DDOS ATTACK USING MACHINE LEARNING

***Abstract*— Leveraging machine learning has emerged as a robust strategy for navigating the ever-evolving landscape of Distributed Denial of Service (DDoS) attacks, which exploit vulnerabilities in networks. Traditional defense mechanisms often fall short in countering these attacks, but machine learning offers a promising avenue for timely detection and response. The exploration of various methodologies, encompassing supervised, unsupervised, and hybrid techniques, proves indispensable in the quest to uncover and thwart DDoS attacks. Supervised approaches such as Random Forests, XGboost, Support Vector Machines, and neural networks rely on labeled datasets to differentiate between legitimate and malicious network traffic. Feature engineering plays a pivotal role by extracting relevant attributes from network data, facilitating algorithms in detecting deviations from expected traffic patterns. Deep Learning, featuring Long Short-Term Memory and Multi-Layer Perceptron excels in identifying subtle attack patterns that may evade traditional defenses. Despite challenges such as imbalanced data and the constantly evolving nature of attack strategies, machine learning's adaptability and pattern recognition capabilities empower proactive cybersecurity measures against DDoS threats.**

Keywords—DDos, Random Forest, Decision Tree, Long Short-Term Memory, Multi-Layer Perceptron

I. INDRODUCTION

DDoS attacks, also known as distributed denial of service (DDoS) attacks, target vulnerabilities present in a variety of assets, including the website infrastructure of an organization. A DDoS attack involves flooding a target web asset with requests, frequently using IP spoofing to conceal the attacker's identity. The goal is to handle too many concurrent requests for the resource, making it useless for even authorized users. Web applications and commercial websites are the main targets of these attacks, which are motivated by various factors.

Numerous strategies have emerged for addressing the classification and prevention of DDoS attacks. In a study utilizing the CIC-DDoS2019 dataset, a deep learning framework for intrusion detection was proposed, showcasing models with impressive overall performance. These models included MLP (Multi-Layer Perceptron), LSTM (Long Short-Term Memory), Random Forest, and Decision Tree. Among them, MLP and LSTM demonstrated strong effectiveness, achieving commendable accuracy rates. Another research effort introduced a hybrid deep learning model aimed at enhancing intrusion detection using the same CIC-DDoS2019 dataset. This model combined the strengths of these various algorithms, resulting in significant improvements in accuracy. Notably, multiple deep learning models have found application in DDoS attack detection using the CIC-DDoS2019 dataset, highlighting the dataset's suitability for testing and validation across a range of models. The fusion of machine learning and DDoS attack mitigation holds promise for advancing network security strategies. Machine learning models have demonstrated their aptitude for spotting odd patterns and anomalies in network traffic, assisting in the early detection of potential dangers like DDoS attacks. These models use historical data to forecast and prevent potential attacks, reducing the harm they do to vital services. Because malicious actors are becoming more skilled as technology advances, defense mechanisms must continue to evolve. In addition to providing a way to understand complex network behavior, machine learning also offers an adaptive defense against new dangers like DDoS attacks. The combination of DDoS attack mitigation and machine learning presents a potent shield, fortifying network infrastructures against evolving threats, in an interconnected digital environment.

A. TYPES

DDoS attacks refer to malevolent efforts aimed at interferin g with the regular operations of a network or online service by inundating it with an excessive amount of traffic. The me thods and targets determine the categories in which these att acks can be classified. A critical aspect in effectively defend ing against these attacks lies in comprehending these categor izations.

HTTP Flood Attacks: refer to deliberate attempts to overlo ad a webserver by inundating it with an excessive volume of HTTP requests. It is challenging to identify these attacks si nce they are frequently executed by botnets, which makes th e traffic seem genuine. Assailants might employ various tool s or scripts to create a vast number of requests, resulting in t he depletion of the webserver's resources, such as CPU and memory.

Slow- and low-rate distributed denial-of-service (DDoS) attacks involve the act of attackers sending traffic to a webserver and deliberately maintaining open con nections without providing any responses. As time goes on, this gradually depletes the server's resources, making it imp ossible for authorized users to access. Attacks of this nature are often carried out using tools such as Slowloris and Socks tress.

