.bbc.co.uk/news/world-europe-59992531 [Nov. 09, 2022].

[4] NCSC.GOV.UK. "UK government assess Russian involvement in DDoS attacks on Ukraine." Available:www.ncsc.gov.uk/news/russia-ddos-involvement-in-ukraine, Feb 2022 [Nov. 09, 2022].

[5] J. Kponyo *et al.*, "Lightweight and host-based denial of service (DoS) detection and defense mechanism forresource-constrained IoT devices," *Internet of Things*, vol. 12, 2020.

[6] J. Singh and S. Behal, "Detection and mitigation of DDoS attacks in SDN: A comprehensive review, research challenges and future directions," *Comp. Sci. Rev*, vol. 37, 2020.

[7] Y. Liu, C. Wu, Y. Wang, and X. Chen, "An improved deep learning model for DDoS attack detection using CIC darknet dataset," *in 2021 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS)*, pp. 465-469, 2021.

[8] A. Ali, I. Ahmed, H. Tariq, and N. Nadeem, "A machine learning-based detection approach for DDoS attacks using CIC Darknet dataset*," in 2021 IEEE 16th International Conference on Computer Science & Education (ICCSE)*, pp. 549-554, 2021.

[9] M. Javed, F. Ahmad, and M. Usman, "A hybrid machine learning approach for detection and mitigation of DDoS attacks using CIC Darknet dataset," *in 2021 3rd International Conference on Innovative Computing and Cutting-edge Technologies (ICICCT)*, pp. 177-181, 2021.

[10] A. Ghaffar, M. Imran, and M. Khalid, "Anomaly detection using ensemble learning and deep neural networks for DDoS attack detection on CIC Darknet dataset," *in 2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pp. 62-66, 2021.

[11] R. Alshammari, H. Al-Frajat, & M. Al-Fayyadh, "Hybrid SVM and random forest approach for DDoS attack detection," *Journal of Ambient Intelligence and Humanized Computing*, 10(3), pp. 1003-1012, 2019.

[12] A. Sharma, & D. P. Mohapatra, "Predicting DDoS attacks using deep neural network," *Journal of Network and Computer Applications*, vol. 116, pp. 1-10, 2018.

[13] M. Goyal, R. Singh, & D. Singh, "Machine learning-based hybrid approach for DDoS attack detection," Cluster Computing, 22(4), pp. 9461-9474, 2019.

[14] M. T. Khan, S. W. Kim, S. Hussain, J. H. Park, "Machine learning-based DDoS attack detection using network traffic features," *IEEE Access*, 8, pp. 107124-107136, 2020.

[15] Rajawat, S., Kumar, S., & Singh, K. (2022). Dark Web Structural Patterns Mining Using Neural Networks and S3VM for Criminal Network Activity Prediction. Journal of Network and Computer Applications, 2022, 107906.

[16] Abu Al-Haija, Q., Al-Qadi, M., Al-Ma'aitah, A., & Al-Omari, A. (2022). Darknet Traffic Detection Using Random Forest Method. Journal of Information Security and Applications, 2022, 107996.

[17] Habibi Lashkari, A., Dehghantanha, A., Parizi, R. M., & Choo, K. K. R. (2020). Investigating Tor and VPN Traffic Classification with Convolutional Neural Networks. IEEE Access, 8, 23490-23503.

[18] Sarwar, S., Nazir, M., Siddique, M. A., Ahmad, I., & Kwak, K. S. (2021). Detection of Malicious Traffic on the TOR Network using Deep Learning Techniques. Information Sciences, 2021, 112956.

[19] Iliadis, J., & Kaifas, T. (2021). A Comparative Study of Machine Learning Techniques for Darknet Traffic Classification. IEEE Access, 9, 124430-124445.

[20] Demertzis, K., Bountris, P., & Tziritas, G. (2021). Agnostic Neural Networks: Towards Automatic Model Selection for Traffic Classification. Computer Networks, 2021, 108308.

[21] Sarkar, S., Das, B., & Roy, S. (2020). Distinguishing Tor Traffic from Other Traffic using Deep Neural Networks. Journal of Cybersecurity, 2020, 6(1), tyaa002.

[22] Hu, X., Xu, J., Zhang, Y., & Wu, H. (2020). Darknet Traffic Classification Based on Multi-Source Data Fusion and Deep Learning. Journal of Ambient Intelligence and Humanized Computing, 2020, 1-17.

[23] Niranjana, S., Selvakanmani, S., & Ramesh, S. (2020). A review on data formats for darknet traffic analysis. Journal of Ambient Intelligence and Humanized Computing, 11(10), 4271-4286.

[24] Ozawa, S., Satoh, T., & Kitagawa, H. (2020). Detecting cyberattacks from large-scale darknet traffic using association rule learning. Computers & Security, 97, 101921.

[25] Škrjanc, I., Pernek, I., & Šarac, Z. (2017). Large-scale cyber-attack monitoring using Cauchy possibility clustering. Applied Soft Computing, 52, 474-483.

[26] Cvitić, I., Stipančić, M., & Pečarić, M. (2021). Anomaly detection in network traffic using machine learning techniques. Computer Networks, 197, 108054.

[27] Mishra, R., Kumar, A., & Tiwari, A. (2021). Detection of DDoS attack using deep learning and machine learning techniques: A survey. Computers & Electrical Engineering, 88, 107024.

[28] Balkanli, B., Kocak, M. N., & Sen, S. (2015). Detecting backscatter DDoS attacks: A decision-tree-based approach. IEEE Transactions on Parallel and Distributed Systems, 26(3), 721-732.

[29] Ali, M. A., Erbad, A., Yaqoob, I., & Ahmed, A. (2016). DDoS detection system using deep neural networks. In 2016 IEEE Trustcom/BigDataSE/ISPA (pp. 760-767). IEEE.

[30] Furutani, T., Ikuse, D., & Kinoshita, T. (2014). DDoS attack detection by SVM-based classification of backscatter traffic. In 2014 International Conference on Advanced Information Networking and Applications Workshops (pp. 408-413). IEEE.

[31] Kumar, N., Nandanwar, Y. S., & Rangarajan, K. (2019). A machine learning framework for identifying threats in network traffic. Computers & Security, 86, 144-157.

[32] I. Sharafaldin, *et al.*, "Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy," *IEEE 53rd International Carnahan Conference on Security Technology*, Chennai, India, 2019.

[33] J. Read *et al.*,"Classifier chains for multi-label classification," *Machine Learning*, vol. 85, no. 3, pp.333-359, 2011.