iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: 52788

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Machine Learning Algorithms for Intrusion Detection in Cybersecurity

Prof. Dipali Mane[1], Chaitanya Chaudhari[2], Saurabh Shitole[3], Mubin Shaikh[4], Shivam Sashte[5]
[1]*Assistant Professor, Department Of Computer Engineering, ZCOER, Pune*
[2, 3, 4, 5] *BE Students, Zeal College of Engineering and Research, Pune, Maharashtra, India*

***Abstract**: Computer networks and virtual machine security are very necessary in today's time. An Intrusion Detection System (IDS) is a security mechanism designed to monitor computer networks or systems for malicious activities or unauthorized access attempts. The primary function of an IDS is to detect and respond to potential security breaches in real time. Tasks performed by an IDS are anomaly detection, Signature detection, security alert generation, etc… Various researchers are actively working on different ideas for increasing the performance of the IDS. We have used a machine-learning approach for intrusion detection. We have used SVM, Random Forests, and Decision trees for detecting intrusions.*

Keywords: Intrusion Detection System, SVM, Random Forest, Decision Trees, Host Intrusion Detection System, Network Intrusion Detection System.

I. INTRODUCTION

This IDS is exclusively for detecting six types of attacks. which are Denial of service, Bot attack, Web attack, Port Scanning, Brute force, and beingn attack. The IDS helps administrators and users to take preventive measures. It helps to protect the infrastructure from unauthorized users[3].IDS has two main approaches which are anomaly-based intrusion detection and Signature-based intrusion detection.

Signature Based intrusion detection:- Signature-based intrusion detection systems (IDS) are a type of IDS that use pre-defined signatures or patterns to detect known attacks or malicious behaviour. These signatures are created by analyzing past attacks or exploits and identifying their characteristics or behaviours. Once the signatures are created, they can be stored in a database and used to match against incoming traffic in real time.

Anomaly Based intrusion detection:- Anomaly-based intrusion detection systems (IDS) are a type of IDS that detect malicious activities by identifying deviations from normal or expected behaviour. Instead of relying on pre-defined signatures like signature-based IDS, anomaly-based IDS establishes a baseline of normal network or system behaviour and flags any anomalies or deviations from that baseline as potential security threats.

IDS is Broadly classified into Network-based intrusion detection systems and Host Based intrusion detection systems. Both of them work differently from each other and the organization can decide for themselves to choose among them. Each of them serves different purposes and has unique benefits.

II. LITERATURE SURVEY

According to [1] and [11], In this paper, Monika D. Rokade and Yogesh Kumar Sharma used machine learning algorithms like SVM and Naïve Bayes to detect intrusions in a network. They have developed multiple subsystems to detect a single type of attack. The main aim of this paper was to increase the detection rate of attacks using machine learning algorithms like SVM, naïve Bayes, and ANN.

According to [2], In simpler terms, the Therminator wants to make sense of the data on a network. It wants to convert the data into a format that is easy to understand and analyze. By doing this, it becomes easier to detect any strange or suspicious behavior happening on the network. So, the Theeminator helps identify any unusual activities that could potentially be harmful or threatening to the network's security.

According to [3], The authors also conducted tests to measure the performance of the DIDS server when different amounts of data were processed. They wanted to see how well the system handled increasing workloads.

However, the authors acknowledge that there are other important factors to consider. They mentioned issues such as communication delays, the additional work required by blockchain technology, the costs associated with implementing the system, and more. They highlight the need to discuss and analyze these factors further to better understand the overall effectiveness and feasibility of the DIDS system.

According to [4], When it comes to achieving high accuracy and effectively detecting intrusions, it can be challenging to rely solely on classifiers (algorithms that make decisions based on data). However, in the proposed method described in this research, they found a way to improve the performance of three classifiers by implementing a technique called boosting.
Additionally, they found that the Naïve Bayesian classifier was efficient in terms of processing speed.
According to [8], This research focused on using classification algorithms to detect hidden or disguised techniques that attackers use when infiltrating a network. By analyzing the network data, the researchers developed a neural network approach that achieved an average accuracy of 81.73%, with the highest accuracy reaching 95%.
According to [10], Based on the analysis conducted in this research, it can be concluded that using statistical analysis, specifically measures like Mean, Variance, and Standard Deviation, in combination with a technique called Backpropagation Cross-Entropy Artificial Neural Network (ANN), is an effective approach for intrusion detection.

