Real Time DDos Attack Detection System Using Apache Flink and Gradient Boosting Algorithm

Author: Rupal D. More, Kartik Shikhare, Rahulkumar Choudhary, Akanksha Mate, Sanket Gawali.

Department of Computer Engineering

Sandip Institute of Technology & Research Center®, Nashik, India

Abstract

When large amounts of traffic from hundreds, thousands, or even millions of other computers are routed to a network or server to crash the system and disrupt its function to make the system crash.A DDoS attack aims to overwhelm the devices, services, and network of its intended target with fake internet traffic, rendering them inaccessible to or useless for legitimate users.DDoS attacks are designed to look like a flood of calls, or requests, made by browsers asking a web page to load.Ddos attacks are very crucial as they make the website or webapp down which leads to losses to the company.Early DDoS detection is critical for businesses because it can help protect the functioning and security of a network.Networks without a robust DDoS defense strategy may have trouble defending against the wide range of DDoS attacks, which can be difficult to trace.The problems need to be addressed with models that can manage the time information contained in network traffic flows.We can detect the attack while the initial requests are being made to the server and block the requests made by such IP addresses or if they are false dods attack warnings we can scale our app to handle the traffic.The system to be proposed should be as real time as it can be as it will help in preventing system crashes. In this System we will be using apache flink to give real time detection and for better classification of attack the gradient boosting is proposed.

Keywords: Distributed Denial of Service attack detection; machine learning; network security .

Literature Survey

As in numerous other areas in network security, several different deep learning algorithms continue to be used tirelessly to establish secure Internet communication between devices. Especially in recent years, early detection and response to cyber-attacks have become very important. Due to the pandemic, the smooth operation of servers related to the provision of services and products via the Internet is more critical than ever. This has made such servers even bigger targets for attacks, which has led to the publication of various related tasks in a short period.

Barati et al. [13] suggested a DDoS assault detection system framework. In a hybrid technique, a Genetic Algorithm and an Artificial Neural Network were used for characteristic identification and threat detection, respectively. The most efficient features are chosen using a layering technique based on GA, and the recognition rate of DDoS attacks was increased using ANN's Multilayer Perceptron (MLP). The findings showed that the suggested approach could identify DDoS attacks with excellent precision and a low chance of false alarm. They intended to use similar themes to conduct further tests on other data sets to assess the experiment's resilience.

Hosseini and Azizi [14] provided a method for identifying DDoS attacks using gradual training depending on a data stream methodology. To quickly organize the activity, they devised a method that allocated the computation responsibility between the client and proxy components based on the resources available to

each of those portions. The consumer side had three stages: first, the client service's data collection; second, component recovery based on forwarding classification for each method; and finally, the differentiation test. As a result, if the deviation exceeded a certain threshold, the assault was detected; otherwise, data were sent to the intermediary side. On the proxy side, they employed the nave Bayes, random forest, decision tree, multilayer perceptron (MLP), and k-nearest neighbors (K-NN) to get superior results. Distinct assaults have diverse tendencies, and the necessary efficiency for identifying assaults and more capacity to differentiate novel threat patterns is obtained thanks to different chosen characteristics for each method. The findings suggest that the random forest method outperforms the other techniques.

Shurman et al. [1] offered two approaches for detecting Distributed Reflection Denial of Service (DrDoS) assaults on the Internet of Things. The first way is a hybrid-based IDS for IoT networks, which involves providing an IDS framework scheme specified as an application capable of detecting abnormal data traffic from any network node and running IP datasets against it. It was able to detect strange IP packets and ban unaccepted IPs before they escalated into possible DoS threats. The second technique used a deep training system based on LSTM that was trained on the CICDDoS2019 dataset containing different types of DrDoS assaults and was able to identify them. Their findings showed that the suggested approaches could identify malicious behavior, ensuring the safety of the IoT network. They wanted to create a new deep learning model to identify the second kind of DDoS assault in the CICDDoS2019 dataset and evaluate the performance of these approaches in a real-world system.

