DDoS Attack Detection using Machine Learning

K. Akash Reddy[1], B. Sourabh[2], D. Rohith Reddy[3], Mr. M. Sreenu[4]
[1,2,3] *Student, Dept. of Computer Science and Engineering (Cyber Security), Geethanjali College of Engineering and Technology, Cheeryal, Keesara, Hyderabad - 501301*
[4] *Associate Professor, Dept. of Computer Science and Engineering (Cyber Security), Geethanjali College of Engineering and Technology, Cheeryal, Keesara, Hyderabad - 501301*

***Abstract*— Distributed Denial of Service (DDoS) attacks continue to pose significant threats to the availability and integrity of online services. Addressing this challenge necessitates the development of robust and efficient detection systems capable of swiftly identifying and mitigating malicious traffic. In this paper, we propose a hybrid DDoS detection system leveraging a combination of machine learning algorithms, including Random Forest, Logistic Regression, Decision Tree, and Support Vector Machine (SVM). The system operates in real-time, continuously monitoring incoming network traffic for anomalous patterns indicative of DDoS attacks. Leveraging the versatility of Random Forest, Logistic Regression, Decision Tree, and SVM algorithms, the system constructs a comprehensive feature space encompassing various network traffic attributes such as packet size, packet frequency, source IP reputation, and traffic volume. The proposed system employs Random Forest to handle complex and nonlinear relationships within the feature space, Logistic Regression for probabilistic modeling and binary classification, Decision Tree for hierarchical partitioning of data, and SVM for effective separation of normal and anomalous traffic patterns in high-dimensional spaces. By integrating these diverse algorithms, the system achieves enhanced accuracy, scalability, and resilience against sophisticated DDoS attacks.**

I. INTRODUCTION

In an era marked by the continual evolution of Internet technologies, the digital landscape has become a hub of convenience, offering a myriad of services to users worldwide. However, this interconnectedness also exposes us to a host of security threats, ranging from network viruses to malicious attacks, thus thrusting network security into the forefront of societal and governmental concerns. DDoS (Distributed Denial of Service) attacks, in particular, pose significant challenges, disrupting the availability of online services and causing financial losses to businesses.

The delineation of network traffic into two overarching categories—normal and malicious—further underscores the need for robust DDoS detection mechanisms. Moreover, the subdivision of network traffic into five distinct categories—Normal, DoS (Denial of Service), R2L (Root to Local), U2R (User to Root), and Probe (Probing attacks)—underscores the intricate nature of the classification task inherent in DDoS detection. As such, the efficacy of DDoS detection hinges upon the accurate identification of malicious activities within the network.

Traditionally, machine learning techniques have served as the cornerstone of DDoS detection systems. Approaches such as Naive Bayes, Support Vector Machines (SVM), neural networks, ensemble methods, and decision trees have been instrumental in addressing the challenges posed by malicious network traffic. These algorithms offer varying degrees of performance and flexibility in handling the diverse nature of network data.

However, existing methodologies have largely treated network traffic as monolithic entities, neglecting the hierarchical structure inherent within. Network traffic, comprising traffic units and data packets, possesses distinct features that necessitate nuanced analysis. By recognizing the hierarchical nature of network traffic and extracting sequential features between data packets, we can harness the full potential of machine learning techniques in enhancing DDoS detection accuracy.

This project seeks to address these limitations by proposing a novel approach that leverages machine learning algorithms while embracing the hierarchical structure of network traffic. Through a comprehensive analysis of various machine learning

models, including Naive Bayes, SVM, neural networks, ensemble methods, and decision trees, we aim to develop a robust DDoS detection system capable of accurately classifying malicious network traffic. By integrating domain knowledge of network traffic characteristics, we endeavour to enhance the efficacy and reliability of DDoS detection systems in safeguarding network integrity and ensuring the seamless operation of digital infrastructures.

II. LITERATURE SURVEY

1. A DDoS Attack Detection Method Based on SVM in Software Defined Network by Jin Ye, Xiangyang Cheng , Jian Zhu,Luting Feng and Ling Song.

