International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Detection of Distributed Denial of Service (DDOS) Attack Using Logistic Regression and K Nearest Neighbor Algorithms

G. Gnana Priya[1*], S. Harini Shriram[2] , S. Jeeva[3], G. Sakthi Priya[4], Dr. K. Balasubadra[5]

Submitted: 04/12/2023 **Revised**: 12/01/2024 **Accepted**: 27/01/2024

Abstract: SDN (Software Defined Network) devices are controlled in a centralized manner and it is better when compared to all other traditional networks. Some advantages of SDN such as greater scalability, high programmability, security features and management. In SDN, DDOs attack occurs certainly. Attacks such as DDOS (Distributed Denial of Service) pose a foremost risk in maintaining the security of the network and it also shut down the network fully. Traditional techniques do not work as well to identify the DDOS attack. Hence, in order to identify the DDOS attack, we employ certain machine learning algorithms. In our work, we compare two algorithms of Machine Learning (ML) such as Logistic Regression (LR) and K-Nearest Neighbors (KNN) and the accuracy is also compared. The accuracy of the two algorithms differs in our experimental results. The accuracy of Logistic Regression is roughly 91% and the accuracy of the KNN algorithm is roughly 99%. From the analysis KNN is better rather than Logistic Regression.

Keywords: *Software Defined Networking (SDN), Logistic Regression (LR), Distributed Denial of Service (DDOS), K-Nearest Neighbors (KNN).*

1. Introduction

When a website is flooded with traffic congestion from numerous sources and due to this the users become inaccessible to the network is referred to as a DDOS assault. By overloading a website or service with traffic, which may cause it to crash or take longer to react, a DDOS attack seeks to prevent it from performing its normal functions. Fig. 1 shows the DDoS attack.

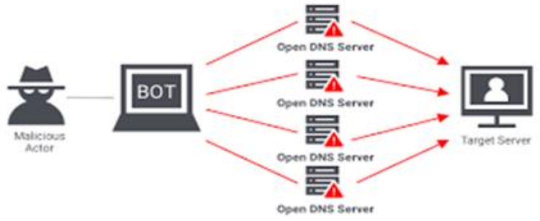


Fig.1. DDoS attack

Machine learning is classified into numerous forms, including unsupervised learning, supervised learning and semi-supervised learning. The algorithm used in supervised learning has been trained using labelled data. This means that the expected output for each input is already known. Unsupervised learning, on the other hand, is training an algorithm on unprocessed data and letting it to identify patterns or structures on its own. Combination of supervised and unsupervised learning is said to be semi-supervised learning. Fraud detection, image with audio identification, recommendation systems, natural language processing and many more applications employ machine learning techniques. The following listed algorithms are employed here.

- Logistic Regression (LR)
- K-Nearest Neighbors (KNN)

1.1. Logistic Regression

Logistic regression is a machine learning method used for categorization jobs. The result of logistic regression is a binary variable, such as whether or not an email is spam. Any real-valued input is transformed into a value using the logistic function that lies within the range of 0 to 1. As a result, it adopts a "S" curve form also known as the sigmoid function or logistic function. If the anticipated probability exceeds a threshold value typically 0.5 then the input is classified as fitting to the positive class; if not, it is classified as fitting to the negative class. The S-form Curve for Logistic Regression is shown in Fig. 2.

The logistic function has the following form:

[1*] *Assistant Professor(Sr.Gr.),*
Department of Electronics and Communication Engineering,
Ramco Institute of Technology, Rajapalayam – 626117, Tamilnadu,, India
ORCID ID : 0000-0002-6240-5420
[2] *Assistant Professor,*
Department of Electronics and Communication Engineering,
Ramco Institute of Technology, Rajapalayam – 626117, Tamilnadu,, India
[3] *Assistant Professor,*
Department of Electronics and Communication Engineering,
Ramco Institute of Technology, Rajapalayam – 626117, Tamilnadu,, India
[4] *Assistant Professor,*
Department of Computer Science and Engineering,
Ramco Institute of Technology, Rajapalayam – 626117, Tamilnadu,, India
[5] *Professor & Head, Department of Information Technology,*
R.M.D Engineering College, Chennai– 601206, Tamilnadu,, India
[1]* *Corresponding Author Email: ggpriya2019@gmail.com*

$$sigmoid = \frac{1}{(1+e^{-z})}$$

□□

□□□□□□□□□□□□□□□□□□

□□□□□□□□□□□□□□□□□□□□□□□□ (1)

where z represents the linear combination of the coefficients and input variables.

