IDENTIFICATION AND PREVENTION OF DDOS ATTACKS USING MACHINE LEARNING ALGORITHMS

G. Mokshagna
School of Computer science & Engineering
REVA University
Bengaluru, India
govindumoksha@gmail.com

E. Vamshi Krishna
School of Computer science & Engineering
REVA University
Bengaluru, India
vamkrishna9122@gmail.com

G. Bharath
School of Computer science & Engineering
REVA University
Bengaluru, India
gotlabharath3@gmail.com

G. Nithin
School of Computer science & Engineering
REVA University
Bengaluru, India
gotlabharath3@gmail.com

Kiran Kumar A
School of Computer science & Engineering
(Assistant Professor)
REVA University
Bengaluru, India
kiransachi.a@gmail.com

***Abstract*-- A distributed denial of service (DDoS) attack involves flooding a server machine with an excessive amount of requests via the internet or intranet, with the intent of making a computer or networking available partially or completely unavailable to legitimate users. The primary objective of a DDoS assault is to overwhelm the system with unexpected traffic and shut down its services. Software-defined networking (SDN) is a network architecture that allows the creation and virtualization of hardware components, enabling the dynamic modification of network connections. Unlike traditional networks, SDN offers flexibility in network configuration. Despite this, SDN is still susceptible to DDoS attacks, which are a significant threat to the internet. To mitigate such attacks, machine learning algorithms can be utilized. This paper presents an approach for identifying and preventing DDoS attacks by utilizing methods for machine learning algorithms includes Gradient Boosting Classifier, Random Forest, Logistic Regression and Stacking Classifier. Each attack class's performance of these models is assessed separately in terms of accuracy.**

Keywords—DDoS, Software Defined Network, Logistic Regression, Random Forest, Gradient Boosting and Stacking Classifier.

1. INTRODUCTION

The objective of a DDoS attack is to disrupt the access of legitimate applications to network infrastructure through the flooding of an excessive amount of traffic with no intention of retrieving sensitive information or compromising credentials. In this malicious activity, the targeted system's resources are overwhelmed while also experiencing network congestion because of unsuccessful traffic delivery. The use of client/server technology in a DDoS(Distributed Denial of-Service) attack allows multiple computers to be employed as an attack platform, which is then directed at one or more targets, increasing the attack's impact. This attack has altered the traditional peer-to-peer attack model and may be difficult to differentiate from regular behavior by analyzing the protocols and services utilized. Due to the difficulty of detecting a DDoS attack, most defense technologies rely on network intrusion detection methods. Researchers have identified a few characteristics to describe DDoS attacks attributes, providing the flow density, the amount of source IP addresses, and the destination ports.

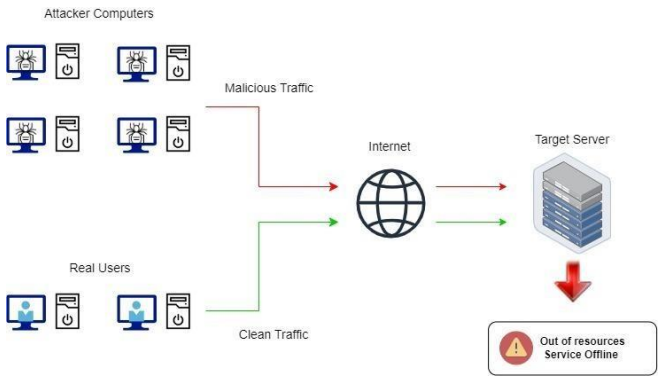

Fig 1.1 Operation of a DDoS attacks

Software-Defined Networking (SDN), a revolutionary networking technique, separates the control and data planes of network devices, overcomes the drawbacks of conventional network architectures. The data plane, control plane, and application plane are the three planes that make up the SDN architecture. The control plane governs traffic flow by creating routing tables, and the controller's decisions determine how the data plane delivers network traffic. The

Additional applications like load balancers, firewalls, and quality of service, or QoS, apps are managed by the application plane. By separating the forwarding and network control duties, SDN design improves network performance. Numerous routers are managed by control software running on a logically centralized controller.