SIP attacks are specifically directed towards the SIP protoc ol, which plays a crucial role in managing communication si gnaling among various parties. Legitimate clients are unable to access SIP services due to these attacks causing disruptio ns. SIP servers often get overwhelmed by malicious activity caused by attackers utilizing tools like SIPp-DD.

Reflector attacks pertain to the utilization of external "refle ctors" to conceal the identity of the attacker. The perpetrator directs data through these reflectors, giving the impression t hat the data is originating from authentic sources. This result s in a substantial volume of harmful traffic that depletes the network resources of the target.

DNS amplification attacks utilize weaknesses in the DNS s ystem to convert small messages, initially initiated by attack ers, into significantly larger ones. The victim's resources are unable to handle these significant messages, causing their se

rvices to become unavailable. DNS flooder tools are employ ed by assailants to carry out these attacks.
SYN Flooding Attacks: SYN flooding attacks rank high a mong the frequently encountered forms of DDoS attacks. The TCP threeway handshake method is manipulated by atta ckers who flood the server's queue with SYN packets, leadin g to the rejection of genuine requests. The tool frequently e mployed to carry out SYN flooding attacks is Hping3.
Attacks known as UDP flooding aim to overwhelm a serve r by bombarding it with UDP packets on various ports. Whe n the server receives ICMP messages as a response, it uses u p its resources, resulting in it being unresponsive to genuine clients.
ICMP flooding attacks, alternatively referred to as ping att acks, involve overwhelming the target server by sending ma ssive volume of echo requests. Every request necessitates se rver resources for processing causing an overpowering effec t on server and rendering it inaccessible to genuine users.
DHCP flooding attacks, also known as DHCP starvation at tacks, strive to exhaust all the IP addresses that a DHCP serv er can allocate. The server's IP address pool is overwhelmed and legitimate users are unable to connect due to the attacke rs sending multiple DHCP requests.

For organizations and network administrators to successfull y implement security
measures and strategies against DDoS attacks, it is crucial to have a comprehensive knowledge of the various types of the se attacks. Being aware of the changing nature of these attac ks and understanding their categoriesand traits is essential in order to protect online services and networks.

B. MOTIVATION

The authors of the paper suggested various classification algorithms as alternatives to the existing ones, which are deficient in several aspects and have limitations. To begin with, these existing algorithms are unable to operate with values that are not relevant, and the results obtained from the confusion matrix demonstrate the importance of feature engineering. Furthermore, these algorithms can be inaccurate, as there are certain outcomes that are assigned a value of zero, indicating that these algorithms do not perform effectively. Therefore, it is crucial to undergo training in order to model accurately. Another issue arises when certain outcomes demonstrate (Null) that denotes absent values, which are also encompassed within the data, but are not calculated.

LSTM and MLP are distinct algorithms, each suitable for various purposes. LSTM is often used for sequential data, especially in tasks requiring memory and understanding patterns over time, while MLP is employed for a wide range of machine learning tasks. In the study conducted by the authors, both LSTM and MLP were utilized. Specifically, LSTM was employed for intrusion detection due to its capability to handle sequential data effectively. However, it's worth noting that training LSTM models can be time-consuming and resource-intensive. Therefore, it's essential to employ advanced machine learning techniques and optimization methods to facilitate training the best-performing model, resulting in significant outcomes.

In this paper, the primary focus is on intrusion detection, a specific problem domain. Managing these operational algorithms is of utmost importance. One of the current challenges is the absence of a data mining methodology to enhance data quality, which is crucial for effective intrusion detection.

Random forest, a notable machine learning technique, and decision trees, its fundamental building blocks, are pivotal in supervised learning, particularly in classification tasks. Notably, random forest algorithms tend to be significantly slower, approximately 100 times, when compared to single decision trees. Decision trees offer simplicity and interpretability, constructing a transparent model of decisions and consequences. In contrast, random forests, composed of multiple decision trees, excel at enhancing predictive accuracy by reducing overfitting and improving generalization, albeit at the cost of increased computational time. Thus, decision trees are preferred for straightforward tasks where transparency is paramount, while random forests shine in more complex classification problems where predictive performance and efficiency are crucial.