$$z = w_o + w_1 x_1 + w_2 x_2 + ... + w_n x_n \quad (2)$$ □

A technique known as maximum likelihood estimation is used to estimate the coefficients w_0, w_1, w_2, and w_n from the training data. The aim of logistic regression is to determine the coefficient values that maximize the training data's probability.

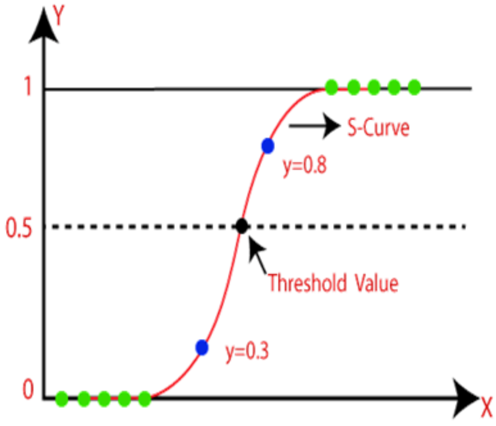


Fig.2. S-form Curve for Logistic Regression

1.2. KNN Algorithm

For problems with classification and regression, the machine learning technique K-Nearest Neighbors (KNN) is utilized. This straightforward technique relies on the notion of "nearest neighbours" to function. A new data point is identified using KNN by locating the k labelled instances that are closest to it in the training dataset, where k is a user-defined value. On the basis of the majority class of its k closest neighbours, the class of the new data point is then predicted. Regression involves making an estimate by averaging the data of its k closest neighbours.

The K value, which establishes the number of neighbours to take into account, is the primary hyper parameter of the KNN algorithm. The decision border becomes more rounded with a higher value of K and more sharp with a lower value of K. The simplicity of the KNN algorithm in terms of implementation and interpretation is one of its benefits.

The distance metrics that KNN algorithms most frequently employ are Euclidean Distance and Manhattan Distance.

Euclidean Distance:

In an n-dimensional space, consider any two points 'p' and 'q'. According to Euclidean distance, the distance between these two points is given by:

$$d(p,q) = \sqrt{\sum (x_i - y_i)^2}$$

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

(3)

where x_i and y_i are, respectively, the ith dimension's values for points p and q.

Fig. 3 illustrates the implementation of KNN algorithm.

Manhattan Distance:

It is also known as the L1 distance. Considering the same two points 'p' and 'q' in an n-dimensional space, the L1 distance is established by:

$$d(p,q) = \sum |x_i - y_i| \quad (4)$$

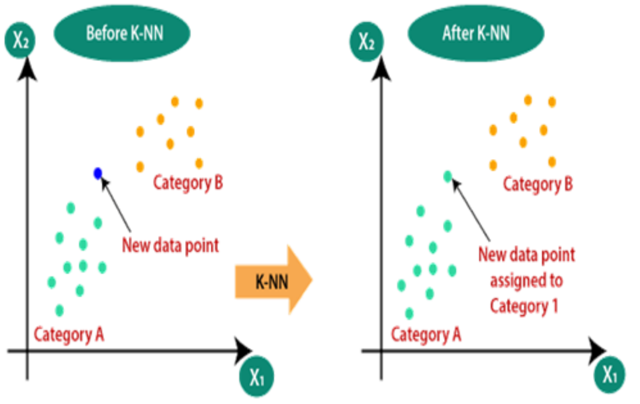


Fig.3. KNN Algorithm

The KNN method calculates the distances between the query point and every other point in the dataset once the distance metric has been chosen, and then chooses the K closest points based on the distance values.