To correctly forecast DDoS assaults utilizing benchmark data, Alghazzawi et al. [15] recommended employing a hybrid deep training (DL) model, specifically a CNN with BiLSTM (bidirectional long memory). Only the most essential characteristics were chosen by rating and selecting the features that rated the best in the supplied data set. The suggested CNN-BI-LSTM achieved an efficiency of up to 94.52 percent utilizing the data set CIC-DDoS2019 throughout training, testing, and validation, according to the results of the experiments. Using a unique data set, a single statistical approach, the chi-squared test to select relevant characteristics, and the utilization of hidden states instead of a pretrained CNN model were all limitations of their system. They intended to evaluate the usage of multiple traffic data sets and alternative feature selection techniques to the chi-squared evaluation and pretrained word encoding algorithms such as autoencoders, Glove, and Fasttext.

In a Software Defined Network setting, Deepa et al. [16] suggested a hybrid deep learning approach to identify DDoS attacks. They've also evaluated their work using three performance criteria: accuracy, precision, and false alert rate. In contrast to the SVM method, SOM is an uncontrolled machine learning algorithm that performs well in detecting assaults. However, when compared to a basic machine learning model, they obtained higher accuracy, detection rate, and reduced false alarm rate utilizing their suggested hybrid machine learning model. By enforcing security restrictions in the flow table, they planned to develop ensemble deep learning models to identify DDoS attacks in the data plane.

Zhang et al. proposed a framework to detect intrusion using a Random Forest algorithm using big data Apache Spark. Random Forest algorithm was implemented in Apache Spark to detect intrusion in the high-speed network data, and their efficiency and accuracy were then evaluated and compared to existing systems. An issue with this approach is that they evaluated few classification algorithms in their research. They implemented an RF algorithm in Apache Spark's design and evaluated their results compared with other fast speed data network models. Their proposed model achieved higher accuracy and can detect intrusion in a shorter amount of time. The results showed that their model could detect intrusion in 0.01 s compared with other existing models. A simple RF model can detect an attack in 1.10 s.

Zekri et al. focused on how Distributed Denial of Service affects cloud performance by utilizing network resources. The attack techniques implemented and evaluated different ML algorithms in a cloud computing environment. The authors presented a DDoS protection design, and the algorithms they implemented and evaluated in cloud environments were Naive Bayes, Decision Tree (C4.5), and K-Means (KM). The drawback of this approach is that they evaluated only a few models to detect DoS attacks.

Li and Yan proposed IDS based on Apache Spark using "MSMOTR" and Adaboost models. The experiment was performed on the KDD99 dataset. The results showed that the proposed approach could reduce the system's error rate and processing time and improve the accuracy rate. The results showed that the traditional model achieved accuracy, error rate, and processing time of 84.28, 17.2, and 18.26 s. The model achieved accuracy, error rate, and processing time of 86.32, 13.6, and 15.24 s, respectively.

Priya et al..proposed an ML-based model for the detection of DDoS attacks. The authors applied three different machine learning models: K-Nearest Neighbors (KNN), Random Forest (RF) and Naive Bayes (NB) classifiers. The proposed approach can detect any type of DDoS attack in the network. The results of the proposed approach showed that the model can detect attacks with an average accuracy of 98.5%.

Mazhar Javed Awan proposed a system with apache spark for some of the big data tools and for classification Random forest was used . The system was efficient but the technology used such as apache spark and random forest can be made efficient with apache flink and gradient boosting.

Li Xinlong1 and Chen Zhibin proposed a system with hybrid deep learning methodologies With RNN and LSTM model trained on Cerebral cortex known as neocortex and The proposed system used HTM. a. Experiments showed that the proposed methodology successfully resolved the issue of accurately detecting DDoS attacks.

Hosseini and Azizi [14] provided a method for identifying DDoS attacks using gradual training depending on a data stream methodology. a method that allocated the computation responsibility between the client and proxy components based on the resources available to each of those portions.