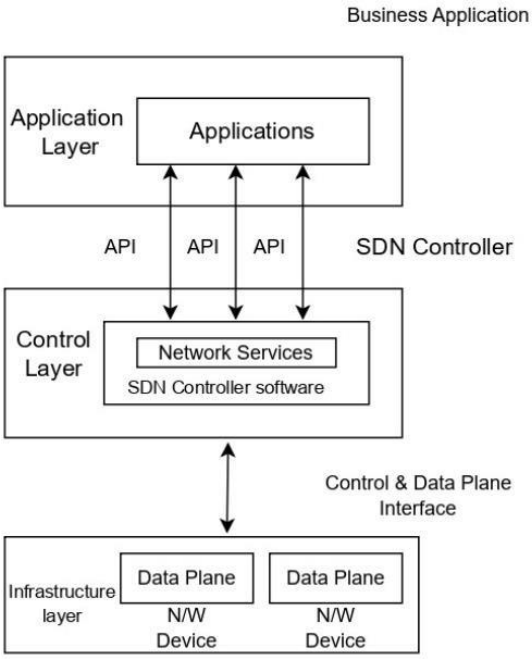

Fig 1.2 SDN Architecture

2. RELATED WORK

The internet has become a necessary component of daily life, with uses in everything from banking, education, and transportation to healthcare, entertainment, and e-commerce. This study specializes in applying machine learning approaches to identify application layer DoS/DDoS threats. Given the difficulty of manually monitoring network data, a sophisticated defence system that can recognize attacks is required. The proposed offers a straightforward but effective way to get improved results, making it more effective than current methods.

In this particular section, we will be analyzing the prior research studies that have been a study that was done in the area of Software-Defined Networking (SDN) with the goal of detecting DDoS attacks. Experiments take place on an extensive variety of machine learning models, such as Random Forest (RF), and Logistic regression (LR), Gradient Boosting(GB),stacking classifier (SC). Applying innovative techniques, assessment of performance is carried out accurately. The accuracy scores achieved by the various machine learning models on the DDoS attack dataset were as follows: Random Forest achieved a score of 0.97, while Logistic Regression achieved 0.982, and Gradient Boosting achieved 0.991.

A hybrid machine learning technique was presented in the paper for identifying DDoS attacks in an SDN environment. The approach involved utilizing a Stacking Classifier to identify malicious activities. In this, Stacking Classifier is the combination of Random Forest, Logistic Regression, and Gradient Boosting. According to the results, the approach was able to achieve a detection rate ranging from 96% to 99% while utilizing the 30% sampling rate.

Studies in the literature have demonstrated that deploying complex machine learning models can yield high accuracy results in detecting DDoS attacks, but some of the models come with a high computational cost. Conversely, while simple methods have a low computational cost, their attack detection accuracy is often poor or moderate. Our endeavor is to encounter these limitations and eradicate them.

3. PROPOSED METHODOLOGY

The framework contains an early detection mechanism based on machine learning that is intended to effectively and collaboratively resist multiple attacks at the application level.

To achieve this, the approach involves using network traffic data from a DDoS attack ideal for experimenting with and training a dataset. The investigation's starting point process is to preprocess the dataset by eliminating null values and transforming actual numbers to values in integers.

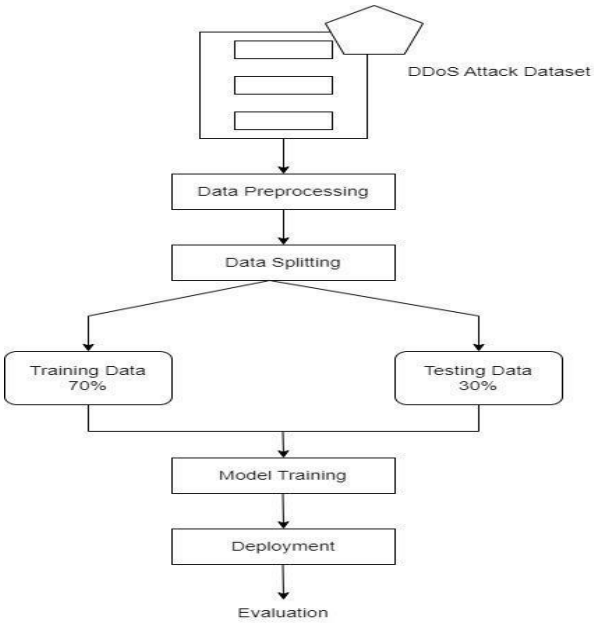

Fig 3.1 Flowchart illustrating the suggested approach.

