Sixth Int. Conf. on Computer Systems, Data Communication and Security, CDCS

Machine Learning based Ensemble Technique for DDoS Attack Detection in Software-Defined Networking

C. Srinivas[1], Dr. P S Avadhani[2] and Dr. P. Prapoona Roja[3]
[1]Research Scholar, JNTU-K-Kakinada
Email: csrinivas@gvpcew.ac.in
[2]Professor, Computer Science and Systems Engineering, Andhra University
Email: psavadhani@yahoo.com
[3]Professor, Dept. of Computer Science and Engineering, GVP College of Engineering (A), Visakhapatnam, AP, INDIA
Email: ap_roja@gvpce.ac.in

***Abstract*—Next-generation networks can benefit from a more dynamic and successfully controlled network design because of a new network paradigm termed the Software-Defined Network (SDN). Network administrators may simply monitor and manage the entire network using the design of the customizable centralized controller. A number of attack vectors simultaneously target it because of its centralized nature. DDoS attacks are the most efficient type of attack against the SDN. The goal of this work is to classify SDN flow as either normal or assault traffic using ML techniques. We manage a public "DDoS attack SDN Dataset" with 23 characteristics in total. The dataset comprises both legitimate and malicious traffic for the TCP, UDP, and ICMP (TCP). The dataset, which includes over 100,000 recordings, offers statistical statistics such byte count, time sec, packet rate, and packet per flow, with the exclusion of characteristics that define source and target devices. In this paper DDoS attack was detected using Various ML Algorithms such as K-Nearest Neighbor (KNN), Decision Tree (DT), Support Vector Machine (SVM) and Random Forest (RF) algorithms The experimental results demonstrate that an Ensemble Random Forest algorithm was given 99.99% classification accuracy than the other methods.**

***Index Terms*— SDN, Distributed Denial of Service attacks, machine learning.**

I. Introduction

SDN is a modern paradigm whose dynamic and programmable structure makes network administration easier SDN separates the control and data planes, and the network is overseen by a central controller.

As a result, the controller, which has the ability to administer the whole network from a single location, may easily implement various network regulations across the board. However, in addition to the benefits it offers, this growing new technique also has security issues. SDN is vulnerable to attacks unique to itself in addition to threats seen in conventional network topologies. Attacks against the controller are perhaps the riskiest of these since the attacker in control of the controller may be able to manipulate or interrupt all network traffic.

The majority of attacks on the controller are DDoS assaults, which prevent users from accessing N/W benefits. The intruders intend to create a significant amount of traffic.using several machines, reduce the resources of the

Grenze ID: 02.CDCS.2023.6. 514

target machine, and eventually stop it from functioning by using DDoS assaults. Attackers make advantage of "botnets" made up of zombie devices that have been hacked online. DDoS assaults use several devices, making it incredibly challenging to identify and stop them. DDoS assaults are becoming more frequent and more severe, and they have the potential to completely destroy many network services.

For network service providers and administrators, one of the most pressing issues is the timely detection and prevention of DDoS assaults Even though SDN's central administration and programmable architecture provide IDSs additional possibilities,The efficiency of these detecting systems is reliant on the calibre of training datasets. Different datasets, including KDD Cup'99, NSL-KDD, CICIDS 2017, CAIDA 2016, UNB-ISCX, and CIC DoS, were employed in recent research we mentioned above. The fact that these datasets are outdated is their main flaw. The demand for current datasets is growing as attack characteristics change. The datasets currently utilised to identify DDoS assaults include LITNET-2020 and Boaziçi University databases.

However, much like the other datasets, these datasets were also developed utilising conventional network platforms. We were motivated to conduct this study because we wanted to use the most recent datasets available from SDN network systems. For anomaly detection systems used in SDN networks, there are a few publicly accessible datasets that may be utilised right away. In our work, we made use of the "DDOS attack SDN Dataset," a brand-new dataset that is available for use in machine learning by academics.

II. Literature Survey

A method for identifying distributed denial of service (DDoS) assaults was proposed by Deepa, V., K. Muthamil Sudar, and P. Deepalakshmi [1].They implemented four unique ML models to identify suspicious traffic in the Sdn network. The SVM-SOM approach fared better than the other ML algorithms, with an accuracy rate of 98.12%.

In [2], the authors presented a Method for detecting DDoS attacks on SDN. The system employed 2 levels of privacy. In order to identify signature based assaults, they used Snort at initially. The DNN, ML and the SVM classifier were then used to characterise assaults. According to the experimental findings, The classification accuracy percentage for DNN is 92.30 percent, which is higher than SVM's.

