**Detection of DDoS attack using Machine Learning**

**UROP REPORT**

**SRM UNIVERSITY AMARVATHI**

PRESENTED BY:

M. Hemanth

K. Laharish

M. Shashank

K. Shiva Chaitanya

**Contents**

[**ABSTARCT:** 1](#_Toc71022178)

[**INRODUCTION:-** 2](#_Toc71022179)

[**RELATED WORK:** 3](#_Toc71022180)

[**PROPOSAL IDEA:** 5](#_Toc71022181)

[**K Nearest Neighbours:-** 5](#_Toc71022182)

[**Logistic Regression:** 6](#_Toc71022183)

[**Random forest:** 7](#_Toc71022184)

[**XGBoost**  : 7](#_Toc71022185)

[**Formula:** 8](#_Toc71022186)

[**Classification Block Diagram:-** 9](#_Toc71022187)

[**Evaluation of result:** 10](#_Toc71022188)

[**Accuracy:** 11](#_Toc71022189)

[**CONCLUSION:** 13](#_Toc71022190)

[**FUTURE WORK :** 13](#_Toc71022191)

[**Reference links:** 14](#_Toc71022192)

Detection of ddos attack using machine learing technique : comparitive analysis

# **ABSTARCT:**

In online Services, Distributed Denial of Service (DDoS) stays as one of the principle dangers. attackers can execute DDoS by the means which is simpler and with the high efficiency, to slow the process for the client's entrance. To recognize the DDoS attack, ML calculations are utilized. The administered machine learning calculations like k-nearest neighbours (k-NN), logistic regression and random forest are utilized for recognition and alleviation of attack. There are three stages: data gathering, Pre-processing and feature extraction in the grouping calculation for recognition of "Normal or DDoS" attack utilizing the NSL-KDD dataset.it required the most effective model to distinguish malicious activities as quick as could really be expected and accurate. Various calculations show distinctive conduct dependent on the select features. The presentation of DDOS attack detection is looked at and best algorithm is proposed.

# **INRODUCTION:-**

Distributed denial of service (DDoS) may be a special sort of denial-of-service attack. Here the DDoS detection model and defence system software based on in-depth learning in the environment. Models can learn patterns from network traffic and trace network attack activities. The results demonstrate the much better performance of the model compared with conventional machine learning ways.

Despite the large number of traditional recognition solutions currently available, DDoS attacks still increase with frequency and intensity. Many defence methods are conducted to detect and reduce the DoS attacks in these methods are based on the collection of flow characteristics received. Recently, machine learning (ML) and data processing techniques are playing an important role within the detection and therefore the classification of intrusion attacks. Several machine learning studies are conducted in several domains. There are many issues that can influence machine learning performance, such as the feature selection methods, the dataset used etc. However, there are not any related works that perform a scientific analysis of those machine-learning techniques.

DDoS attacks which are against the controller and against the OpenFlow switch. Denial of Service attack fall on the classification of basic attack that borrow the accessibility of the system. These attacks have gotten refined and keep it up developing at a fast pace. So to identify and counter these attacks has become a difficult to undertaking.