In this paper, the status information of the network traffic is collected on the switch by the controller. We extracted the six-tuple characteristic values related to DDoS attack and then used the support vector machine algorithm to judge the traffic and carry out DDoS attack detection. We focus on the analysis of the changes of the characteristic values of traffic and verify the feasibility of this method by deploying the SDN experimental environment.Te detection accuracy rate of the experiment is high and the false alarm rate is low, which has obtained our expected results. In comparison, the test detection accuracy rate of ICMP attacks is relatively low. By analyzing the ICMP traffic, we have come to the conclusion that the ICMP flow has no source port and destination port, so SSP and RPF are zero, which makes the six-tuple characteristic values matrix change into a fourtuple characteristic values matrix, whether attacked or not.

Limitations:

1. Limited Robustness: SVMs may struggle to generalize well to diverse DDoS attack scenarios, leading to reduced accuracy in detecting novel or evolving threats.
2. Sensitivity to Parameters: SVM performance heavily relies on parameter tuning, and selecting optimal parameters may be challenging, especially in dynamic network environments.
3. Scalability Concerns: SVMs can become computationally expensive and memory-intensive, particularly when dealing with large-scale network traffic datasets, potentially limiting their practical applicability in real-time DDoS detection systems

2.High-Speed Network DDoS Attack Detection: A Survey by Rana M. Abdul Haseeb-ur-rehman , Azana Hafizah Mohd Aman.

This paper primarily classified DDoS attacks and the types that can occur in a highspeed network. The DDoS issue is a rapidly growing problem. This research also examined the various existing solutions for detecting DDoS attacks, including traceback mechanisms, which are classified into proactive and reactive approaches, packet marking such as PPM and DPM, and application layer protocol analyses to improve the detection accuracy in terms of the monitoring and filtering of affected data packets using an express data path. This article detailed the growing differences between regular and irregular traffic in terms of DDoS attacks. Additionally, the high-speed vulnerabilities, problems, and challenges of the network layer for maximum packet processing were also explored. High-speed packet processing, detecting, and preventing a DDoS is difficult, and the packet drop ratio is high. DDoS mitigation in high-speed networks is progressing quickly, and researchers are developing efficient and innovative solutions. The open issues and challenges discussed above provide an ideal picture for future directions regarding DDoS detection in high-speed networks. Different studies have been proposed to process data quickly based on 100 Gbe.

Limitations:

1. The conclusion lacks specific solutions tailored to high-speed networks, overlooking practical implementation challenges and performance metrics essential for evaluating effectiveness.
2. The paper fails to delve into implementation challenges and lacks thorough analysis of key performance metrics, limiting its practical applicability in real-world high-speed network environments.

3. DDoS attacks and machine-learning-based detection methods: A survey and taxonomy by Mohammad Najafimehr Sajjad Zarifzadeh Seyedakbar Mostafav.

This paper has provided an in-depth analysis of machine learning-based approaches used to identify various types of DDoS attacks. Our investigation reveals that while supervised learning methods are effective, they require pre-labeled datasets and training, which is unfeasible for not-yet-known

attacks. In contrast, unsupervised methods can be applied more widely to distinguish DDoS attack traffic from benign traffic under unknown circumstances, albeit with less accuracy and detection ability than supervised methods. Combining both supervised and unsupervised methods, along with non-ML methods, may offer the most effective approach to identify known or unknown attacks. However, due to emerging novel and unknown types of DDoS attacks, there are noticeable differences between known and lab-based train datasets and the unforeseen factors that occur in real DDoS attacks. Consequently, the recall is low while the false-negative rate is high. We recommend further research on developing resilient and effective methods that accurately detect malicious traffic under real attack scenarios and different test datasets. Furthermore, the present datasets employed for DDoS research have certain limitations. For instance, the KDDCup99 and NSL-KDD datasets have become outdated and do not encompass the latest innovative and advanced DDoS attacks. Similarly, the CICIDS2017 and Edge-IIoTset lack several novel types of DDoS attacks, rendering them inadequate for such detection purposes. Moreover, the CICDDoS2019 dataset is not suitable for identifying slow and low-rate attacks, and it is also imbalanced with benign-labeled records accounting for less than 1%. These limitations in current datasets underline the need for further and sustained research to provide future-oriented and up-to-date datasets that can assist in the detection and mitigation of DDoS attacks in diverse network environments. In addition to conducting comprehensive research to address the limitations of existing methods and datasets, we propose that researchers focus on developing novel forms of DDoS attacks to proactively anticipate the malicious techniques that may be employed by attackers. Introducing innovative attack types, such as the SlowDrop attack,25 may serve as a crucial measure towards preparing for and mitigating future DDoS attacks.