In their research, the authors of [3] showed that the DDoS attacks were successfully identified and categorised using the DL model. On two separate samples selected from the CICDDoS2019 dataset,the DNN model was used. The attack detection scenario was applied to the first dataset, while the attack traffic categorization scenario was applied to the second dataset. A DDoS protection system leveraging the SDN architecture to identify attack flows was proposed by Nam, Tran Manh, and colleagues [4]. Algorithms from KNN and SOM are incorporated in their hybrid solution. Using flow statistics gathered from SDN switches and car sensors, they divided the traffic into legitimate and malicious traffic. Adhikary et al. [5] concentrated on a hybrid approach that combines DT and neural network techniques to counteract various DDoS assaults in vehicular ad hoc networks (VANET).

Results from the proposed hybrid algorithm are superior than those from the DT and neural network standalone models. To identify and stop the DDoS assault, Hosseini and Azizi [6] presented a hybrid methodology. In their structure, the parties were divided into proxy and client.To find the attack flows, they integrated six distinct machine learning algorithms. In comparison to the other ML approaches, the Random Forest classifier produces superior results. A security technique was put out by Ravi, Nagarathna, and S. Mercy Shalinie [8] to identify and mitigate DDoS assaults in Internet of Things networks. Their system, Learning-driven Detection Mitigation (LEDEM), employed a semi-supervised machine learning model to identify fraudulent communications.

III. Methodology

Figure 1 illustrates the process that was used to get the desired outcomes. The initial action is seen in Fig. is to import the SDN dataset and perform Data Cleaning. We then partition the dataset into training, testing, and validation dataset. After then, several Machine Learning methods are employed training of the model on the train dataset. The model's performance is predicted using the training metrics. If the performance is poor on the training set or validation set, then we go back to the training step to adjust the models's parameters. Then we make use of the test dataset to validate the model and finally predict the results and give the analysis.

A. Dataset

This work utilises the publicly accessible "DDOS attack SDN Dataset," which was resulted from the SDN architecture and made available to scientists to be use in deep learning and machine learning research. The dataset has 104345 traffic flows, and there are 23 characteristics in it. The classes normal and attack traffic are

TABLE I. THE COMPARISON OF THE RELATED STUDIES

Related Studies and Datasets	Feature Selection	ML Algorithms	Accuracy (%)
CIC DoS dataset.	Without feature selection	Random Tree, J48, REP Tree, SVM, Random Forest, MLP	95.00
NSL-KDD.	Without feature selection	K-Means and kNN	98.85
Their dataset.	Without feature selection	Stacked Autoencoders (SAE) deep learning model	95.00
UNB-ISCX.	Without feature selection	Semi supervised machine-learning algorithm	96.28
Their dataset.	Without feature selection	Polynomial SVM- Linear SVM	95.00
CICIDS2017.	Without feature selection	CNN	98.98
Their dataset.	Without feature selection	ALM	97.00
CAIDA 2016.	Without feature selection	KNN, Naive Bayes, SVM, and SOM	98.12
KDD Cup'99.	Without feature selection	SVM classifier and DNN	92.30
CAIDA "DDoS Attack 2007".	Entropy-based selection	SOM+KNN, SOM distributed-centre	98.24
Their dataset.	Without feature selection	Hybrid algorithm of DT and Neural Network	96.40
NSL-KDD, the introduced dataset in	KNIME forward feature selection	Random Forest, Naive Bayes, DT, kNN, MLP	98.63