# **RELATED WORK:**

An AI device called WEKA is utilized to order different sorts of assaults. It has been seen that J48 calculation created best outcomes when contrasted with Random Forest and Naïve Bayes algorithms.[1]. Profound learning approach can consequently remove undeniable level highlights from low-level ones and gain incredible portrayal and induction. We plan a repetitive profound neural organization to take in designs from successions of organization traffic and follow network assault activities.[2]. This framework use measurable data from both the cloud worker's hypervisor and the virtual machines, to forestall network bundles from being conveyed to the external organization. We assess nine AI calculations and cautiously analyse their exhibition. Our trial results show that over 99.7% of four sorts of DOS assaults are effectively distinguished. [3]. This may possess a significant segment of organization transmission capacity of the casualty cloud foundations or devour a large part of the workers time. Accordingly, in this work, we planned a DDoS location framework dependent on the C.4.5 calculation to relieve the DDoS danger. [4]. The paper using the blend of the neural organization and the help vector machine presents the identification and the characterization strategy for the DDOS assaults in the media transmission network.[5]. There are three stages: data gathering, Pre-preparing and highlight extraction in the characterization calculation for discovery of "Typical or DDoS" assault utilizing the NSL-KDD dataset. Various calculations show distinctive conduct dependent on the chose highlights. [6]. To get the higher True Negative Rate (TNR), exactness, and accuracy and to ensure the heartiness, dependability, and comprehensiveness of recognition framework, in this paper, we propose a DDoS assault identification technique dependent on crossover heterogeneous multi classifier troupe learning and plan a heuristic discovery calculation dependent on Singular Value Decomposition (SVD) to develop our location framework. [7]. The task is centered around distinguishing a portion of the likely properties of a DoS assault dependent on figured load for every one of the characteristics utilizing entropy computation. Likewise, the determination of potential credits dependent on client characterized picked granulation is additionally given utilizing NSL KDD dataset.[8]. All through we'll accomplish some work by hand that should be possible in more successful manners utilizing conveyed usefulness inside sci-pack. The goal here is to be more purposeful about the way toward understanding what we're doing and why.[9]. In fact, this technique can recognize even exceptionally unobtrusive assaults just somewhat not the same as the typical practices. The direct intricacy of the strategy makes its ongoing identification practical[10]. We thought about Deep Neural Network (DNN) for recognizing the assaults in IoT. Astute interruption discovery framework must be constructed if there is accessibility of a compelling information set.[11]. DDoS assault was performed utilizing ping of death strategy and identified utilizing AI procedure by utilizing WEKA apparatus [12]. This work joins different AI calculations: Support Vector Machine, Naive Bayes, and Random Forest for grouping [13]. The exactness of the counter phishing arrangement relies upon highlights set, preparing information and AI calculation. This paper presents an exhaustive examination of Phishing assaults, their misuse, a portion of the new AI based methodologies for phishing identification and their near investigation [14]. Be that as it may, it should be appropriately shielded from security dangers. One significant assault that torment the SDN network is the circulated forswearing of-administration (DDoS) assault. There are a few ways to deal with forestall the DDoS assault in a SDN network.[15]. They creates better methodology of half and half Machine Learning procedures instead of existing KNN on a similar informational collection giving more exactness of identifying DDOS assaults on higher accuracy rate[16]. The distinctive AI calculations received for achieving the undertaking are Naive Bayes, K-Nearest neighbor (KNN) and Support vector machine (SVM) to recognize the strange conduct of the information traffic. These three calculations are contrasted agreeing with their exhibitions and KNN is discovered to be the reasonable one over other two [17]. As bots have constantly developed and turn out to be progressively complex, just utilizing stream based traffic examination or chart based investigation for the identification would bring about bogus negatives or bogus positives, or can even be avoided [18]. ADIN Suite proposes a guide to defeat these difficulties with multi-module arrangement. It tackles the requirement for certifiable organization traffic, a versatile cross breed examination to diminish mistake rates [19]. At that point, another dataset was made utilizing highlight choice techniques on the current dataset. Highlight determination strategies were liked to improve on the models, work with their understanding, and give a more limited preparing time[20].

# **PROPOSAL IDEA:**

To detect DDoS attacks, machine learning models supported with feature selection methods are utilized in this study. Machine learning may be a method that pulls implications from existing data by using mathematical and statistical methods, and makes predictions about the unknown with these implications. In the literature, a spread of machine learning models are proposed. The taxonomy of models are often summarized with the kernel-based, distance-based, neural network- based and probability-based characteristics. The continuing research suggests that there's no universal best model for all classification tasks. The literature suggests that random forest, KNN, and logistic regression models have shown good performance for solutions of classification problems. In this study, as candidates of their classification group, the performance rate for detecting DDoS attacks from random forest, KNN, and logistic regression models were investigated.

## **K Nearest Neighbours:-**

The KNN algorithm may be a simple, easily applicable, and supervised machine learning algorithm which will be used for solving both classification and regression problems. When new data comes in, it determines the category of the new data by watching its nearest K neighbours. Manhattan, Minkowski, and Euclidean distance functions are used for the space between two data. The Euclidean distance function was utilized in this study. The similarity between the sample to be classified and therefore the samples found within the classes was detected. When new data was encountered, the distance of this data to the data in the training set was calculated individually by using the Euclidean function. Then, the classification set was created by selecting the k dataset from the littlest distance. The number of neighbouring KNN (k) is predicated on the worth of classification. During the classification, k was determined as 10.