III. REQUIREMENT SPECIFICATION

A. *Hardware Requirement*

1) *Processor (CPU):* Intel 5 or above processor is needed. A multi-core processor or a dedicated high-performance processor is recommended. The SVM algorithm can be computationally intensive, especially when dealing with large datasets or complex feature spaces.
2) *Memory (RAM):* The amount of RAM required depends on the size of your dataset and the complexity of the SVM model. Though a minimum of 4GB of memory is required.
3) *Storage:* A minimum of 40GB of hard disk is required. Sufficient storage is required to store the dataset, feature vectors, and the SVM model.

B. *Software Requirement*

1) *Programming Language:* You'll need a programming language that supports SVM implementation and offers libraries or packages for machine learning. We have used Python programming language.
2) *Integrated Development Environment (IDE):* An IDE can provide a user-friendly development environment with features like code editing, debugging, and project management. We have used PyCharm, Anaconda (with Jupyter Notebook), and sometimes Visual Studio Code.

IV. METHODLOGY

In our research, we used various machine learning algorithms for detecting the intrusion. After pre-processing our dataset we kept five features for training or model for detecting the intrusion. These five features are Total_Fwd_Packets, Total_Backward_packets,Down_Up_Ratio,Act_Data_Pkt_fwd,Min_seg_size_fwd.
The model will take the value of these five features and will show which type of intrusion has been detected. It will show an alert message to the administrator or security analysts.

A. *Total_Fwd_Pkts*

"Total_Fwd_Packets" is a term typically used in network analysis and refers to the total number of packets that have been forwarded by a network device or system. In computer networking, data is divided into smaller units called packets for transmission over a network. These packets contain the necessary information for data transfer, including the source and destination addresses, data payload, and control information. When monitoring network traffic, analyzing the "Total_Fwd_Packets" metric can provide insights into the volume of packets being forwarded by a specific device or system. By tracking the total number of forwarded packets, network administrators can evaluate the load on a network device and make informed decisions to optimize network resources or troubleshoot any issues that may arise.

B. *Total_Backward_Pkt*

"Total_Backward_Packets" is a term used in network analysis to represent the total number of packets that have been sent in the backward direction by a network device or system.
While "Total_Fwd_Packets" represents the count of packets forwarded in the forward direction, "Total_Backward_Packets" focuses on packets sent in the opposite or backward direction. These packets typically contain responses, acknowledgments, or other types of network traffic that are sent back from the destination device to the source.

By analyzing the number of backward packets, network administrators can gain insights into the behavior of network devices, detect anomalies or potential issues, and optimize network performance accordingly.

C. Down_Up_Ratio:

"Down_Up_Ratio" refers to the ratio of download speed to upload speed in a network connection. It is often used to describe the asymmetry of a broadband or internet connection, where the download speed (data received from the internet) is typically higher than the upload speed (data sent to the internet).

The Down_Up_Ratio is an important factor to consider when evaluating the performance and suitability of an internet connection for specific applications.

In the context of IDS, the focus is on analyzing network traffic patterns, identifying suspicious or anomalous behaviour, and detecting potential security threats. The IDS dose not directly deal with measuring or evaluating network performance metrics such as download and upload speeds.

D. Act_Data_Pkt_fwd

"Act_Data_Pkt_fwd" is a term that could potentially refer to the count or number of actual data packets forwarded in a network. In network analysis, data packets are divided into different types, including control packets and data packets. Control packets are responsible for managing and controlling network traffic, while data packets carry the actual payload or information being transmitted. The "Act_Data_Pkt_fwd" metric specifically focuses on the count of data packets that have been successfully forwarded in the forward direction by a network device or system. It excludes control packets, acknowledgments, and other non-data packets. Monitoring the "Act_Data_Pkt_fwd" metric can provide insights into the volume of actual data being transmitted in a network, separate from control or management traffic. This metric can be useful for assessing network performance, measuring data transfer efficiency, identifying congestion or bottlenecks, and optimizing network resources.