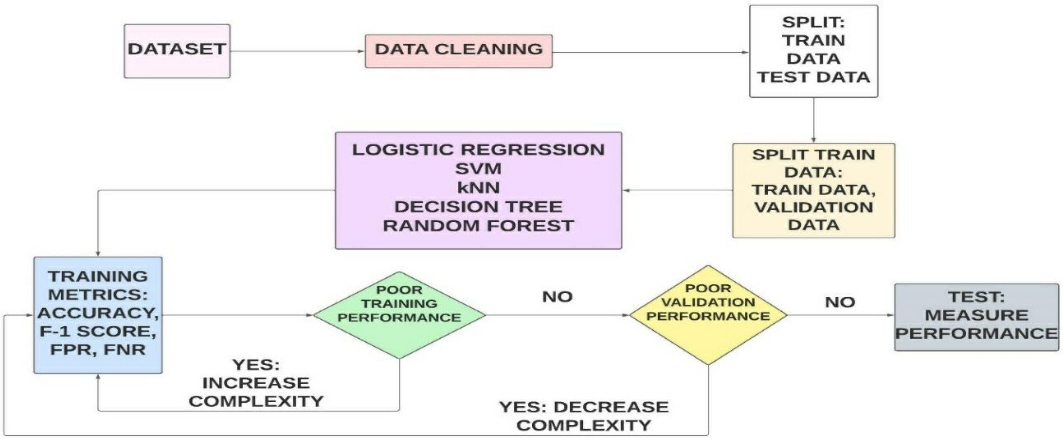
DATASET
DATA CLEANING
SPLIT: TRAIN DATA TEST DATA
SPLIT TRAIN DATA: TRAIN DATA, VALIDATION DATA
LOGISTIC REGRESSION SVM kNN DECISION TREE RANDOM FOREST
TRAINING METRICS: ACCURACY, F-1 SCORE, FPR, FNR
POOR TRAINING PERFORMANCE
NO
POOR VALIDATION PERFORMANCE
NO
TEST: MEASURE PERFORMANCE
YES: INCREASE COMPLEXITY
YES: DECREASE COMPLEXITY

Fig.1 Proposed System

used to show the dataset, which consists of TCP, UDP, and ICMP traffic. The dataset includes statistical characteristics such as byte count, time sec, packet rate, and packet per flow, with the exception of those that identify the source and target computers. Before starting the machine learning model training operation, the data must be pre-processed. Currently, the dataset does not include the attributes of packet rate, byte per-flow, or packet per flow since they have duplicate values.

B. Dataset Partitioning

The dataset is segmented into testing and training sets for each experiment. More precisely, 70% of the dataset was used for model training, and 30% was used for testing

C. Machine Learning Algorithms

Logistic Regression

For categorization and predictive analytics, this kind of statistical model is frequently employed. Data must be pre-processed before beginning the machine learning model training activity... In logistic regression, we fit a "S" shaped logistic function instead of a regression line, which predicts two maximum values (0 or 1).

KNN

K-NN is one of the most well-known machine learning algorithms. It was first used in 1951 and is a non-parametric, distance-based, and supervised method. This algorithm calculates the dataset's similarities while taking a distance function into account. Depends on the democratic most of its k-nearest neighbours, the test data are categorised.

Decision Tree

The regression and categorization of real-world issues are both done using the decision tree machine learning technique. The tree structure is the model's source of inspiration. The tree's base, however, is discovered at the top. The decision tree is moderately formed, and the branches are made in consideration of the dataset's properties and objective rules. Following the steps outlined below will enable you to design a decision tree.

1. The whole dataset has been segmented into train and test sets.
2. The train data set is utilised as input at the tree's root.
3. The information theory as explained is used to find the root.
4. The prone technique is used.
5. Continue repeating steps 1 through 4 until enough nodes have turned into leaf nodes.

Support Vector Machine

One among the most powerful machine learning methods for classification and regression issues is the SVM algorithm. A hyper plane which may divide the space into two or more distinct classes is determined using SVM. Data points within this boundary are called support vectors, and the boundary is kept as wide as possible.

Random Forest

Supervised machine learning algorithms like random forests are commonly used for regression and classification problems. Use different samples to build a decision tree, use their mean for classification, and majority vote for regression. One of the most important properties of random forest algorithms is their ability to process data sets with both continuous variables, such as regression, and categorical data, such as classification. When it comes to classification problems, it gives excellent results.

The random forest method involves the following steps:

Step 1: In Random Forest, n randomly chosen records are selected from a data collection of k number of records.
Step 2: For each sample, a different decision tree is built.
Step 3: The output that each decision tree produces.
Step 4: For classification and regression, the final product is evaluated using the majority vote or an average.

Evaluation Metrics

The proposed method's performance is evaluated using the following metrics: accuracy, precision, sensitivity (recall), specificity, and F1 score. These are the formulae used to calculate these metrics

$$\text{Accuracy} = (TN + TP)/TS$$
$$\text{Precision} = TP/\ TP + FP$$
$$\text{Sensitivity (recall)} = TP\ /TP + FN$$
$$\text{Specificity} = TN\ /TN + FP$$
$$\text{F1\_score} = 2 * \text{Precision} \times \text{Recall}\ /\text{Precision} + \text{Recall}$$

True positive, total samples, false positive, true negative, and false negative are denoted by the letters TP, TS, FP, TN, and FN.