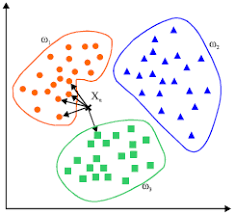


Fig: 1

## **Logistic Regression:**

Logistic regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical.

For example,

* To predict whether an email is spam (1) or (0)
* Whether the tumor is malignant (1) or not (0)

Consider a scenario where we need to classify whether an email is spam or not. If we use linear regression for this problem, there is a need for setting up a threshold based on which classification can be done. Say if the actual class is malignant, predicted continuous value 0.4 and the threshold value is 0.5, the data point will be classified as not malignant which can lead to serious consequence in real time.

From this example, it can be inferred that linear regression is not suitable for classification problem. Linear regression is unbounded, and this brings logistic regression into picture. Their value strictly ranges from 0 to 1.[21]

## **Random forest:**

Random forest is a [supervised learning algorithm](https://builtin.com/data-science/supervised-learning-python). The "forest" it builds, is an ensemble of decision trees, usually trained with the “xgboost” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

**Put simply: random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.**

One big advantage of random forest is that it can be used for both classification and regression problems, which form the majority of current machine learning systems. Let's look at random forest in classification, since classification is sometimes considered the building block of machine learning[22].

## [**XGBoost**](https://xgboost.ai/)**:**

[XGBoost](https://xgboost.ai/)is a decision-tree-based ensemble Machine Learning algorithm that uses a [gradient boosting](https://en.wikipedia.org/wiki/Gradient_boosting) framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now. Please see the chart below for the evolution of tree-based algorithms over the years.

XGBoost and Gradient Boosting Machines (GBMs) are both ensemble tree methods that apply the principle of boosting weak learners ([CARTs](https://www.datasciencecentral.com/profiles/blogs/introduction-to-classification-regression-trees-cart) generally) using the gradient descent architecture. However, XGBoost improves upon the base GBM framework through systems optimization and algorithmic enhancements. XGboost makes highest boosting techniques among of all and it time efficient as well. So, its chosen for many datasets for boosting technique.