By tracking the count of forwarded data packets, network administrators can gain a better understanding of the actual data flow within the network, identify any anomalies or issues, and take appropriate actions to ensure smooth and efficient data transmission.

E. Min_seg_size_fwd

"min_seg_size_fwd" refers to the minimum segment size for forwarded packets in a network. In computer networking, data is divided into smaller units called segments for transmission over a network. The segment size is determined by various factors, including network protocols, configurations, and performance considerations.

By setting a minimum segment size, IDS systems can filter out small or fragmented packets that may not contain enough meaningful data to indicate malicious activity. This parameter helps to reduce false positives by focusing on larger segments that are more likely to carry meaningful network traffic.

We have used three algorithms for comparisons. These algorithms are SVM, Random Forest, and Decision Trees.

F. SVM

It is widely used in IDS for the classification of intrusions.[11], One of the most popular supervised learning algorithms, Support Vector Machine, or SVM, is used to solve Classification and Regression problems. However, it's largely employed for Machine Learning Classification problems.

SVM selects the sharp vectors and points that make it easier to create the hyperplane. The algorithmic software is called a "Support Vector Machine" because these extreme examples are known as support vectors. Consider the diagram below, in which there are two distinct classes that are separated by a call boundary or hyperplane.

G. Random Forest

It is a Machine Learning algorithm used for classification and regression problems. It has robustness, it can handle high dimensional data.

It combines the predictions of many decision trees. The algorithm works by creating a forest of decision trees, where each tree is trained on a random subset of the data and a random subset of the features. During training, each tree in the forest learns from the data and makes predictions independently.

H. Decision Trees

Decision trees are a popular machine learning algorithm used for both classification and regression tasks. They provide a structured and intuitive approach to making predictions based on a set of input features.

The decision tree algorithm learns from the data by finding the best features and thresholds to split the data at each node, aiming to maximize the information gain or decrease the impurity of the data. Different impurity measures can be used, such as Gini impurity or entropy, depending on the task at hand. It can handle non-linear Relationships.

V. IMPLEMENTATION

1) *Dataset Preparation:* To train an SVM model for intrusion detection, a labelled dataset is required. The dataset typically consists of network traffic data, where each data point is associated with a label indicating whether it is normal or malicious. The data can include features such as source and destination IP addresses, port numbers, packet sizes, protocol types, etc.
2) *Feature Extraction:* Before training an SVM model, it's common to perform feature extraction or selection to identify relevant and informative features from the dataset. Various techniques such as statistical analysis, information gain, or principal component analysis (PCA) can be applied to extract the most discriminative features.
3) *Data Pre-Processing:* Once the features are extracted, the dataset needs to be pre-processed. This step involves normalizing or scaling the features to ensure they are on a similar scale, which can improve the SVM's performance. Additionally, data cleaning techniques like removing duplicates, handling missing values, and handling outliers may be applied.
4) *Model Training:* After pre-processing, the dataset is divided into training and testing sets. The training set is used to train the SVM model by finding an optimal hyperplane that separates the normal and malicious instances with a maximum margin. SVM employs a kernel function, such as linear, polynomial, or radial basis function (RBF), to transform the data into a higher-dimensional feature space if necessary. The appropriate kernel and its parameters are selected through cross-validation or grid search.
5) *Model Evaluation:* Once the SVM model is trained, it is evaluated on the testing set to assess its performance. Common evaluation metrics for intrusion detection include accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC AUC). The performance of the model can be further enhanced through techniques like ensemble learning or adjusting the decision threshold based on the application's requirements.
6) *Deployment and Monitoring:* After the SVM model demonstrates satisfactory performance, it can be deployed in a production environment for real-time intrusion detection. The model continuously analyzes the network traffic and classifies instances as normal or malicious. It's crucial to monitor the performance of the deployed model regularly and update it as new data becomes available to adapt to evolving attack patterns