C. RELATED WORK

In this section, recent advancements in DDoS detection systems, specifically tailored for the CICDDoS2019 dataset, are explored. Notably, the authors of [5] have introduced an innovative approach employing a hybrid machine learning methodology. This system combines the robust Extreme Learning Machine (ELM) algorithm with the precision of the black-hole optimization algorithm. Rigorous experimentation across diverse datasets underscores the effectiveness of this hybrid approach in recognizing DDoS attacks. Remarkably, this hybrid system has found successful application in the domain of cloud computing, achieving an impressive detection accuracy rate of 99.80% when applied to the CICDDoS2019 dataset. This amalgamation of advanced machine learning techniques and optimization strategies represents a promising avenue in the ongoing quest for robust DDoS detection, particularly within the dynamic realm of cloud computing.

Part of the research presented in [7] involves a meticulous examination of the impact of data balancing algorithms on the classification of network traffic. This study addresses the challenge of imbalanced data in the context of various DDoS attack types, utilizing the comprehensive CICDDoS2019 dataset, which encompasses diverse data related to exploitation- and reflection-based attacks. The outcomes of this investigation underscore the significance of

data balancing techniques, including but not limited to naive random sampling, synthetic minority sampling, adaptive synthetic sampling, and their application in network attack

classifying the type of attack with a classification rate of approximately 95%.

Filho et al. [9] introduced an intelligent detection system that leverages machine learning for online DDoS attack detection. This system adeptly classifies network traffic utilizing the Random Forest algorithm. The authors conducted a thorough performance evaluation of this innovative system across three distinct datasets, comparing its efficacy to that of other existing systems. The model's training process encompassed six diverse machine learning algorithms, and salient features were extracted through the utilization of the feature importance technique.

In 2020, Budi K. conducted groundbreaking research on enhancing DDoS attack detection. The study leveraged Feature Selection with Information Gain in Random Forest, achieving an unparalleled 99.99% detection accuracy and an exceptionally low false alarm rate of 0.001, as documented in (Budi, K., 2020) [10].

MODEL

In this study, we have meticulously designed a structured framework specifically tailored for the classification and prediction of DDoS attacks using the CIC-DDoS2019 dataset. This framework harnesses advanced machine learning techniques, including MLP (Multi-Layer Perceptron), LSTM (Long Short-Term Memory), Random Forest, and Decision Tree, to make the most of the valuable information contained within this dataset.

Our proposed model, aptly named the "Enhanced Intrusion Detection Framework," capitalizes on a potent ensemble of these machine learning algorithms, exclusively drawing insights from the rich CIC-DDoS2019 dataset. Through meticulous feature engineering and data preprocessing, it extracts critical information to ensure optimal performance. Moreover, our model offers visualization capabilities for insightful result comparisons against the four aforementioned models. This framework signifies a significant leap forward in intrusion detection, offering superior accuracy and robustness specifically tailored to the CIC-DDoS2019 dataset, ultimately advancing the state of the art in intrusion detection.

DATASET

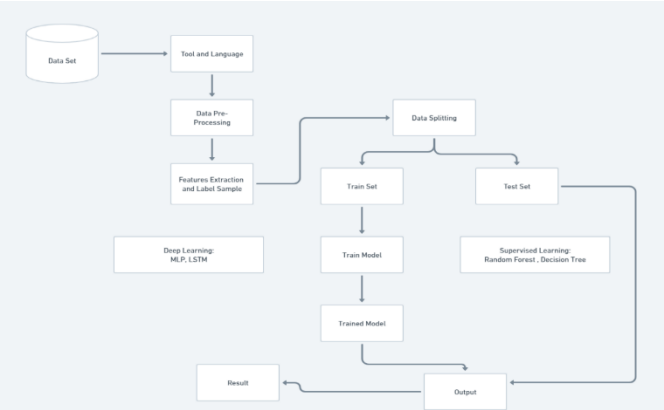
In this experiment, the CIC DDoS2019 datasets were employed for the purpose of training and evaluating various machine learning (ML) algorithms. The objective was to identify the best-performing ML algorithm(s) based on rigorous evaluation criteria. These datasets were sourced from internet resources and served as both test cases and valuable learning tools.