2. Literature Survey

If you are using Word, use either the Microsoft Equation Editor or the MathType add-on (http://www.mathtype.com) for equations in your paper (Insert | Object | Create New | Microsoft Equation or MathType Equation). "Float over text" should not be selected.

2.1. Equations

Low-rate denial of service (LDOS) attacks transmit high-intensity burst data streams to targets in order to lower TCP traffic and limit network service capabilities, according to Low-Rate DOS Attack Detection based on Improved Logistic Regression [1]. Even though various LDOS attack detection techniques have been suggested, these techniques suffer from poor real-time performance, large overhead, and low efficiency. Since the TCP traffic during the LDOS assault is lower than the ordinary average value and its distribution is more discontinuous, this approach makes use of the network traffic to extract the eigenvalues such as variance, average TCP and sample entropy as the foundation to categorize the traffic data. In order to assess whether an LDOS attack has taken place in the network, regression analysis is utilized to detect the presence/absence of aberrant traffic in accordance with the derived classifier. Experiments on NS-2 and the test-bed show that the method in this work may efficiently and instantly identify LDOS assaults with low false negative rate, high accuracy and false positive rate. Its complexity is also reduced.

One common tactic used in security hacking to disrupt geographical networks or render computational resources unavailable is denial of service (DOS), according to Denial of Attack (DOS) Detection: Performance Comparison of Supervised Machine Learning Algorithms [2]. In this, they used data that was available to the public to identify DOS attacks using the Naive Bayes approach, Logistic Regression and Artificial Neural Networks. The tests' findings show that given a dataset with a little unbalanced distribution, artificial neural networks performed better in terms of balanced accuracy and ROC curve than the Naive Bayes method and also logistic regression.

According to Machine Learning Approaches for Combating Distributed Denial of Service Attacks in Modern Networking

Environments [3], a DDOS attack is a huge risk to service providers. A DDOS attack, attacks a target by flooding it with an overwhelming amount of malicious requests in an attempt to disrupt and deny services to legitimate users. Through the use of ML approaches, many defense systems have, in fact, been turned into smart and intelligent systems that can resist DDOS attacks. In light of recent discoveries, this study examines how the DDOS detection techniques are updated for application in single and hybrid ML methods in modern networking conditions. The paper also explores machine learning (ML) techniques as security solutions against denial-of-service (DDOS) assaults in IOT contexts, as the growth of the Internet of Things (IOT) has garnered substantial scholarly attention in recent years. The report also suggests other lines of inquiry for further study. This effort objects to aid the research community in designing and creating defenses systems that are successful against various DDOS attacks.

In DDOS Attack Detection Method Based on Improved KNN with the Degree of DDOS Attack in Software-Defined Networks [4], network availability has been greatly reduced by DDOS attacks, and there is presently no effective security against them. However, a novel strategy for DDOS assault defense is provided by the recently developed Software Defined Networking (SDN). Two techniques for DDOS attack detection in SDN are presented in this research. To gauge the level of the DDOS attack, one technique uses its intensity. The alternative approach locates the DDOS attack by utilizing the machine learning (ML)-based, enhanced K-Nearest Neighbors (KNN) technique. The outcomes of the theoretical and the actual results on datasets show that the proposed methodologies are better than the currently used schemes for identifying DDOS attacks.

In Automated DDOS attack detection in software defined networking [5], by enabling the programmability of network devices, the networking paradigm known as "Software-Defined Networking" (SDN) has reframed the term "network." Network engineers can swiftly monitor networks, administer networks centrally, and detect fraudulent traffic and connection failures with pinpoint accuracy. Along with the freedom that SDN provides, additionally it is susceptible to denial-of-service assaults (DDOS) that could crash the network as a whole. The research suggests utilising machine learning to separate DDOS attack traffic from benign traffic in order to counteract this assault. This paper's main contribution is the discovery of new characteristics for DDOS attack detections. The categorization is carried out using a brand-new hybrid machine learning model

3. Proposed Methodology

The proposed methodology steps are as follows:

- The pandas software is used to read the input dataset first.

- The data is pre-processed by eliminating null values after feature selection, which entails choosing input features to transmit into the module.