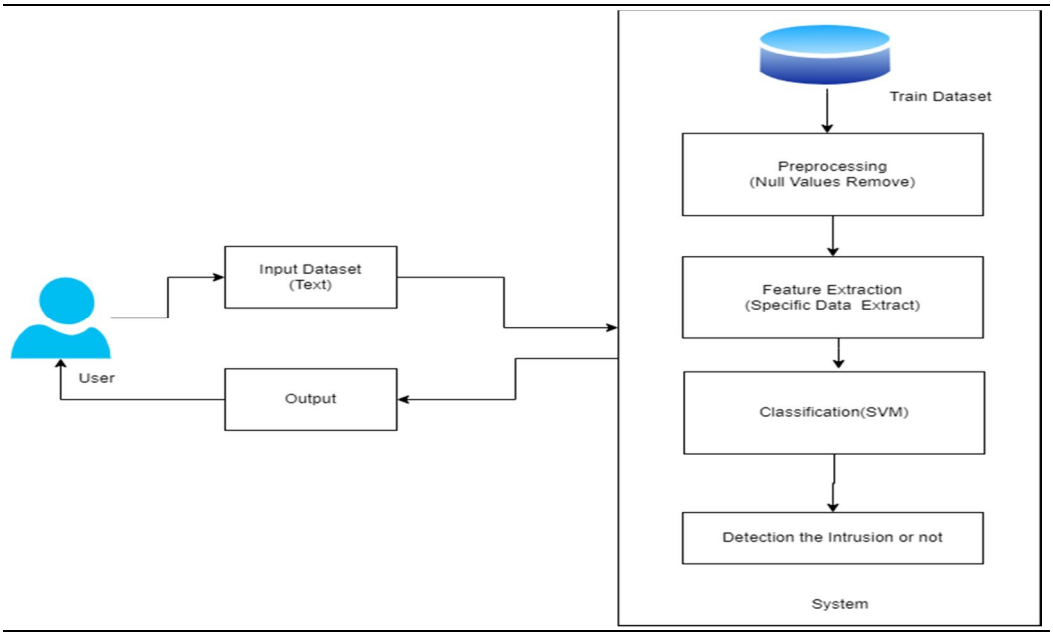


Fig 1: System Architecture

VI. RESULTS AND DISCUSSIONS

After the successful implementation of the system, we calculated its confusion matrix.

TABLE 1: Performance Evaluation With SVM, RF, and DT

	SVM	RF	DT
Accuracy	0.96	0.94	0.92
Precision	0.96	0.90	0.88
Recall	0.94	0.88	0.85
F1-Score	0.95	0.85	0.86

We can see that SVM has better accuracy than Random Forests and Decision Trees. It is also Robust against Overfitting. It is effective with small training sets. and it is effective with high-dimensional spaces.

By learning from labelled data, SVM-based IDS can accurately classify network traffic into normal and malicious categories. This leads to reliable detection of various types of attacks, such as DoS, DDoS, intrusion attempts, or malware communication. It Reduced False Positives: False positives occur when legitimate network traffic is incorrectly identified as malicious. SVMs provide an interpretable decision boundary, which can help in understanding the impact of different features on the classification decision. Decision trees and Random forests are sensitive to noise and outliers potentially leading to overfitting.

We have to add the values of how many packets are flowing in the system.and it will give us in result that which type of attack is detected.