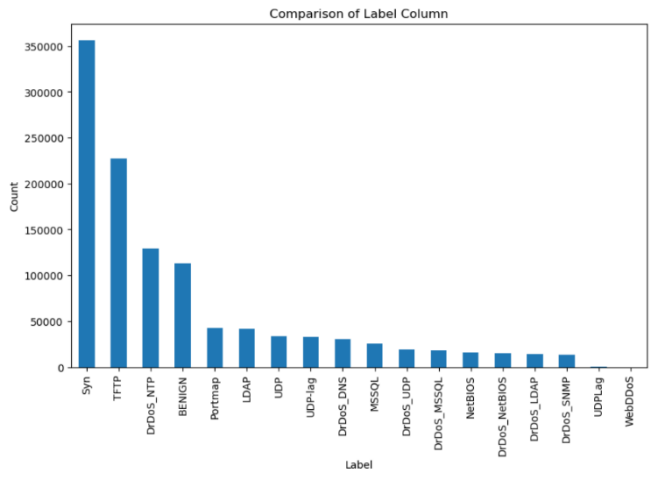
The CIC-DDoS 2019 dataset, provided by the Canadian Centre for Cybersecurity (CIC), represents a significant milestone in the field of cybersecurity. It was meticulously designed to emulate real-world testing scenarios, incorporating insights from actual network traffic assessments. This extensive dataset comprises a staggering 30,480,823 entries, encompassing 56,863 benign cases and 30,423,960 instances of DDoS attacks. However, due to its substantial size, approximately 17.2 GB, it presents challenges for processing on standard computing devices. To make effective use of this data for training artificial intelligence (AI) and machine learning (ML) models, it becomes essential to reduce its size. This reduction not only optimizes computational resource utilization but also streamlines data processing, enabling more efficient model training. Validated with the CIC-DDoS2019 dataset from the Canadian Institute of Cybersecurity, FLAD assesses recent DDoS attacks, encompassing diverse network activities over several days, including benign traffic and 13 types of attacks.

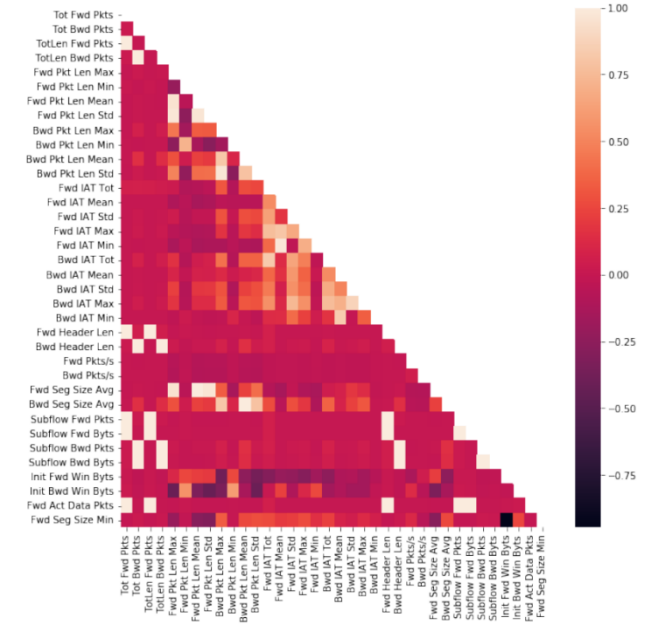
DATA PRE-PROCESSING

Data cleansing stands as a time-consuming yet indispensable facet of data analysis. It serves as the initial phase where superfluous and irrelevant data are meticulously purged, and the remaining information is refined into a high-quality format. This process involves the application of statistical methods to weed out unnecessary values, ensuring that only pertinent data remains for our experimental analysis. Once this crucial cleansing phase is completed, the data can be transformed into a cohesive and structured form, setting the stage for an in-depth examination of its intrinsic value. Ultimately, this refined information can be effectively presented and analyzed through the lens of various graphical styles.