The dataset is divided into training and testing in proportions of 70% to 30% after cleaning. 30% of the dataset was utilized to test the models, while 70% was used to train the models. Then the training of model (Random Forest, Logistic Regression, Gradient Boosting) takes place. After training, deployment of model takes place. The models' performance is determined in terms of precision.

Machine learning algorithms:

Because machine learning models are widely used, various modifications have been documented in the literature, all of which are capable of obtaining acceptable classification performance. Furthermore, the Sci-Kit package provides useful utilities for implementing these models.

This study employs a number of Machine learning classifications that can differentiate between hazardous and legitimate traffic, namely Random Forest, Logistic Regression, Gradient Boosting, and Stacking Classifier. A brief description of each model is included for completeness.

A. Random Forest:

A popular machine learning method for addressing problems with regression and classification is random forest. To solve complicated issues, it leverages ensemble learning, which combines several classifiers.

The "forest" created by the random forest approach is taught from a number of decision trees. using a technique called bagging, which helps machine learning algorithms be more precise. The random forest method uses the decision trees' predictions to determine an outcome. It makes predictions by averaging or increasing the amount of trees and using the mean of the results from several trees can improve the accuracy of the result.

Random forest is an upgraded variant of decision tree algorithms since it lowers dataset overfitting and hence increases precision. Furthermore, unlike tools such as Scikit learn, it delivers predictions without requiring substantial configuration. The random forest algorithm offers various advantages over the decision tree approach, including being more accurate, providing an effective solution to manage missing data, producing fair predictions without hyperparameter tuning, and overcoming the decision tree overfitting issue. Additionally, a subset of features is randomly selected in each random forest tree at the node's splitting point.

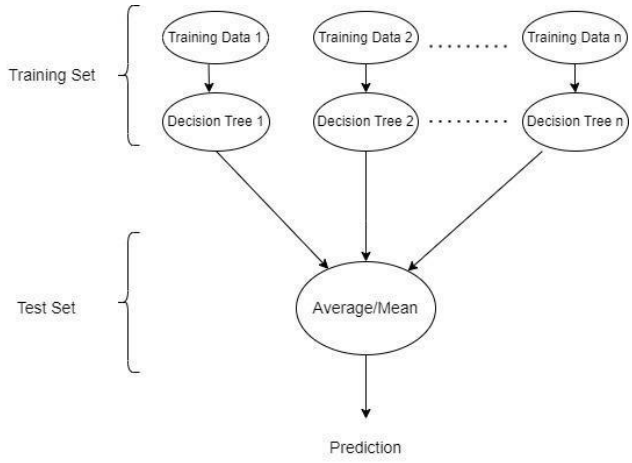

Fig 3.2 Algorithm of random forest

B. Logistic Regression:

Classification issues are frequently addressed using the logistic regression (LR) method. A probabilistic framework that is frequently utilized for density estimation issues, the maximal likelihood estimation (MLE) framework, is used to ascertain the LR parameters. In a dataset, LR establishes a link between both independent and dependent variables. The logistic function, which is additionally referred to as the function of sigmoid, is used to assess the probability that any number of independent variables would have an impact on the dependent variable. The sigmoid function translates real integers from 0 to 1 to produce an S-shaped curve. $\sigma(Z)\frac{1}{1+e^{-z}}$

The output of the logistic function(σ(z)) is a probability estimate, which is bound between 0 and 1. The input of the logistic function is the prediction made by the algorithm, denoted as 'z'. The natural logarithm with base e is also known as the Euler quantity.

The logistic function maps a curve that either tends to +∞ or -∞, given that the real number range is +∞ to -∞. The value becomes 0 if it drops beneath negative infinity and 1 otherwise.