- Finally, we build a very accurate model, add the gathered attributes to the model, and train the computer.

- After training, the model is fed test data in order to anticipate DDOS data assault.

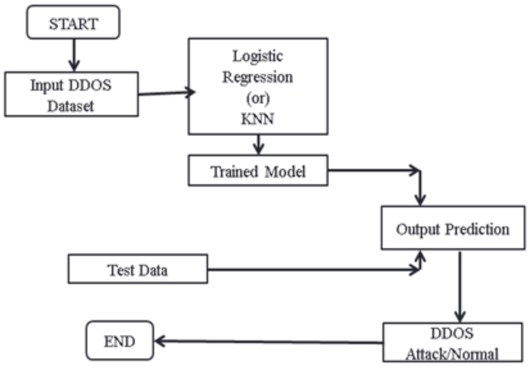


Fig. 4. Block Diagram of the proposed model

The Data Pre-Processing, Feature Selection, Trained

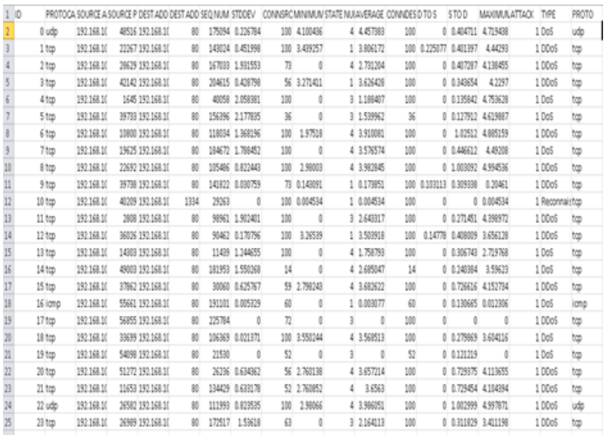
Model, and Test Data make up the proposed flow diagram. Fig. 4 shows the block diagram of the proposed model.

The dataset is first retrieved using the pandas library, and then it is saved inside a pandas data frame. As the dataset initially has a lot of null values, it is removed completely because the machine learning model is unable to handle them.

4. Dataset

Kaggle is a well-known platform where data scientists and machine learning enthusiasts may compete and cooperate on projects. The "DDOS Attack Detection" dataset is one of those accessible on Kaggle.

This dataset comprises network traffic data obtained during a DDOS assault on a web server. A botnet is a network of compromised computers that is run by hackers, and it was used to carry out the attack.