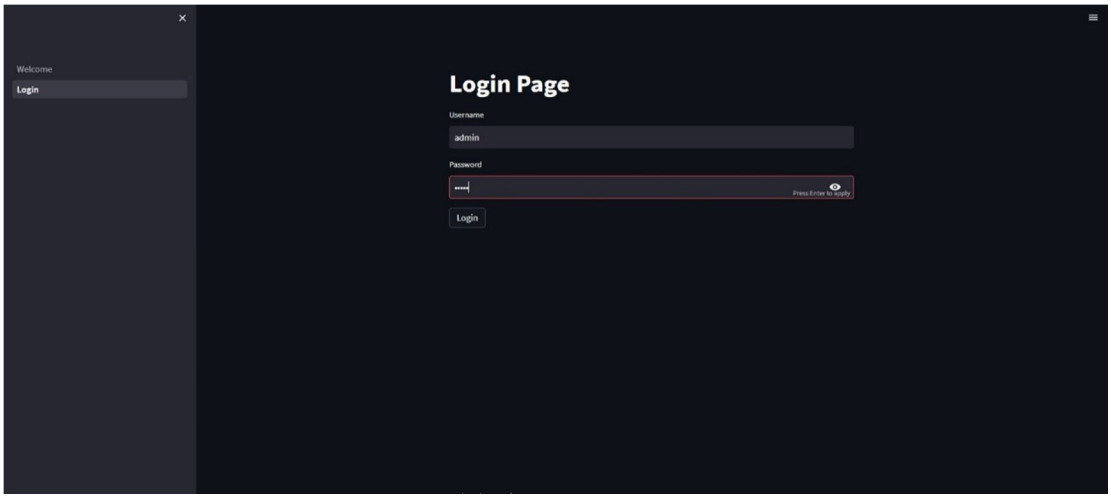


Fig 2: Login Page

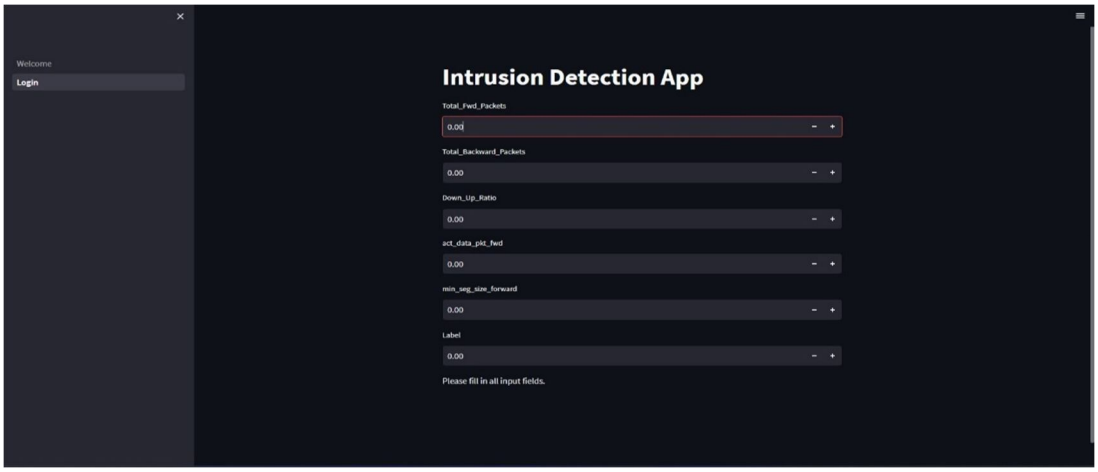


Fig 3: interface for Intrusion Detection web-app