Data preprocessing is an essential precursor to insightful analysis. Beyond cleansing, it includes steps like normalization and feature scaling to ensure uniformity. Transformation techniques enhance data structure for comprehensive examination. This refined data, formatted cohesively, facilitates robust analysis, enabling meaningful insights when visualized through diverse graphical representations.Data Preprocessing is an important step in the Data Preparation stage of a Data Science development lifecycle that will ensure reliable, robust, and consistent results. The main objective of this step is to ensure and check the quality of data before applying any Machine Learning or Data Mining methods.







SUPERVISED LEARNING

Supervised learning is a machine learning approach that relies on labeled training data to instruct an algorithm in making independent predictions or decisions. It earns its "supervised" moniker because a knowledgeable guide oversees the learning process, providing the correct answers during training.

In supervised learning, the model establishes a mathematical or computational representation of the relationship between input features and output labels. Various algorithms serve as models in this context, including decision trees, linear regression, logistic regression, support vector machines, k-nearest neighbors, and neural networks.

The choice of model hinges on the nature of the problem and the characteristics of the data. Some models excel at capturing straightforward, linear patterns, while others thrive in deciphering intricate, non-linear relationships.

RANDOM FOREST

One of the decision tree's techniques is the random forest. The root node is used for data collection, the inner node is used for data analysis, and the leaf node is used for problem solving and decision-making. A decision tree, or decision tree diagram, is a flow diagram in the shape of a tree. Each healthy tree is combined into a single model via a random forest. Each of the decision trees in Random Forest has a maximum depth and is based on the use of a random vector value with the same distribution across all of them. A random forest is a type of classifier that consists of a classifier in the form of a tree, h(x, k), where k is a randomly distributed vector, and each tree in the forest.

DECISION TREE

A decision tree is a versatile and intuitive model used in machine learning and decision analysis. Imagine it as a flowchart for decision-making. It begins with a single node, representing a choice or decision, and then branches out into multiple nodes, each symbolizing potential outcomes. These branches continue to split until they lead to leaf nodes, which signify final decisions or outcomes. Essentially, a decision tree is a hierarchical structure where internal nodes correspond to features or attributes, and branches signify decisions based on these features. It's a supervised learning technique, meaning it learns from labeled data to make predictions or classifications. Decision trees are well-suited for tasks like classification and regression, accommodating both categorical and numerical data. The key concept is to select the most informative features at each node to effectively categorize data or predict numerical values.

DEEP LEANING

Deep learning is a fascinating field within machine learning (ML) that's all about training artificial neural networks to learn and make sense of complex data, much like the human brain does. Imagine you're teaching a computer to recognize cats in photos. Traditional ML might involve manually picking out cat-like features, like ears or whiskers, and instructing the computer to look for them.

Deep learning, on the other hand, works more like how we learn. It starts with a "neural network," which is like a virtual brain made up of layers of interconnected "neurons." These neurons are mathematical functions that process information.

LSTM

Long Short-Term Memory, or LSTM, is a specialized type of neural network renowned for its ability to work with sequential data. Picture it as a smart system capable of both remembering past information and making predictions about what comes next. LSTMs find extensive applications in tasks such as speech recognition, language translation, and stock market forecasting. What sets them apart is their knack for capturing long-range dependencies in data. This means they can connect information from earlier in a sequence to what's happening in the present, offering valuable context for accurate predictions.

Essentially, LSTMs comprise memory cells designed to store and retrieve information over extended periods. These cells learn when to remember, forget, or focus on different aspects of the data. This remarkable ability makes LSTMs exceptionally powerful for handling time series data, natural language processing, and any scenario where recognizing patterns over time is vital. In summary, LSTMs serve as the memory experts of the neural network world, excelling at deciphering sequential data and delivering impressive results in various applications.