1	ID	PROTOCA	SOURCE A	SOURCE P	DEST ADD	DEST ADD	SEQ.NUM	STDDEV	CONNSRC	MINIMUM	STATE NU	AVERAGE	CONNDES	D TO S	S TO D	MAXIMUM	ATTACK	TYPE	PROTO
2	0	udp	192.168.1(	48516	192.168.1(	80	175094	0.226784	100	4.100436	4	4.457383	100	0	0.404711	4.719438	1	DoS	udp
3	1	tcp	192.168.1(	22267	192.168.1(	80	143024	0.451998	100	3.439257	1	3.806172	100	0.225077	0.401397	4.44293	1	DDoS	tcp
4	2	tcp	192.168.1(	28629	192.168.1(	80	167033	1.931553	73	0	4	2.731204	100	0	0.407287	4.138455	1	DDoS	tcp
5	3	tcp	192.168.1(	42142	192.168.1(	80	204615	0.428798	56	3.271411	1	3.626428	100	0	0.343654	4.2297	1	DDoS	tcp
6	4	tcp	192.168.1(	1645	192.168.1(	80	40058	2.058381	100	0	3	1.188407	100	0	0.135842	4.753628	1	DoS	tcp
7	5	tcp	192.168.1(	39733	192.168.1(	80	156396	2.177835	36	0	3	1.539962	36	0	0.127912	4.619887	1	DoS	tcp
8	6	tcp	192.168.1(	10800	192.168.1(	80	118034	1.368196	100	1.97518	4	3.910081	100	0	1.02512	4.885159	1	DDoS	tcp
9	7	tcp	192.168.1(	19625	192.168.1(	80	184672	1.788452	100	0	4	3.576574	100	0	0.446612	4.49208	1	DoS	tcp
10	8	tcp	192.168.1(	22692	192.168.1(	80	105486	0.822443	100	2.98003	4	3.982845	100	0	1.003092	4.954536	1	DDoS	tcp
11	9	tcp	192.168.1(	39738	192.168.1(	80	141822	0.030759	73	0.143091	1	0.179851	100	0.103113	0.309338	0.20461	1	DDoS	tcp
12	10	tcp	192.168.1(	40209	192.168.1(	1334	29263	0	100	0.004534	1	0.004534	100	0	0	0.004534	1	Reconnai:	tcp
13	11	tcp	192.168.1(	2808	192.168.1(	80	98961	1.902401	100	0	3	2.643317	100	0	0.271451	4.398972	1	DDoS	tcp
14	12	tcp	192.168.1(	36026	192.168.1(	80	90462	0.170796	100	3.26539	1	3.503918	100	0.14778	0.408009	3.656128	1	DDoS	tcp
15	13	tcp	192.168.1(	14303	192.168.1(	80	11439	1.244655	100	0	4	1.758793	100	0	0.306743	2.719768	1	DoS	tcp
16	14	tcp	192.168.1(	49003	192.168.1(	80	181953	1.550268	14	0	4	2.685047	14	0	0.240384	3.59623	1	DoS	tcp
17	15	tcp	192.168.1(	37862	192.168.1(	80	30060	0.625767	59	2.798243	4	3.682622	100	0	0.726616	4.152734	1	DDoS	tcp
18	16	icmp	192.168.1(	55661	192.168.1(	80	191101	0.005329	60	0	1	0.003077	60	0	0.130665	0.012306	1	DoS	icmp
19	17	tcp	192.168.1(	56855	192.168.1(	80	225784	0	72	0	3	0	100	0	0	0	1	DDoS	tcp
20	18	tcp	192.168.1(	33699	192.168.1(	80	106369	0.021371	100	3.550244	4	3.568513	100	0	0.279869	3.604116	1	DoS	tcp
21	19	tcp	192.168.1(	54098	192.168.1(	80	21530	0	52	0	3	0	52	0	0.121219	0	1	DoS	tcp
22	20	tcp	192.168.1(	51272	192.168.1(	80	26236	0.634362	56	2.760138	4	3.657214	100	0	0.729375	4.113655	1	DDoS	tcp
23	21	tcp	192.168.1(	11653	192.168.1(	80	134429	0.633178	52	2.760852	4	3.6563	100	0	0.729454	4.104394	1	DDoS	tcp
24	22	udp	192.168.1(	26582	192.168.1(	80	111993	0.823535	100	2.98066	4	3.986051	100	0	1.002999	4.997871	1	DDoS	udp
25	23	tcp	192.168.1(	26989	192.168.1(	80	172517	1.53618	63	0	3	2.164113	100	0	0.311829	3.411198	1	DDoS	tcp

Fig. 5 DDOS Attack Dataset

Fig. 5 shows the dataset that includes 17 features in total, including the source IP address, the protocol, packet size and the destination IP address. The data has been labeled, with each event labeled as either normal or attack.

The training set and test set of the dataset are separated, accordingly. Eighty percent of the samples originate from the dataset's training set, with the remaining twenty percent coming from the test set.

5. Confusion Matrix

To evaluate how effectively a machine learning model works in a supervised learning setting, a confusion matrix is used. It is a matrix of actual and expected classes, where the diagonal indicates the proportion of accurate forecasts and the off-diagonal parts the proportion of false predictions.

The following are the four categories of the confusion matrix:

5.1. True Positive (TP)

The quantity of observations that the model correctly predicted to be positives and that are also true positives.

5.2. False Positive (FP)

The amount of observations that the model incorrectly forecasts as positive when they are actually negative is known as False Positive (FP).

5.3. True Negative (TN)

The proportion of data that precisely match the model's expectation that a negative value will exist is known as True Negative (TN).