III. SYSTEM ANALYSIS

Existing system:

In existing methods, while researchers have explored the viability of logistic regression and other traditional machine learning algorithms for DDoS detection, there has been a notable lack of emphasis on the integration of deep learning techniques. Deep learning, with its ability to automatically extract intricate features from raw data, offers a promising avenue for enhancing the accuracy and efficacy of DDoS detection systems. However, existing systems often overlook the potential benefits of deep learning approaches due to concerns regarding computational complexity and the requirement for extensive labeled datasets. As a result, the full potential of deep learning in augmenting DDoS detection capabilities remains largely untapped within the current landscape of existing methodologies.

Disadvantages:

Existing DDoS detection methods face notable drawbacks that hinder their effectiveness in accurately identifying and mitigating attacks. One significant limitation lies in their tendency to treat network traffic as uniform sequences of bytes, disregarding the wealth of domain knowledge available regarding network behaviors and protocols. By failing to fully utilize this information, these methods may overlook critical indicators of DDoS attacks, leading to potentially inaccurate or incomplete detection results.

Furthermore, current approaches often treat network traffic as independent entities, ignoring the intricate internal relationships and dependencies that exist between different packets and flows within the network. This oversight prevents a comprehensive understanding of the complex interactions that characterize DDoS attacks, limiting the ability to detect and mitigate them effectively.

Proposed System:

We utilize the classic NSL-KDD dataset alongside contemporary benchmark datasets, conducting thorough analysis and data cleaning. Our approach addresses the class imbalance problem in intrusion detection by employing a machine learning algorithm that reduces majority samples while augmenting minority samples in the challenging set, enhancing the classifier's ability to discern differences during training. The classification model incorporates Random Forest (RF), Support Vector Machine (SVM), and NLP alongside other methods, leading to a comprehensive experiment.

Furthermore, we introduce an end-to-end deep learning model integrated with traditional machine

learning techniques such as logistic regression and attention mechanism. This hybrid model, leveraging CNN, effectively tackles issues related to Software Defined Networks and offers a novel approach for Early Warning Proactive Systems. By comparing the performance of machine learning models with traditional deep learning methods, we demonstrate the ability of our approach to extract information from each packet, capturing features comprehensively by leveraging the structural information of network traffic.

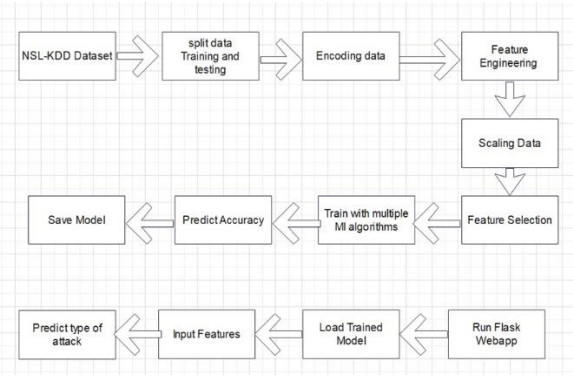
Finally, we evaluate our proposed network using a real NSL-KDD dataset. Experimental results indicate superior performance compared to traditional methods, highlighting the effectiveness of our algorithm in enhancing intrusion detection capabilities.

Advantages:

This method employs an analysis of packet vectors to determine the significance of individual components, thereby extracting fine-grained features that are particularly relevant for the detection of malicious traffic. These features, identified through an attention mechanism, are then fed into a fully connected layer at the output stage. This process facilitates feature fusion, enabling the extraction of key characteristics that precisely depict network traffic behavior.

IV. IMPLEMENTATION

System Architecture



A system architecture, also known as systems architecture, is a conceptual framework that outlines the arrangement, functionality, and various perspectives of a system. It encompasses a formal description and representation of the system, structured to facilitate understanding and analysis of its structures and functionalities.