MLP

A Multi-Layer Perceptron (MLP) is a versatile neural network used extensively in machine learning. It consists of layers of interconnected nodes and is renowned for its ability to address diverse problems, from image and speech recognition to financial forecasting. In an MLP, there are typically three types of layers: the input layer, one or more hidden layers, and the output layer. The input layer receives data, which is then processed through the hidden layers where complex computations occur. These hidden layers are responsible for learning intricate patterns within the data. Finally, the output layer provides the network's prediction or classification. MLPs are potent because they learn and adapt from data. During training, the network fine-tunes the connections (weights) between neurons to minimize prediction errors. This process, known as backpropagation, enables MLPs to discern intricate patterns and make precise predictions. Nonetheless, designing an effective MLP requires careful consideration of its architecture, including the number of hidden layers and neurons. In summary, MLPs are essential tools in machine learning, employing layers of interconnected neurons to solve a wide array of complex tasks through data-driven learning.

RESULT AND COMPARISON

In the evaluation of Table 1 shows several prominent machine-learning algorithms for a specific task, the Random Forest Classifier emerged as the standout performer, boasting an impressive accuracy score of 99.10%. This result underscores its efficacy as a highly effective algorithm for the given task. Following closely behind, the Decision Tree Classifier demonstrated substantial competence with an accuracy rate of 97.70%. Equally commendable, the Multi Layer Perceptron (MLP) delivered strong results, achieving an accuracy level of 98.46%. Additionally, the Long Short Term Memory (LSTM) algorithm showcased its potential with an accuracy rating of 94.78%, particularly indicating its suitability for tasks involving sequential data. These four algorithms, Random Forest, Decision Tree, MLP, and LSTM, were the focal points of this analysis.

	Machine Learning Model		**Deep Learning Model**	
	Random Forest	Decision tree	Multilayer Perceptron	Long Short Term Memory
Accuracy	99.10%	97.70%	98.46%	94.78%
Precision	0.9902	0.9729	0.984	0.9127
Recall	0.9907	0.9745	0.984	0.9256

Table 1: Result Analysis Summary

In the comparative analysis of deep learning and machine learning models, both paradigms showcase remarkable capabilities in addressing a specific task. Within the realm of machine learning, the Random Forest model delivers exceptional results, boasting a high accuracy of 99.10%, coupled with a precision of 0.9902 and a recall of 0.9907. Similarly, the Decision Tree model achieves a commendable accuracy of 97.70%, supported by a precision of 0.9729 and a recall of 0.9745. These machine learning models excel in providing accurate classifications and exhibit robust performance. In contrast, the deep learning models demonstrate their prowess. The Multilayer Perceptron (MLP) yields an impressive accuracy of 98.46%, accompanied by a precision and recall both at 0.984, showcasing its ability to capture intricate patterns. The Long Short-Term Memory (LSTM) model, tailored for sequential data, maintains a solid accuracy of 94.78%. It showcases a precision of 0.9127 and a recall of 0.9256, underscoring its proficiency in handling data with temporal dependencies.

This analysis underscores the strengths of both deep learning and machine learning approaches, with the former excelling in intricate pattern recognition, and the latter demonstrating strong accuracy and robustness in classification tasks.

CONCLUSION

In conclusion, this comprehensive survey paper has delved deeply into the crucial domain of Distributed Denial of Service (DDoS) attack detection and mitigation, elucidating the paramount role that machine learning and deep learning models play in fortifying network security. The rapidly evolving cyber threat landscape necessitates innovative approaches, and machine learning has emerged as a formidable strategy for countering the pernicious scourge of DDoS attacks. Within the machine learning paradigm, discernible models such as Random Forest and Decision Trees have demonstrated robust capabilities in effectively discerning benign from malevolent network traffic, with Random Forest, in particular, achieving an extraordinary accuracy rate of 99.10%. These models exhibit a penchant for delivering precise classifications, thereby serving as invaluable instruments in fortifying network defenses.