C. Gradient Boosting:

The gradient boosting algorithm (GB) is a strong method for building predictive models that reduces the mean square error (MSE) by combining many weak learning models. Boosting algorithms, as opposed to bagging algorithms, can successfully deal with both Variability and bias trade-offs. To solve the disadvantages of weak learners, the gradient boosting approach employs a gradient in the loss function. The validation set is just one of several variables that the MSE considers while calculating the average difference between anticipated and

actual values. To assess how well the model coefficients fit the underlying data, the loss function is used.

The gradient boosting method (as a regressor or classifier) may predict both continuous and categorical target variables. When used as a regressor, Mean Square Error (MSE) is the cost function,when used as a classifier, Log loss is the cost function.

$$\text{Loss} = \sum(y_i - y_i^p)^2$$

Where $f(y_i, y_i^p)$ is the loss function and y_i is the ith targeted value and y_i^p is the ith anticipated value.

D. Stacking Classification:

A common strategy in machine learning is ensemble modelling, which combines various models to enhance performance. There are many ensemble strategies, including bagging, boosting, and stacking, the last of which is especially well-liked. To improve the output prediction model, stacking combines the predictions made by a number of mediocre learners using Meta learners. In this method, the algorithm discovers the most effective way to combine the output predictions of sub-models to get a prediction that is more precise. As an extension of the Model Averaging Ensemble approach, stacking is also known as stacked generalization. All sub-models contribute to the new model based on their performance weights.

Multiple base or learner models are used in an architecture which layers models and a kind of meta-model that integrates their predictions is also used. The stacking architecture is composed of the base models, also known as level 0 models, and the meta-model, also known as the firstlevel model. original data for training, primary level models, forecasts at the primary level, a second-level model, and final prediction are all included in the ensemble method known as stacking.

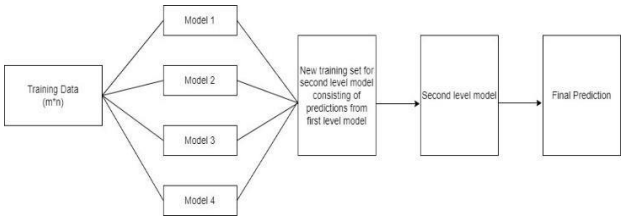

Fig 3.3 Stacking classification architecture

4. PERFORMANCE OF MACHINE LEARING MODELS

The effectiveness of the planned job is assessed using precision. The precision score of a model indicates its correctness in terms of accurately classifying attacks. It is determined by dividing the total number of accurate forecasts by the entire number of predictions.

Model	Testing Size	Accuracy
Logistic Regression	30%	97.95%
Random Forest	30%	98.30%
Gradient Boosting	30%	99.80%
Stacking Classifier	30%	97.10%

5. RESULT

The findings of the studies performed to identify DoS/DDoS attacks are reported in this section. Using data from DoS/DDoS attacks at the application layer, numerous experiments were carried out utilizing machine learning models to accomplish this goal.

5.1 Accuracy without stacking classifier.

The initial experiments aimed to classify accuracy results of a dataset of DoS/DDoS attacks using machine learning models without any Stacking Classifier. The studies used a total of 47 qualities, and the results are shown in the table above.

The accuracy attained by three diverse approaches, such as logistic regression, random forest and gradient boosting is shown in the graph below. The graph shows that the accuracy range for Logistic Regression is between 96 and 98%, whereas the accuracy range for Random Forest is between 97 and 99%. Gradient Boosting, on the other hand, performs the best out of the three algorithms, with an accuracy range of 98 to 99.7%.

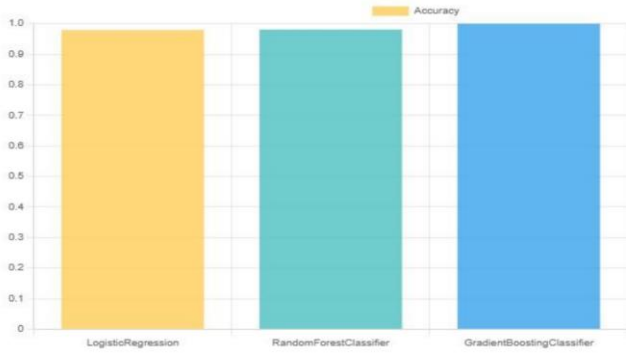

Fig 5.1 Accuracy without stacking classifier.