The three-tier software architecture emerged in the 1990s to address limitations of the two-tier architecture. It involves three layers: the client (user interface), middle tier server (process management), and server (data management). The middle tier executes business logic and rules, improving scalability and performance compared to two-tier systems. It's favored for distributed client/server designs, offering benefits like flexibility, maintainability, and shielding users from distributed processing complexities. This architecture is widely used in Internet applications and net-centric information systems.

Modules

Data Collection

The NSL-KDD dataset features symbolic data types like protocol type, flag, and service. These are converted into numerical features using one-hot encoding, enabling efficient processing.

Pre-Processing

Duplicate entries are removed to keep only one valid instance of each sample. Outliers, such as samples with missing or infinite values, are excluded. Irrelevant attributes like timestamps and addresses are removed. For specific features like 'Init Bwd Win Byts' and 'Init Fwd Win Byts', binary dimensions are introduced to mark occurrences of -1. Categorical features like protocol types are encoded into binary vectors. Numerical data is standardized using Z-Score normalization to ensure uniform impact across all samples.

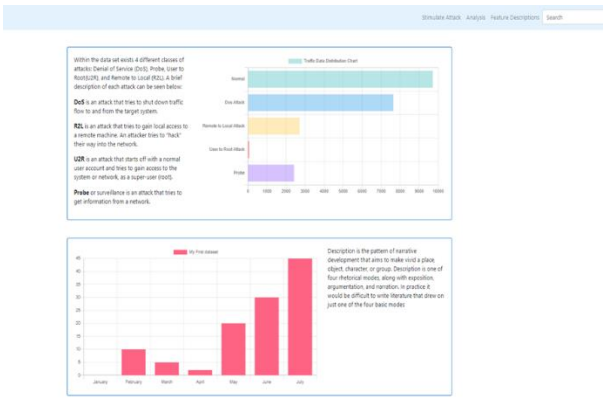
Train-Test Split and Model Fitting

The dataset is split into training and testing subsets to evaluate model performance. Models are then fitted to the training data.

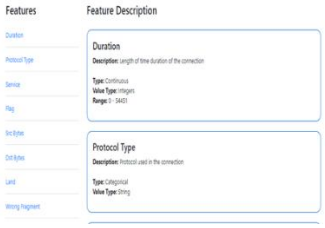
Model Evaluation and Predictions

Model performance is assessed on testing data using accuracy score as the evaluation metric. Predictions are generated for testing data and compared with actual labels to compute accuracy scores across various classification algorithms.