IV. Results and Discussion

In the first part of the pilot study, SDN data were immediately classified by a machine learning algorithm without preprocessing and feature selection. For successful classification, hyperparameter optimization methods were used to automatically compute the hyperparameters of the machine learning algorithm. The sample dataset was split into training at a rate of 0.7, and testing at a rate of 0.3. To use the KNN method for classification, The distance function was chosen as Euclidean, and the value of k, which is the number of neighbors to be checked, was set at 1. The Gini algorithm computes the division criterion in the Decision Tree approach.

Performance Measurement

So when the results obtained have been analysed, after the SDN records were given into the machine learning algorithms, the Random Forest method had the highest accuracy rate of 99.99%, while the Logistic Regression, SVM, Decision Tree, and KNN methods had accuracy rates of 76.64%, 97%, 98.22%, and 98%, respectively. Table 2 shows the additional metrics. Based on the data, we may conclude that the KNN algorithm outperforms all other performance metrics. Figure 4 depicts a graphical depiction of the table below. Table2 Performance measurement of different algorithms.

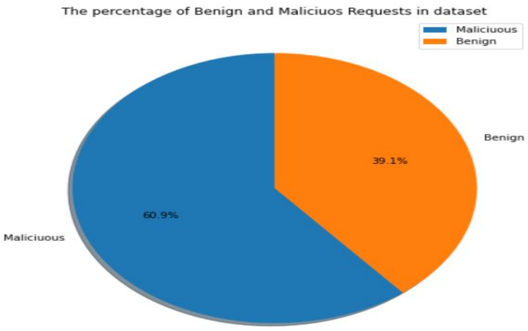

Fig.2 Types of Requests in Dataset

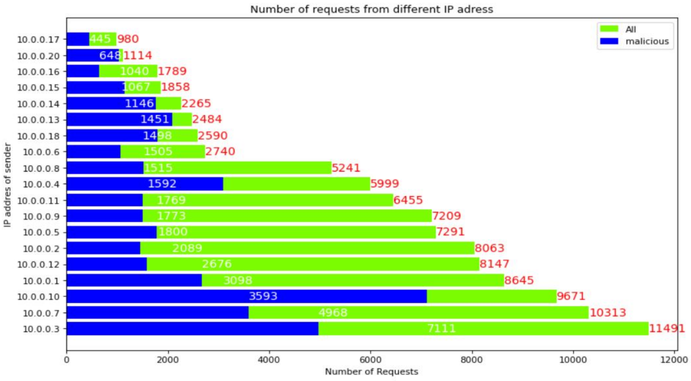

Fig.3 Number of Requests from Different IP address

ML algorithms	**Accuracy(%)**	**Precision(%)**	**Recall(%)**	**F1-score(%)**
Logistic Regression	76.64	75	76	75
SVM	97	97	96	97
Decision Tree	98.22	98	98	98
Random Forest	99.99	100	100	100
kNN	98	98	98	98

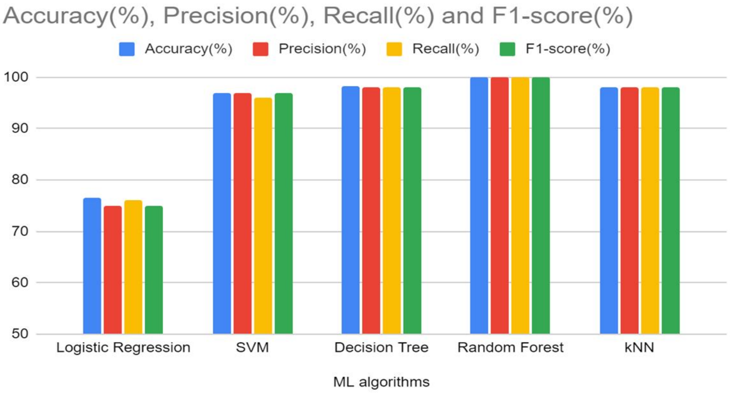

Fig.4 Performance Evaluation

V. CONCLUSION

In this work, the dataset received from the SDN environment was categorized into normal and attack traffic using ML techniques. The specialized SDN-based dataset includes of both legitimate and malicious TCP, UDP, and ICMP traffic. With the exception of attributes that identify the source and target machines, the dataset contains statistical characteristics such as byte count, time sec, packet rate, and packet per flow. 22 network characteristics are examined, and they are then used as data for machine learning techniques. After pre-processing, more than 100,000 network records were categorized using the Logistic Regression, KNN, DT, RF, and SVM algorithms. According to the experimental findings, RF has a 99.99% accuracy rate, which is higher than that of the other algorithms. Future studies will broaden the number of attacks are used to examine the classification results of machine learning models using feature selection approaches.