The consumer side had three stages: first, the client service's data collection; second, component recovery based on forwarding classification for each method; and finally, the differentiation test. As a result, if the deviation exceeded a certain threshold, the assault was detected; otherwise, data were sent to the intermediary side.proxy side,

They employed the naive Bayes, random forest, decision tree, multilayer perceptron (MLP), and k-nearest neighbors (KNN) to get superior results. The findings suggest that the random forest method outperforms the other techniques.

Panpan Qi, Wei Wang, Lei Zhu and See Kiong Ng propose a new weighting scheme integrated into GBDT for sampling instances in each boosting round to reduce the negative impact of wrongly labeled target instances. Experiments on two large malware datasets demonstrate the superiority of our proposed method.

Sriraam Natarajan, Saket Joshi, Prasad Tadepalli, Kristian Kersting and Jude Shavlik implemented a functional gradient boosting approach to imitation learning in relational domains. In particular, given a set of traces from the human teacher, the system learns a policy in the form of a set of relational regression trees that additively approximate the functional gradients. The use of multiple additive trees combined with relational representation allows for learning more expressive policies than what has been done before. Demonstrate the usefulness of our approach in several different domains.

Tianqi Chen, Sameer Singh, Ben Taskar, Carlos Guestrin incorporated second-order information by deriving a Markov Chain mixing rate bound to quantify the dependencies, and introduce a gradient boosting algorithm that iteratively optimizes an adaptive upper bound of the objective function. The resulting algorithm induces and selects features for CRFs via functional space optimization, with provable convergence guarantees. Experimental results on three real world datasets demonstrate that the mixing rate based upper bound is effective for learning CRFs with non-linear potentials.

Jeany Son, Ilchae Jung, Kayoung Park, and Bohyung Han proposed an online tracking algorithm that adaptively models target appearances based on an online gradient boosting decision tree. The algorithm is particularly useful for non-rigid and/or articulated objects since it handles various deformations of the target effectively by integrating a classifier operating on individual patches and provides segmentation masks of the target as final results.

Alina Beygelzimer, Elad Hazan, Satyen Kale and Haipeng Luo extend the theory of boosting for regression problems to the online learning setting. Generalizing from the batch setting for boosting, the notion of a weak learning algorithm is modeled as an online learning algorithm with linear loss functions that competes with a base class of regression functions, while a strong learning algorithm is an online learning algorithm with convex loss functions that competes with a larger class of regression functions.

Chen Cheng, Fen Xia, Tong Zhang, Irwin King, and Michael R. Lyu proposed a greedy interaction feature selection algorithm based on gradient boosting. A novel Gradient Boosting Factorization Machine (GBFM) model to incorporate feature selection algorithms with Factorization Machines into a unified framework. The experimental results on both synthetic and real datasets demonstrate the efficiency and effectiveness of algorithms compared to other state-of-the-art methods.

I. Introduction

When large amounts of traffic from hundreds, thousands, or even millions of other computers are routed to a network or server to crash the system and disrupt its function to make the system crash.A DDoS attack aims to overwhelm the devices, services, and network of its intended target with fake internet traffic, rendering them inaccessible to or useless for legitimate users.DDoS attacks are designed to look like a flood of calls, or requests, made by browsers asking a web page to load.

Ddos attacks are very crucial as they make the website or webapp down which leads to losses to the company.Early DDoS detection is critical for businesses because it can help protect the functioning and security of a network.Networks without a robust DDoS defense strategy may have trouble defending against the wide range of DDoS attacks, which can be difficult to trace.The problems need to be

addressed with models that can manage the time information contained in network traffic flows.