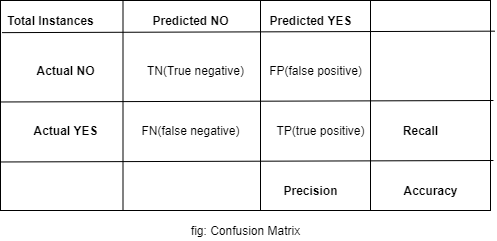


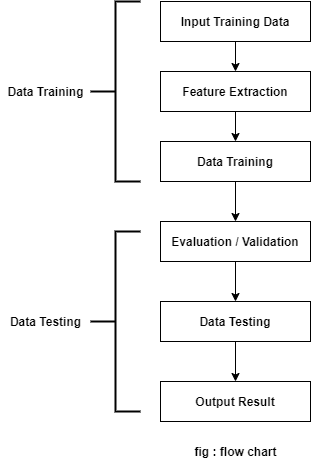
Fig : 2

### **Formula:**

Accuracy = T P + T N / Totalinstances

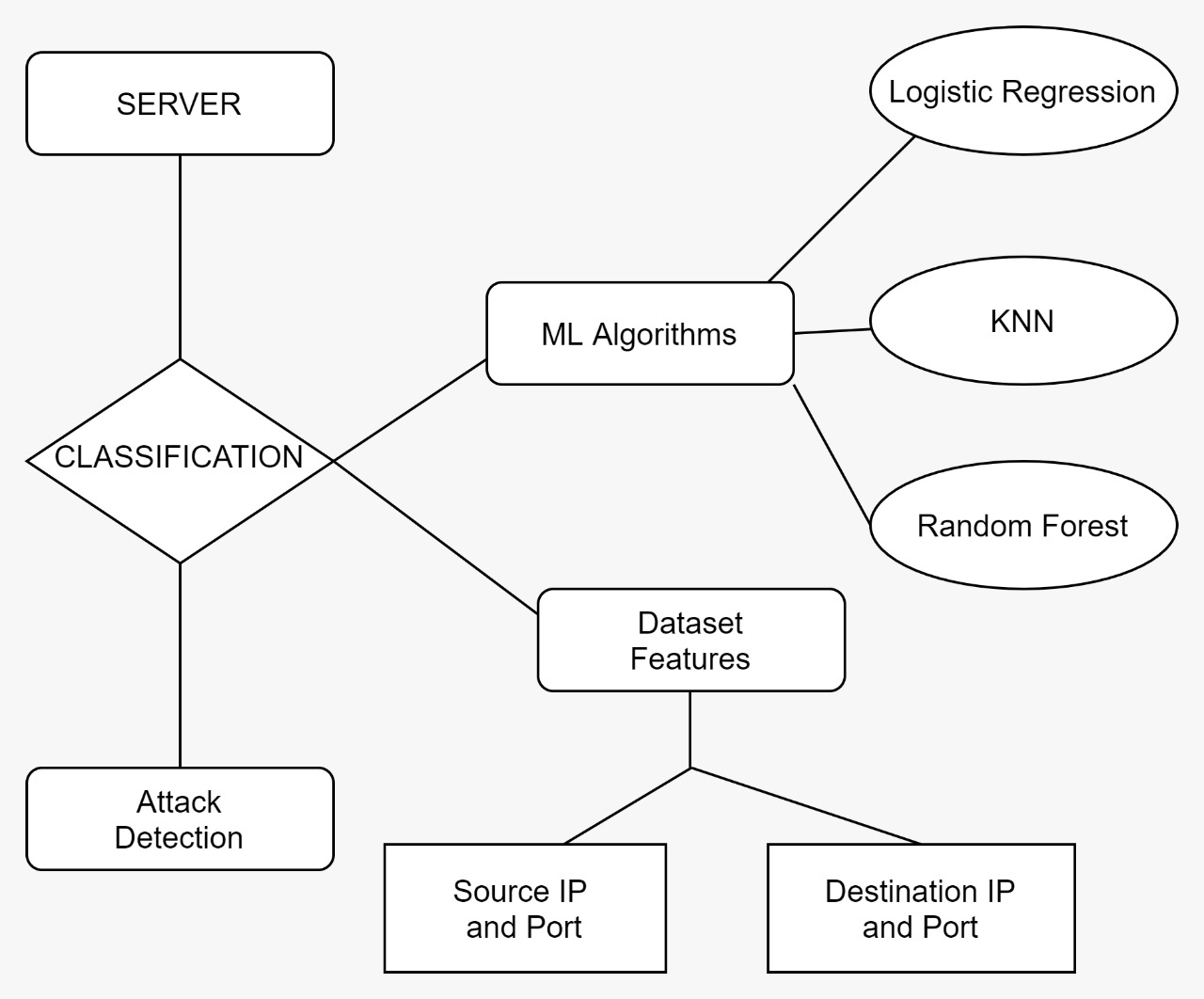
Recall = T P / TotalActuallyYES

Precision = T P / TotalPredictedYES



**Fig : 3**

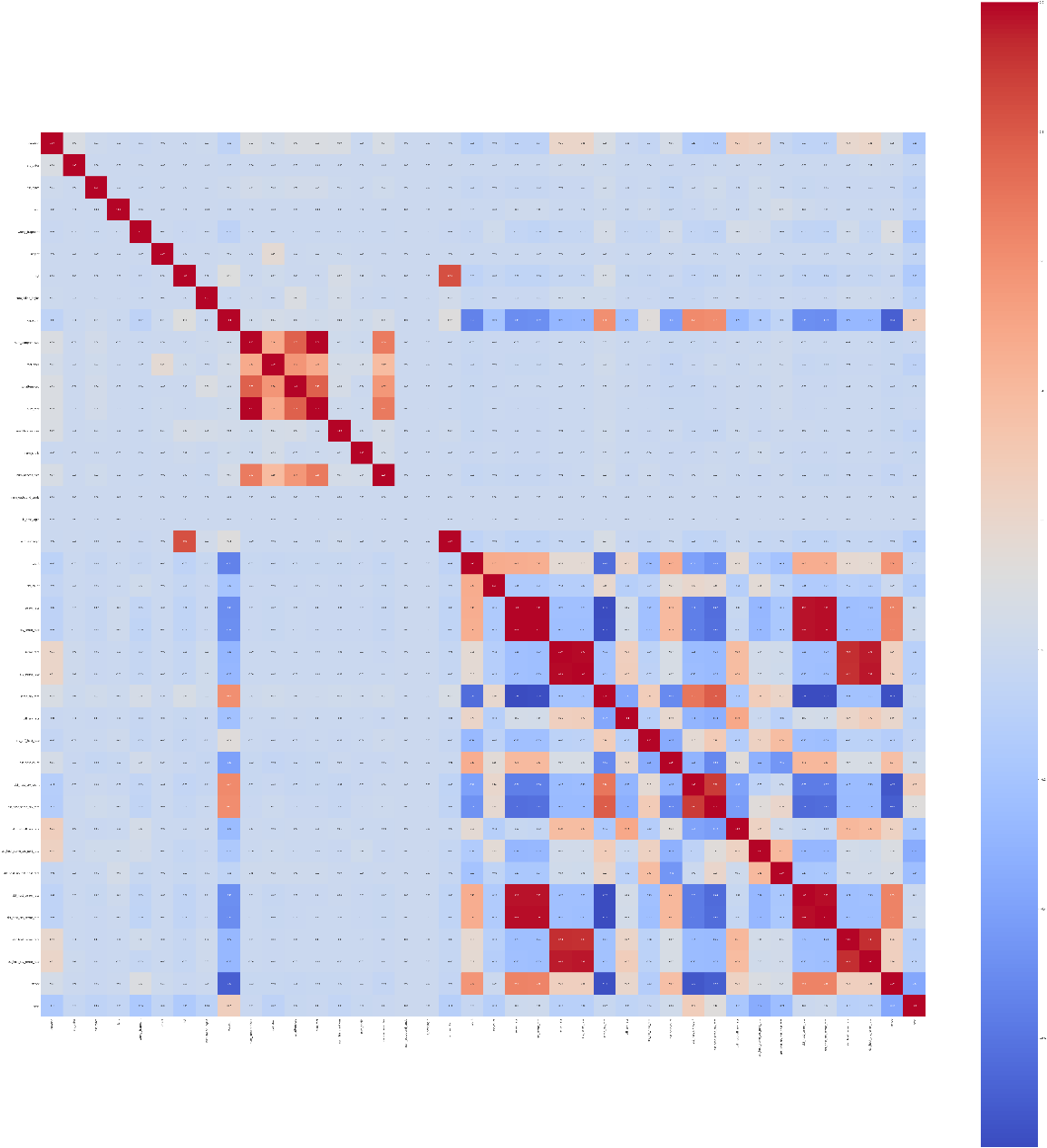
## **Classification Block Diagram:-**



**Fig : 4**

# **Evaluation of result:**

Feature mining is tough in the starting, then figure out to eliminate the non-numerical columns and applied a confusion matrix to find the best features to train the model. We took all the features of attack value is more than 0.6 then we got the features.



**Fig : 5**

## **Accuracy:**

This is the result after training and testing for different classifiers to know the which model is producing the highest accuracy percentage. we trained the features that derived from confusion matrix above. From the below result its random forest classifier and KNN classifier. The low accuracy is with Logistic Regression classifier so we applied a boosting technique for logistic regression for better result. In this process cross validation is also done

“[('RandomForestClassifier', 0, 0.9982139313355824)

('RandomForestClassifier', 1, 0.9974201230402858)

('RandomForestClassifier', 2, 0.9980154792617583)

('RandomForestClassifier', 3, 0.9984122258608713)

('RandomForestClassifier', 4, 0.9974198670239158)

('LogisticRegression', 0, 0.4986108354832308)

('LogisticRegression', 1, 0.4674538598928359)

('LogisticRegression', 2, 0.46417940067473706)

('LogisticRegression', 3, 0.46412622804406073)

('LogisticRegression', 4, 0.46412622804406073)

('KNeighborsClassifier', 0, 0.9949394721174837)