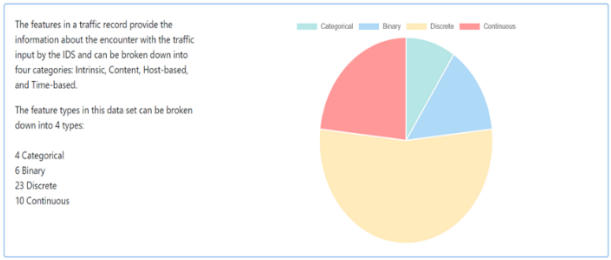
V. RESULTS



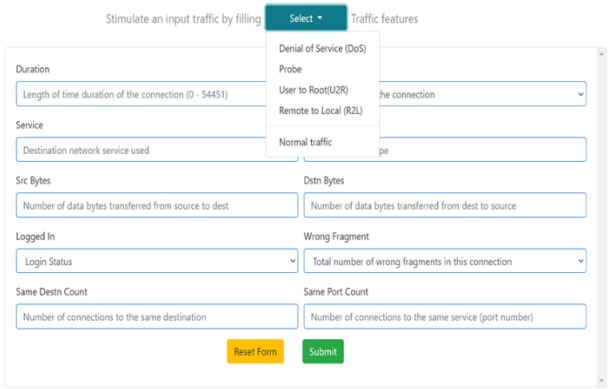
Preliminary Data Analysis



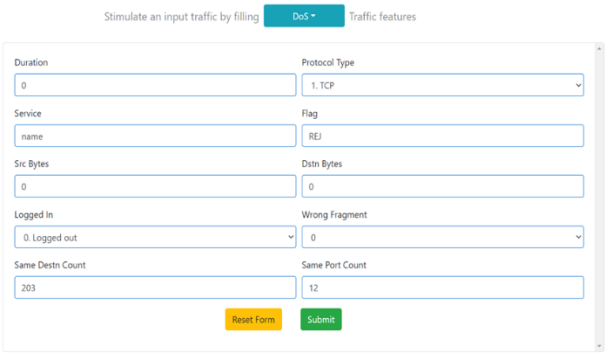
Feature Description



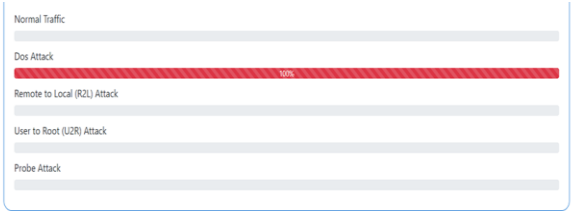
Training Data Analysis



Selecting Attack Type



Entering Packet Information



DDoS Attack Detected

VI. CONCLUSION

As network intrusion continues to evolve, the pressure on network intrusion detection is also increasing. In particular, the problems caused by imbalanced network traffic make it difficult for intrusion detection systems to predict the distribution of malicious attacks, making cyberspace security face a considerable threat. This paper proposed a novel Difficult Set Sampling Technique, which enables the classification model to strengthen imbalanced network data learning. A targeted increase in the number of minority samples that need to be learned can reduce the imbalance of network traffic and strengthen the minority's learning under challenging samples to improve the classification accuracy. We used six classical classification methods in machine learning and deep learning and combined them with other sampling techniques. Experiments show that our method can accurately determine the samples that need to be expanded in the imbalanced network traffic and improve the attack recognition more effectively. In the experiment, we found that deep learning performance is better than machine learning after sampling the imbalanced training set samples through the MLP algorithm. Although the neural networks strengthen data expression, the current public datasets have already extracted the data

features in advance, which is more limited for deep learning to learn the preprocessed features and cannot take advantage of its automatic feature extraction. Therefore, in the next step, we plan to directly use the deep learning model for feature extraction and model training on the original network traffic data, perform the advantages of deep learning in feature extraction, reduce the impact of imbalanced data and achieve more accurate classification.

Further Enhancements

1. Integration of advanced machine learning algorithms.
2. Real-time monitoring and response capabilities.
3. Scalability and performance optimization.
4. Enhanced visualization and reporting features.

REFERENCES

[1] "DDoS Attacks: Evolution, Detection, Prevention, Reaction, and Tolerance" by Abdelkader Lahmadi, Raul Landa, and Guy Pujolle.

[2] "Distributed Denial of Service Attack and Defense" by Roger D. Jones and Richard Boddington.

[3] "Practical Intrusion Analysis: Prevention and Detection for the Twenty-First Century" by Ryan Trost and Max Kilger. (Contains sections on DDoS detection techniques)

[4] "Botnet Detection: Countering the Largest Security Threat" by Wenke Lee, Cliff Wang, and David Dagon.

[5] "Cyber Security Essentials" by James Graham, Richard Howard, and Ryan Olson. (Includes chapters on DDoS attack detection and mitigation)

[6] "Machine Learning for Cybersecurity Cookbook: Over 60 recipes for applying machine learning techniques to enhance your cybersecurity strategies" by Xi Zhang, Junaid Ahmed Ansari, and Abhijit Mohanta - This book offers a collection of recipes for applying machine learning algorithms to various cybersecurity tasks, including DDoS detection.

[7] "Practical Machine Learning for Computer Vision" by Himanshu Singh - While focused on computer vision, this book offers valuable insights into practical machine learning techniques that can be adapted to cybersecurity tasks, including DDoS attack detection.

[8] "Cybersecurity for Industry 4.0: Analysis for Design and Manufacturing" by Lane Thames and Hector M. Diaz - While focusing on cybersecurity for Industry 4.0, this book discusses machine learning techniques applicable to detecting and mitigating DDoS attacks within industrial settings.