We can detect the attack while the initial requests are being made to the server and block the requests made by such IP addresses or if they are false dods attack warnings we can scale our app to handle the traffic.The system to be proposed should be as real time as it can be as it will help in preventing system crashes. In this System we will be using apache flink to give real time detection and for better classification of attack the gradient boosting is proposed

II. FUNDAMENTALS AND BASIC TERMINOLOGIES USED

A. *DDoS:*

Distributed denial-of-service (DDoS) attack is a malicious attempt to disrupt the normal traffic of a targeted server, service or network by overwhelming the target or its surrounding infrastructure with a flood of Internet traffic. DDoS attacks achieve effectiveness by utilizing multiple compromised computer systems as sources of attack traffic. Exploited machines can include computers and other networked resources such as IoT devices. (Li Xinlong, n.d., [1])

B. *Gradient Boosting:*

Gradient boosting is a machine learning technique used in regression and classification tasks, among others. It gives a prediction model in the form of an ensemble of weak prediction models,which are typically decision trees. (Panpan Qi, n.d., [25]

III. Experimental Analysis

This implementation attempted to detect the DDoS attack. The dataset is massive and complicated, and testing it with a standard processor is nearly difficult due to the vast number of training occurrences. The big dataset necessitates the use of high-performance computers. There is no duplicate records in the proposed test sets. The number of records in the train and test sets are reasonable, which makes it affordable to run the experiments on the complete set without the need to randomly select a small portion. Consequently, evaluation results of different research works will be consistent and comparable. Various parameter values were utilized and tested to achieve varying degrees of accuracy.

A. Data set

The NSL-KDD is a security researcher-popular implementation of the KDD detecting DDoS attack dataset. The data set in this implementation provides 38 distinct sorts of assaults that are aggregated into four fundamental attack classes to give a more visible representation of outcomes. There are attack classes categorized in four parts: DoS (Denial of service), Probing (Surveillance and other probing attacks), U2R (Unauthorized access to local super user), R2L (Unauthorized access from a remote machine). DoS attacks vary from other kinds of cyberattacks including that they take down a resource, whilst others penetrate a network or system.

B. Data Preprocessing

There are 41 features in the KDD 99 Dataset. The model requires many character values, which are continuous values of floating-point numbers. Noise filtering is used to reduce noise from data collected on development then missing values are handled utilizing various policies, such as ignoring data with missing entries, replacing data with a universal, consistent method. Normalization is carried out by using typical scaler. Non-numeric categorical data is used to extract the derived features. Used the sklearn Label Encoder for training purpose to encode the variables

IV. METHODOLOGY & DIAGRAM

The block diagram for DDoS detection is shown in Figure1. In pre-processing the data first, noise filtering is performed on the dataset. Noise filtering is a collection of procedures used to reduce noise from data collected on development then missing values are handled utilizing various policies, such as ignoring data with missing entries, replacing data with a universal, consistent method and explicitly filling in missing attributes depending on your area of expertise and at last we reduced the number of features.

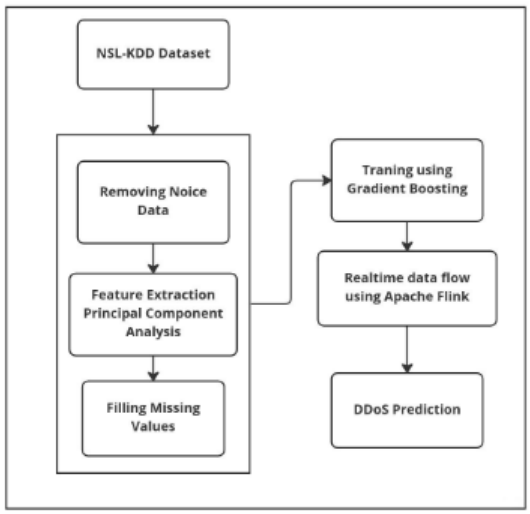

Fig.1. Block diagram of DDoS detection.

Gradient Boosting algorithm for training the data, it relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error. We used Scikit libraries for the modeling purpose and applied Apache flink libraries Flink ML for the training and testing of the dataset. Finally, in the last step we measured the evaluation matrices of all the approaches and compared the results.

V. RESULTS AND DISCUSSIONS

To increase the precision of the model further, the results are processed and examined.

A subset of the KDD data set was used in the analysis to train the models. Data from the test split in KDD was randomly chosen for the testing. The models were initially trained on all characteristics, and then the quantity of features chosen was changed to obtain various degrees of accuracy.