5.4. False Negative (FN)

The percentage of data that the model incorrectly forecasts as negative but are really positive is known as a false negative (FN).

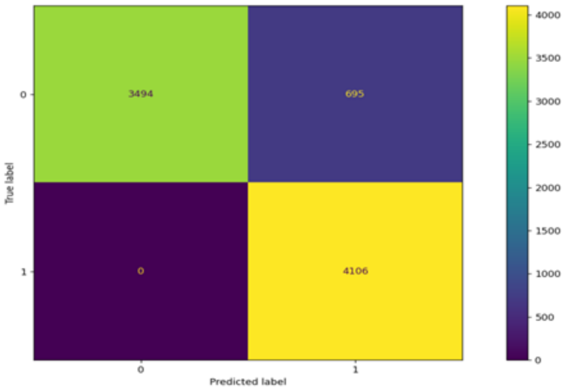
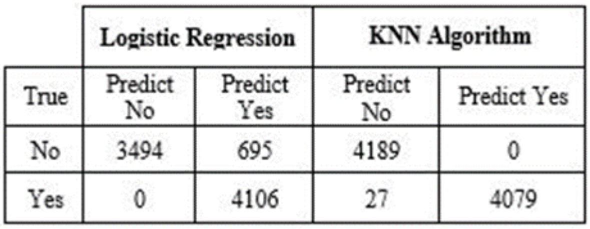


Fig. 6. Confusion Matrix for Logistic Regression Algorithm

Figure 6 Confusion Matrix for Logistic Regression Algorithm shows the predicted outcomes of the model for a binary classification problem and figure 7 Confusion Matrix for KNN Algorithm displays the expected results of the model for a binary classification issue.

Table 1 Result of Confusion Matrix



	Logistic Regression		KNN Algorithm	
True	Predict No	Predict Yes	Predict No	Predict Yes
No	3494	695	4189	0
Yes	0	4106	27	4079

Table 1 Result of Confusion Matrix shows the performance of KNN and logistic regression models in a binary classification problem. The KNN model predicted 4079 true positives, 0 false positives, 4189 true negatives, and 27 false negatives, while the logistic regression model predicted 4106 true positives, 695 false positives, 3494 true negatives, and 0 false negatives.

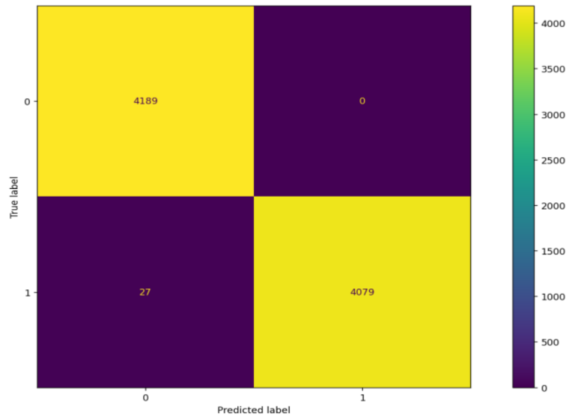
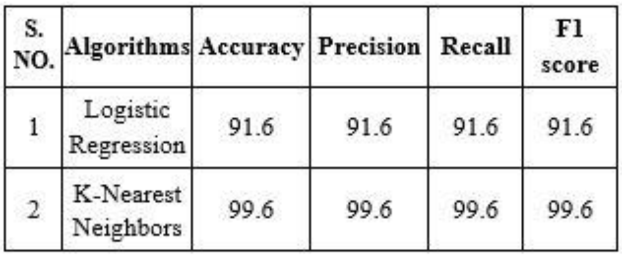


Fig. 7. Confusion matrix for KNN algorithm

6. Accuracy

Table 2 Accuracy Table



S. NO.	Algorithms	Accuracy	Precision	Recall	F1 score
1	Logistic Regression	91.6	91.6	91.6	91.6
2	K-Nearest Neighbors	99.6	99.6	99.6	99.6

From Table 2, K-NN does not always exceed logistic regression in terms of accuracy. However, there could be certain circumstances in which K-NN performs better than logistic regression. A case in point is when the decision border between classes is extremely complicated and non-linear.