REFERENCES

[1] Deepa, V.; Sudar, K.M.; Deepalakshmi, P. Design of Ensemble Learning Methods for DDoS Detection in SDN Environment. In Proceedings of the International Conference on Vision towards Emerging Trends in Communication and Networking (ViTECoN), Vellore, India, 30–31 March 2019.

[2] Karan, B.V.; Narayan, D.G.; Hiremath, P.S. Detection of DDoS Attacks in Software Defined Networks. In Proceedings of the 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 20–22 December 2018; pp. 265–270.

[3] Cil, A.E.; Yildiz, K.; Buldu, A. Detection of DDoS attacks with feed forward based deep neural network model. Expert Syst. Appl.2021, 169, 114520.

[4] Nam, T.M.; Phong, P.H.; Khoa, T.D.; Huong, T.T.; Nam, P.N.; Thanh, N.H.; Thang, L.X.; Tuan, P.A.; Dung, L.Q.; Loi, V.D. Self-organizing map-based approaches in DDoS flooding detection using SDN. In Proceedings of the 2018 International Conference on Information Networking (ICOIN), Chiang Mai, Thailand, 10–12 January 2018; pp. 249–254.

[5] Adhikary, K.; Bhushan, S.; Kumar, S.; Dutta, K. Decision Tree and Neural Network Based Hybrid Algorithm for Detecting and Preventing DDoS Attacks in VANETS. Int. J. Innov. Technol. Explor. Eng. 2020, 9, 669–675.

[6] Hosseini, S.; Azizi, M. The hybrid technique for DDoS detection with supervised learning algorithms. Comput. Netw. 2019, 158, 35–45.

[7] Ujjan, R.M.A.; Pervez, Z.; Dahal, K.; Bashir, A.K.; Mumtaz, R.; González, J. Towards slow and adaptive polling sampling for deep learning based DDoS detection in SDN. Future. Gener. Comput. Syst. 2020, 111, 763–779.

[8] Ravi, N.; Shalinie, S.M. Learning-Driven Detection and Mitigation of DDoS Attack in IoT via SDN-Cloud Architecture. IEEE Internet Things J. 2020, 7, 3559–3570.

[9] Yong, B.; Wei, W.; Li, K.C.; Shen, J.; Zhou, Q.; Wozniak, M.; Połap, D.; Damaševiˇcius, R. Ensemble machine learning approaches for web shell detection in Internet of things environments. Trans. Emerg. Telecomm. Technol. 2020, e4085.

[10] Kushwah, G.S.; Ranga, V. Optimized extreme learning machine for detecting DDoS attacks in cloud computing. Comput. Secur.2021, 105, 102260.

[11] Damasevicius, R.; Venckauskas, A.; Grigaliunas, S.; Toldinas, J.; Morkevicius, N.; Aleliunas, T.; Smuikys, P. Litnet-2020: An annotated real-world network flow dataset for network intrusion detection. Electronics 2020, 9, 800.

[12] Erhan, D.; Anarım, E. Boˇgaziçi] University distributed denial of service dataset. Data Brief 2020, 32, 106187.

[13] Elsayed, M.S.; Le-Khac, N.A.; Jurcut, A.D. InSDN: A novel SDN intrusion dataset. IEEE Access 2020, 8, 165263–165284.

[14] Ahuja, N.; Singal, G.; Mukhopadhyay, D. DLSDN: Deep learning for DDOS attack detection in software defined networking. In Proceedings of the 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 28–29 January 2021; pp. 683–688.

[15] Ahuja, N.; Singal, G.; Mukhopadhyay, D. “DDOS attack SDN Dataset”, Mendeley Data, V1; Bennett University: Greater Noida, India, 2020.

[16] Shao, E. Encoding IP Address as a Feature for Network Intrusion Detection. Ph.D. Dissertation, Purdue University Graduate School, West Lafayette, Indiana, 2019.

[17] Detection of myocardial infarction from ECG signals. In Proceedings of the 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2–5 May 2018; pp. 1–4.