Gradient boosting a machine learning boosting is used as it relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

Where Fm is some imperfect model, in order to improve Fm there is soe new estimator hm, γ_m is the step length and input 'x'.

A. Advantages

To provide legitimate users seamless and uninterrupted access to websites.working of the websites.Prevents the attacker from sending more requests than the victim server can handle.Identify and Block the IP address of the attacker after activity monitoring.

B. Limitations

If the type of ddos attack is not being considered during the creation of the model and algorithm may lead to inaccurate results .Due to limitations in high computational power the detection is limited to a few web apps at the same time.

C. Applications

For security purposes In services like AWS Shield Effective attack mitigation Used by

Companies for making their services available to legitimate users. Can be used in Google Cloud Armor for more secure and advanced ddos attack detection.

VI. FUTURE SCOPE & CONCLUSION

Traditional intrusion detection techniques can only work best on slow-speed data or small data. Still, they are inefficient on big data and are incapable of handling high-speed data, so new methods adapted to work on large data to detect any signs of intrusion are needed. In this paper, we predicted DDoS attacks in real-time with different machine learning models using a big data approach. We used a distributed system, Apache Flink, and a classification algorithm to enhance the algorithms' execution. Apache Flink is a big data tool to detect an attack in real-time with Apache Flink ML libraries. We applied the machine learning approach of Gradient Boosting (GB) through the Scikit ML library and big data framework Flink-ML library for the detection of DoS attacks. In addition to the detection of DoS attacks we have optimized the performance of the models by minimizing the prediction time as compared with other existing approaches using big data framework.

In the future, we will train different models and combine them with deep learning approaches sing neural networks for predicting real-time results from convolutional neural network architectures.

REFERENCES

[1]. Li Xinlong1 and Chen Zhibin 2 , "DDoS Attack Detection by Hybrid Deep Learning Methodologies " Li Xinlong1 and Chen Zhibin 2 .2022.

[2]. Privalov, A.; Lukicheva, V.; Kotenko, I.; Saenko, I. Method of Early Detection of Cyber-Attacks on Telecommunication Networks Based on Traffic Analysis by Extreme Filtering. Energies **2019**, 12, 4768.

[3]. Mubashar, R.; Awan, M.J.; Ahsan, M.; Yasin, A.; Singh, V.P. Efficient Residential Load Forecasting using Deep Learning Approach. Int. J. Comput. Appl. Technol. 2021, in press.

[4] . Mazhar Javed Awan 1,* , Umar Farooq 1 , Hafiz Muhammad Aqeel Babar 1 , Awais Yasin 2 , Haitham Nobanee 3,4,5,* , Muzammil Hussain 6 , Owais Hakeem 6 and Azlan Mohd Zain, 'Real-Time DDoS Attack Detection System Using Big Data Approach 2022 [5]Abdullah, A.; Awan, M.; Shehzad, M.; Ashraf, M. Fake News Classification Bimodal using Convolutional Neural Network and Long Short-Term Memory. Int. J. Emerg .Technol.Learn.2020,11, 209–212.

[6]Ganguly, S.; Garofalakis, M.; Rastogi, R.; Sabnani, K. Streaming algorithms for robust, real-time detection of ddos attacks. InProceedings of the 27th International Conference on Distributed Computing Systems (ICDCS'07), Toronto, ON, Canada, 25–27 June 2007; p. 4.

[7] .Salloum, S.; Dautov, R.; Chen, X.; Peng, P.X.; Huang, J.Z. Big data analytics on Apache Spark. Int. J. Data Sci. Anal. 2016, 1, 145–164.

[8] Alexey Natekin1 * and Alois Knoll 2 'Gradient boosting machines, a tutorial' METHODS article Front. Neurorobot., 04 December 2013

[9] . Syed, N.F.; Baig, Z.; Ibrahim, A.; Valli, C. Denial of service attack detection through machine learning for the IoT. J. Inf. Telecommun. 2020, 4, 482–503.