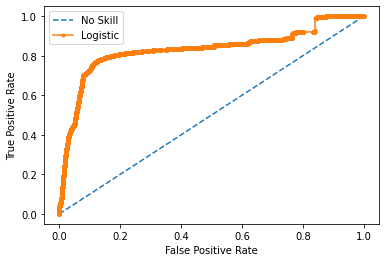
('KNeighborsClassifier', 1, 0.9922603691208574)

('KNeighborsClassifier', 2, 0.9945425679698353)

('KNeighborsClassifier', 3, 0.993251959908703)

('KNeighborsClassifier', 4, 0.993251959908703)]”

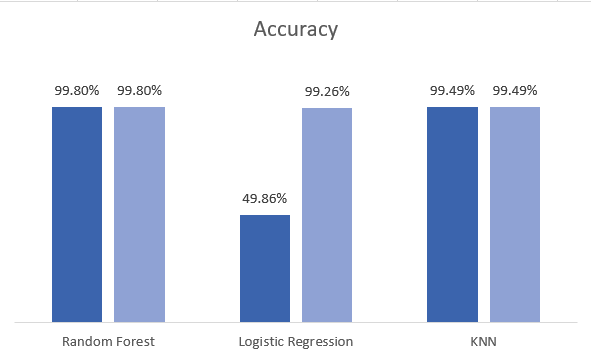
The ROC curve is derived in between the true positive rate and false positive rate . The dotted lines represent the no skill set and the yellow line represents the logistic regression of the trained KDD dataset.



**Fig : 6**

After implementing the XGboosting technique the accuracy of the logistic regression increased nearly 100 percentage. The boosting technique makes the dataset very usable.

“[Testing accuracy is: 0.9926024637910531]”



A small bar graph for compassion before boosting and after boosting technique on logistic regression.

# **CONCLUSION:**

In this work, strategy for include choice called Feature Selection for NSL-KDD Intrusion Dataset is introduced and 11 highlights have been chosen as the main highlights that contribute a lot to improve the exhibition of interruption location frameworks. With a successful examination of highlights chose by two existing strategies and with 10 unique classifiers, it is refuted that rate essentially, expands exactness significantly. Any interruption identification framework, whenever tried with these 11 credits will yield preferable outcomes over with every one of the 41 attributes. The analysis we used is comparative analysis, this helps to find accuracies and cross validation became easy with this method. Then we used a XGboosting technique on logistic regression to compare the previous accuracy. The boosting got accuracy became doubled.

# **FUTURE WORK :**

This work gives the foundations for a number of future projects. First of all, we want to create an interactive web that helps to analyse the state-of-the-art of the DDoS panorama. This repository will include an interactive version of the proposed classifiers, in order to provide an up-to-date reference for DDoS attacks. We will also supply this database with malwares source codes, if available, as well as with exploits that they abuse to infect victims. We also aim to make this database open to other research teams, in order to collect and organize all the useful data. These kinds of information are usually scattered around the web and it takes a lot of time to sort them out; therefore, we hope to simplify the investigation process by joining researchers’ efforts.

# 

# **Reference links:**

[1] P. S. Saini, S. Behal and S. Bhatia, "Detection of DDoS Attacks using Machine Learning Algorithms," 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom), 2020, pp. 16-21, doi: 10.23919/INDIACom49435.2020.9083716.

[2] Yuan, Xiaoyong, Chuanhuang Li, and Xiaolin Li. "DeepDefense: identifying DDoS attack via deep learning." *2017 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, 2017.

[3] He, Zecheng, Tianwei Zhang, and Ruby B. Lee. "Machine learning based DDoS attack detection from source side in cloud." *2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud)*. IEEE, 2017.

[4] Zekri, Marwane, et al. "DDoS attack detection using machine learning techniques in cloud computing environments." *2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech)*. IEEE, 2017.

[5] Smys, S. "DDOS attack detection in telecommunication network using machine learning." *Journal of Ubiquitous Computing and Communication Technologies (UCCT)* 1.01 (2019): 33-44.

[6]https://www.researchgate.net/publication/339680393\_Detection\_of\_Distributed\_Denial\_of\_Service\_Attack\_Using\_NSL-KDD\_Dataset\_-\_A\_Survey

[7] Jia, Bin & Huang, Xiaohong & Liu, Rujun & Ma, Yan. (2017). A DDoS Attack Detection Method Based on Hybrid Heterogeneous Multiclassifier Ensemble Learning. Journal of Electrical and Computer Engineering. 2017. 1-9. 10.1155/2017/4975343.