By taking into account the local density of points surrounding the new data point in these situations, k-NN is able to capture the complicated decision boundary.

On the other hand, logistic regression may be unable to capture the non-linear connections between characteristics since it presumes a linear decision boundary between classes.

7. Graph Results

The graphs for Logistic Regression and KNN Algorithms are shown in Fig. 8 and 9.

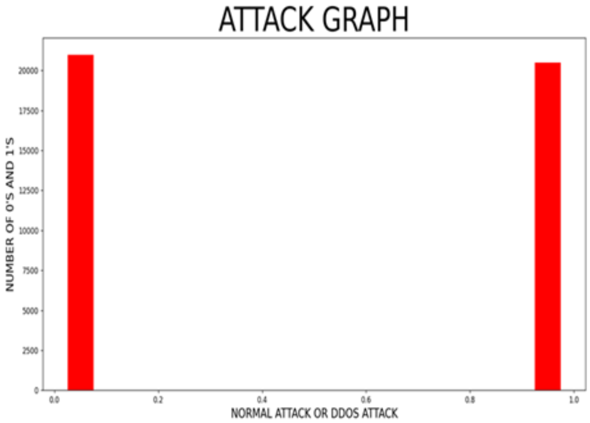


Fig. 8 Attack Graph for Logistic Regression

Both graphs describe the attack's Histogram. The X-axis indicates whether the attack is normal or DDOS. The Y-axis represents the amount of normal or DDOS traffic. The 17th column in the Excel format dataset aids in determining if the traffic is normal or a DDOS attack.

Not only this graph, but a variety of graphs depending on dataset columns show differences between the two methods, which will aid in predicting accuracy.

Both Logistic Regression and KNN algorithms have the similar Attack Graph and some other result also. But both are differ in Accuracy value which is already described in Table 2.

The zero (0) represents number of Normal traffic present in the dataset. The one (1) represents number of DDOS Traffic present in the given dataset. For our experiment, the given dataset contains equal number of 1's and 0's.Therefore, for our dataset 50% attack and 50% of Normal Traffic.

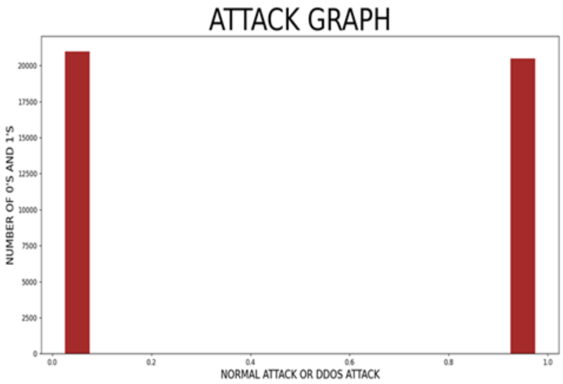


Fig. 9. Attack Graph for KNN Algorithm

8. Conclusion

Despite its numerous benefits, SDN also handles the DDO issue, which is the furthermost prevalent security concern in the network. The central control of SDN has the benefit of making the SDN controller more vulnerable to DDOS assaults, a security risk. In order to address this issue, the

Logistic Regression and KNN machine learning techniques are utilized in this research to analyse the DDOS assault detection and defence mechanisms. A controller for SDN manages networking operations. The trained Python code will be added to the controller. The controller can identify a DDOS assault and stop the network from falling down by receiving the DDOS traffic pattern from any hosts. We can forecast the DDOS assault with 99% accuracy using KNN. We can accurately forecast the DDOS assault 91% of the time using logistic regression.

Acknowledgements

We owe a debt of gratitude to Ramco Institute of Technology and R.M.D Engineering College for generating the concepts and for giving all necessary resources and assistance.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Yudong Yan ,Dan Tang ,Rui Dai , Jingwen Chen, “Low-Rate DOS Attack Detection Based on Improved Logistic Regression”, in IEEE Access, vol. 7, pp. 160536-160545, 2017.

[2] Zhuolin Li, Hao Zhang, Hossain Shahriar, Michael Whitman, Dan Lo1 “Denial of Service (DOS) Attack Detection: Performance Comparison of Supervised Machine Learning Algorithms”, in IEEE Access, Volume 187, November 2021.