[10] .Hu, T., Li, X., and Zhao, Y. (2007). "Gradient boosting learning of Hidden Markov models," in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'06) (Toulouse). doi: 10.1109/ICASSP.2006.1660233

[11] . aved Awan, M.; Shafry Mohd Rahim, M.; Nobanee, H.; Munawar, A.; Yasin, A.; Mohd Zain Azlanmz, A. Social Media and Stock Market Prediction: A Big Data Approach. Comput. Mater. Contin. 2021, 67, 2569–2583, doi:10.32604/cmc.2021.014253

[12] .Saravanan, S. Performance evaluation of classification algorithms in the design of Apache Spark based intrusion detection system. In

Proceedings of the 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 10–12 June 2020; pp. 443–447.
[13] Kato, K.; Klyuev, V. Development of a network intrusion detection system using Apache Hadoop and Spark. In Proceedings of the 2017 IEEE Conference on Dependable and Secure Computing, Taipei, Taiwan, 7–10 August 2017; pp. 416–423.
[14]. Zekri, M.; El Kafhali, S.; Aboutabit, N.; Saadi, Y. DDoS attack detection using machine learning techniques in cloud computing environments. In Proceedings of the 2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech), Rabat, Morocco, 24–26 October 2017; pp. 1–7.
[15] . Zhang, H.; Dai, S.; Li, Y.; Zhang, W. Real-time distributed-random-forest-based network intrusion detection system using Apache spark. In Proceedings of the 2018 IEEE 37th International Performance Computing and Communications Conference (IPCCC), Orlando, FL, USA, 17–19 November 2018; pp. 1–7.
[16] .Friedman, J. (2001). Greedy boosting approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232. doi: 10.1214/aos/1013203451
[17]Polat, H.; Polat, O.; Cetin, A. Detecting DDoS Attacks in Software-Defined Networks Through Feature Selection Methods and Machine Learning Models. Sustainability 2020, 12, 1035.
[18]M. Shurman, R. Khrais, and A. Yateem, "DoS and DDoS attack detection using deep learning and IDS," 0e International Arab Journal of Information Technology, vol. 17, no. 4A, pp. 655–661, 2020.
[19]B. Nugraha, N. Kulkarni, and A. Gopikrishnan, "Detecting adversarial DDoS attacks in software- defined networking using deep learning techniques and adversarial training," 2021 IEEE International Conference on Cyber Security and Resilience (CSR), in Proceedings of the 2021 IEEE International Conference on Cyber Security and Resilience (CSR), pp. 448– 454, Rhodes, Greece, July. 2021.
[20]L. Yao and Y. Guan, "An improved LSTM structure for natural language processing," 2018 IEEE International Conference of Safety Produce Informatization (IICSPI), in Proceedings of the 2018 IEEE International Conference of Safety Produce Informatization (IICSPI), pp. 565–569, Chongqing,China, December 2018.
[21]D. Alghazzawi, O. Bamasag, H. Ullah, and M. Z. Asghar,"Efficient detection of DDoS attacks using a hybrid deep learning model with improved feature selection," Applied Sciences, vol. 11, no. 24, p. 11634, Dec. 2021.
[22]Tilmann Rabl*, Jonas Traub, and Volker Markl , 'Apache Flink in Current Research Projects'
304536933 Researcher.net
[23] .N. Deshai, B.V.D.S. Sekhar, S. Venkataramana , 'Processing Big Data with Apache Flink',
International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-1S3, June 2019.
[24] . Dr. Yusuf Perwej , 'A Comprehend The Apache Flink In Big Data Environments', 2018, IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661, P-ISSN: 2278-8727
[25]. Panpan Qi, Wei Wang, Lei Zhu, and See Kiong Ng. 2021. Unsupervised Domain Adaptation for Static Malware Detection based on Gradient Boosting Trees. In Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia. ACM, New York, NY, USA, 10 pages. M. Shurman, R. Khrais, and A. Yateem, "DoS and DDoS attack detection using deep learning and IDS," *The International Arab Journal of Information Technology*, vol. 17, no. 4A, pp. 655–661, 2020.