[8] Khan, Suleman & Gani, Abdullah & Wahid, Ainuddin & Singh, Prem. (2017). Feature Selection of Denial-of-Service Attacks Using Entropy and Granular Computing. Arabian Journal for Science and Engineering. 43. 499-508. 10.1007/s13369-017-2634-8.

[9] C. Zhang, F. Ruan, L. Yin, X. Chen, L. Zhai and F. Liu, "A Deep Learning Approach for Network Intrusion Detection Based on NSL-KDD Dataset," 2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID), 2019, pp. 41-45, doi: 10.1109/ICASID.2019.8925239.

[10] Shuyuan Jin and D. S. Yeung, "A covariance analysis model for DDoS attack detection," 2004 IEEE International Conference on Communications (IEEE Cat. No.04CH37577), 2004, pp. 1882-1886 Vol.4, doi: 10.1109/ICC.2004.1312847.

[11] Sarika Choudhary, Nishtha Kesswani, Analysis of KDD-Cup’99, NSL-KDD and UNSW-NB15 Datasets using Deep Learning in IoT, procedia Computer Science, Volume 167, 2020, Pages 1561-1573, ISSN 1877-0509, doi.org/10.1016/j.procs.2020.03.367.

[12] Pande S., Khamparia A., Gupta D., Thanh D.N.H. (2021) DDOS Detection Using Machine Learning Technique. In: Khanna A., Singh A.K., Swaroop A. (eds) Recent Studies on Computational Intelligence. Studies in Computational Intelligence, vol 921. Springer, Singapore. doi.org/10.1007/978-981-15-8469-5\_5

[13] A. R. Wani, Q. P. Rana, U. Saxena and N. Pandey, "Analysis and Detection of DDoS Attacks on Cloud Computing Environment using Machine Learning Techniques," 2019 Amity International Conference on Artificial Intelligence (AICAI), 2019, pp. 870-875, doi: 10.1109/AICAI.2019.8701238.

[14] A. K. Jain and B. B. Gupta, "Comparative analysis of features based machine learning approaches for phishing detection," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016, pp. 2125-2130.

[15] O. Rahman, M. A. G. Quraishi and C. Lung, "DDoS Attacks Detection and Mitigation in SDN Using Machine Learning," 2019 IEEE World Congress on Services (SERVICES), 2019, pp. 184-189, doi: 10.1109/SERVICES.2019.00051.

[16] G. Kaur and P. Gupta, "Hybrid Approach for detecting DDOS Attacks in Software Defined Networks," 2019 Twelfth International Conference on Contemporary Computing (IC3), 2019, pp. 1-6, doi: 10.1109/IC3.2019.8844944.

[17] A. Prakash and R. Priyadarshini, "An Intelligent Software defined Network Controller for preventing Distributed Denial of Service Attack," 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 585-589, doi: 10.1109/ICICCT.2018.8473340.

[18] Wei Wang, Yaoyao Shang, Yongzhong He, Yidong Li, Jiqiang Liu, BotMark: Automated botnet detection with hybrid analysis of flow-based and graph-based traffic behaviors, information Sciences, Volume 511, 2020, Pages 284-296, ISSN 0020-0255, doi.org/10.1016/j.ins.2019.09.024.

[19] Meshram, Ankush, and Christian Haas. "Anomaly detection in industrial networks using machine learning: a roadmap." *Machine Learning for Cyber Physical Systems*. Springer Vieweg, Berlin, Heidelberg, 2017. 65-72.

[20] Polat, Huseyin, Onur Polat, and Aydin Cetin. "Detecting DDoS attacks in software-defined networks through feature selection methods and machine learning models." *Sustainability* 12.3 (2020): 1035.

[21][towardsdatascience.com](https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc) , Logistic Regression ,Saishruthi Swaminathan, Mar 15 2018.

[22]builtin.com, data-science, A COMPLE A COMPLETE GUIDE TO THE RANDOM FOREST ALGORITHM, Niklas Donges,June 16, 2019,