5.2 Accuracy with stacking classifier

The accuracy of the stacking classifier, which is a mixture of three algorithms, ranged from 97 to 100%, outperforming individual computer systems learning models like Random

Forest and Logistic Regression. The graph below displays the accuracy of every machine learning model used in the experiment as well as the stacking classifier.

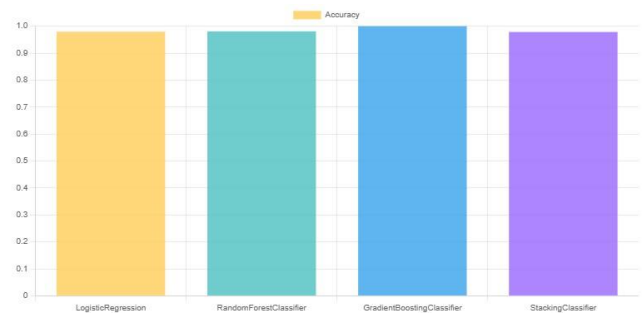

Fig 5.2 Accuracy with stacking classifier.

6. CONCLUSION

We have created a user-friendly website and inserted the DDoS attack Dataset as an input. The data is then subjected to several preprocessing techniques and fed into various machine learning models, including Random Forest, Logistic Regression, and Gradient Boosting. By selecting a specific model, the accuracy of the dataset can be determined. The system is designed to determine whether the network is safe or under attack based on the input data. The web page is built using Python programming and Flask framework. Furthermore, a Stacked Classifier was employed as a combination of all three machine learning models in this study.

7. REFRENCES

[1]. Muthamil Sudar, K., & Deepalakshmi, P. (2020). A two level security mechanism to detect a DDoS flooding attack in software-defined networks using entropy-based and C4. 5 technique. Journal of High Speed Networks, (Preprint), 1- 22.

[2]. Dong, S., Abbas, K., & Jain, R. (2019). A survey on distributed denial of service (DDoS) attacks in SDN and cloud computing environments. IEEE Access, 7, 80813- 80828.

[3]. Muthamil Sudar, K., & Deepalakshmi, P. (2020). A two level security mechanism to detect a DDoS flooding attack in software-defined networks using entropy-based and C4. 5 techniques. Journal of High Speed Networks, (Preprint), 1- 22.

[4] Arshi, M., Nasreen, MD., & Karanam Madhavi. (2020). A Survey of DDOS Attacks Using Machine Learning Techniques. E3S Web of Conferences, 184, 01052. DOI:10.1051/e3sconf/202018401052

[5] Shahraki, A., Abbasi, M., & Haugen, Ø. (2020). Boosting algorithms for network intrusion detection: A comparative evaluation of Real AdaBoost, Gentle AdaBoost and Modest AdaBoost. Engineering Applications of Artificial Intelligence,94,103770. https://doi.org/10.1016/J.ENGAPPAI.2020.103770I

[6] Sharma, K., & Gupta, B., B. (2018). Taxonomy of Distributed Denial of Service (Ddos) Attacks and Defense Mechanisms in Present Era of Smartphone Devices. International Journal of E-Services and Mobile Applications, 10, 2, 58–74. DOI: 10.4018/ijesma.2018040104.

[7]Ghafar A Jaafar., Shahidan M Abdullah., & Saifuladli Ismail. (2019). Review of Recent Detection Methods for HTTP DDoS Attack. Journal of Computer Networks and Communications, 2019, Article ID 1283472, 10. DOI:10.1155/2019/12834727

[8] Batchu, R. K., & Seetha, H. (2021). A generalized machine learning model for DDoS attacks detection using hybrid feature selection and hyperparameter tuning. ComputerNetworks,200,108498. https://doi.org/10.1016/J.COMNET.2021.108498

[9] Dasari, K.B., Devarakonda, N. (2021). Detection of different DDoS attacks using machine learning classification algorithms. Ingénierie des Systèmes d'Information, Vol. 26, No. 5, pp. 461-468. http://dx.doi.org/10.18280/isi.260505