[3] J S.Ahamed Aljuhani “Machine Learning Approaches for Combating Distributed Denial of Service Attacks in Modern Networking Environments”, volume 8, pp. 5039-5048, 2017.

[4] Shi Dong, Mudar Sarem, Dr. Anita Kanavalli,“ DDoS Attack Detection Method Based on Improved KNN With the Degree of DDoS Attack in Software-Defined Networks”, volume 7, 30 May 2019, pp. 46 – 47.

[5] Nisha Ahuja , Gaurav Singal , Debajyoti Mukhopadhyay , Neeraj Kumar “ Automated DDOS attack detection in software defined networking”, Volume 187 ,1 August 2021.

[6] Asmaa A.Elsaeidy, Abbas Jamalipour , “ A Hybrid Deep Learning Approach for DDOS Detection and Classification” 27 Oct 2019 IEEE 2nd International Conference on Computer and Communication Engineering Technology (CCET), 2019, volume 19, pp. 249-253.

[7] Ismail,Muhammad ,Ismail Mohmand ,Hameed Hussain, Ayaz Ali khan ,Ubaid Ullah ,Muhammad Zaka, “A Machine Learning-Based Classification and Prediction Technique for DDoS Attacks” ,International Conference on Information and Communication Technology Convergence,17 Feb2019, volume 19, pp.363-368.

[8] Asmaa A. Elsaeidy, Abbas Jamalipour, Kumudu S ,“A Hybrid Deep Learning Approach for Replay and DDoS Attack Detection in a Smart City”,Volume 12, 29 January 2020.

[9] Victor Ferman, Maiguel Angel Medina Perez,Raul Monroy, ”DNS-ADVP: A Machine Learning Anomaly Detection and Visual Platform to Protect Top-Level Domain Name Servers Against DDoS Attacks” ,in IEEE Access, volume 8, pp. 161908-161919, 2018.

[10] Randy C. Paffenroth, Worcester, MA Chong Zhou “Modern Machine Learning for Cyber-defense and Distributed Denial of Service Attacks”, in IEEE Access, volume 8, pp. 155859-155872, 2018.

[11] Derya Erhan,(Member, Ieee),And Emin Anarim “Hybrid DDoS Detection Framework Using Matching Pursuit Algorithm”, in IEEE Access, volume 8, pp. 118912 – 118923.2020.

[12] Jalal Bhayo ,Sufian Hameed , And Syed Attique Shah “An Efficient Counter-Based DDoS Attack Detection Framework Leveraging Software Defined IoT (SD-IoT)”, in IEEE Access, volume 8, pp. 221612 - 221631, 2020.

[13] Da Yin1 ,Lianming Zhang , And Kun Yang ,(Senior Member,IEEE) “A DDOS Attack Detection and Mitigation With Software-Defined Internet of Things Framework”, in IEEE Access, volume 6, pp. 24694 - 24705, 2018.

[14] Kiwon Hong,Younjun Kim,Hyungoo Choi, and Jinwoo Park, “SDN-Assisted Slow HTTP DDOS Attack Defense Method”, in IEEE Access, volume 22, pp. 688 - 691, 2017.

[15] Rajorshi Biswas ,Sungji Kim, and Jie Wu , Fellow, “Sampling Rate Distribution for Flow Monitoring and DDOS Detection in Datacenter”,in IEEE Access, volume 16, pp. 2524 - 2534, 2021.

[16] Yonghao Gu ,Kaiyue Li, Zhenyang Guo,And Yongfei Wang, “Semi-Supervised K-Means DDoS Detection MethodUsing Hybrid Feature Selection Algorithm”, in IEEE Access, volume 7, pp. 64351 - 64365, 2019.

[17] Zakaria Abou El Houda, Lyes Khoukhi, and Abdelhakim Senhaji Hafid, “Bringing Intelligence to Software Defined Networks: Mitigating DDOS Attacks”, in IEEE Access, volume 17, pp. 2523 - 2535, 2020.