In stark contrast, deep learning models, including Multilayer Perceptrons (MLP) and Long Short-Term Memory (LSTM), manifest a distinctive prowess in capturing intricate patterns, especially in the context of sequential data. Their proficiency is unequivocally evident, with MLP attaining an accuracy rate of 98.46%, and LSTM proving highly effective with an accuracy rating of 94.78%.

In essence, this survey unequivocally underscores that the synergistic alliance between machine learning and DDoS attack mitigation holds immense promise for enhancing network security. These models, underpinned by their adaptability and unparalleled pattern recognition capabilities, empower proactive cybersecurity initiatives. By adroitly leveraging the complementary strengths of both deep learning and machine learning paradigms, organizations can maintain a proactive stance in the ceaseless battle against the ever-evolving and pernicious cyber threats that loom large in our interconnected digital landscape.

References

[1] ISMAIL1, MUHAMMAD ISMAIL MOHMAND UBAID ULLAH HAMEEDHUSSAIN2, AYAZ ALI KHAN3,MUHAMMADZAKARYA ,(Senior Member, IEEE), AFTAB AHMED MUSHTAQ RAZA ,IZAZ UR RAHMAN4,ANDMUHAMMADHALEEM, A Machine Learning Based Classification and Prediction Technique for DDoS Attacks, University of Peshawar, Pakistan; and in part by Kardan University, Kabul, March 2, 2022.

[2] Ebtihal Sameer Alghoson, Onytra Abbass, Detecting Distributed Denial of Service Attacks using Machine Learning Models, Department of Information Technology University of Tabuk, KSA, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 12, 2021

[3] Aratee Giradharbhai Mistry, Adviser: Nahid Ebrahimi Majd Associate Professor Intrusion Detection Classification Using

Machine Learning, Department of Computer Science California State University San Marcos San Marcos, California, United States, 05/03/2023

[4] AHAMED ALJUHANI, Machine Learning Approaches for Combating Distributed Denial of Service Attacks in Modern Networking Environments, Department of Computer Information Systems, University of Tabuk, Tabuk 71491, Saudi Arabia, March 22, 2021

[5] Kushwah, G.S. and Ranga, V., 2021. Optimized extreme learning machine for detecting DDoS attacks in cloud computing. Computers & Security, 105, p.102260

[6] Tewelde Gebremedhin Gebremeskel, Ketema Adere Gemeda and Perumalla Janaki Ramulu , T. Gopi Krishna, DDoSAttack Detection and Classification Using Hybrid Model for Multicontroller SDN, 1Department of Computer Science and Engineering, Center for Electrical System and Electronics, Adama Science and Technology University, Adama, Ethiopia 2Department of Mechanical Engineering, Adama Science and Technology University, Adama, Ethiopia, Volume 2023, 23 June 2023.

[7] Bolodurina, I., Shukhman, A., Parfenov, D., Zhigalov, A. and Zabrodina, L., 2020, November. Investigation of the problem of classifying unbalanced datasets in identifying distributed denial of service attacks. In Journal of Physics: Conference Series (Vol. 1679, No. 4, p. 042020). IOP Publishing.

[8] Shieh, C.-S.; Nguyen, T.-T.;Horng, M.-F. Detection of Unknown DoSAttack Using Convolutional Neural Networks Featuring eometrical Metric. Mathematics 2023,11, 2145. https://doi.org/10.3390/math11092145

[9] Lima Filho, F.S.d., Silveira, F.A., de Medeiros Brito Junior, A., Vargas-Solar, G., Silveira, L.F.:
Smart detection: an online approach for dos/DDoS attack detection using machine learning. Secur. Commun. Netw. 2019 (2019)

[10] Kurniabudi, K., Harris, A., & Rahim, A. 2020. Seleksi Fitur dengan Information Gain untuk Meningkatkan Deteksi Serangan DDoS Menggunakan Random Forest. Techno. Com, 19(1), 56-66.