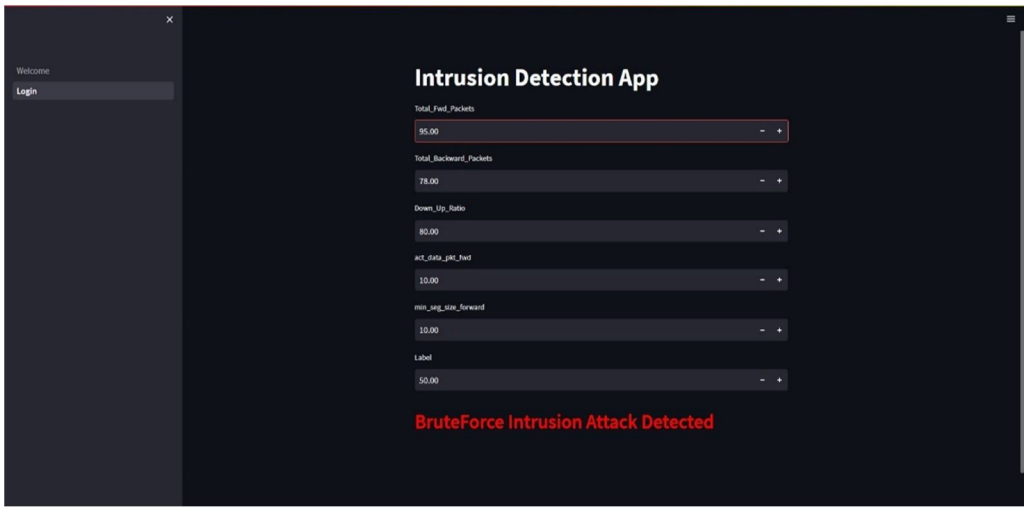


Fig 4: BruteForce intrusion detected

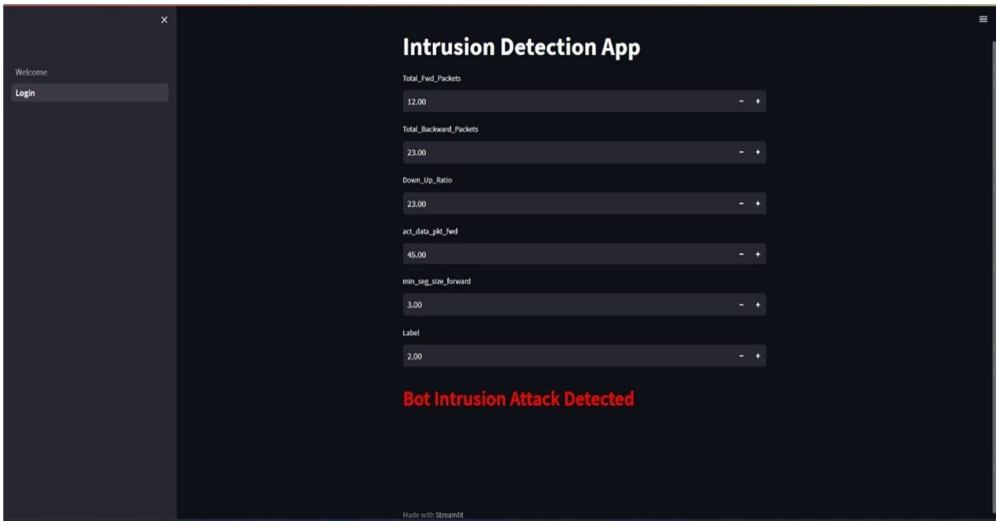


Fig 5: Bot attack detected

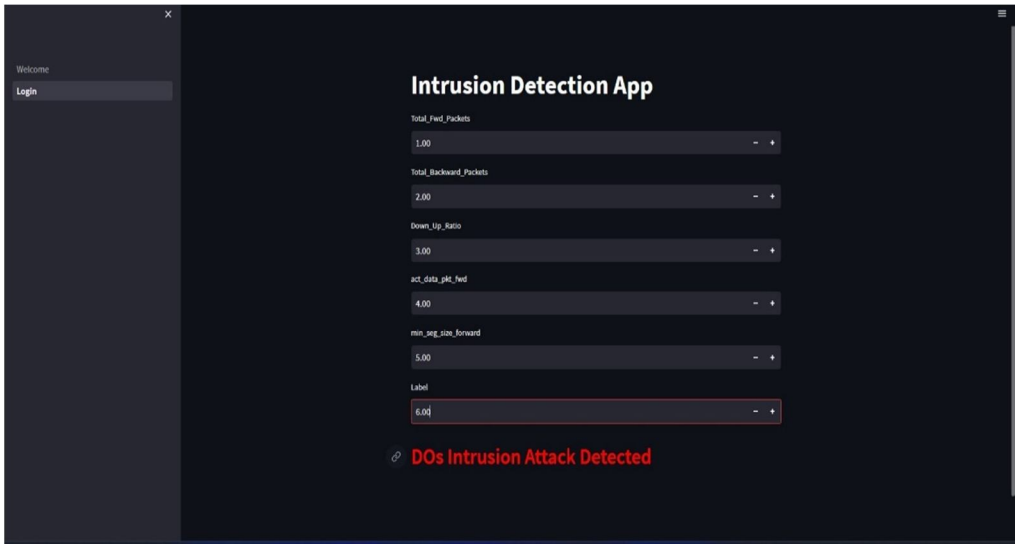


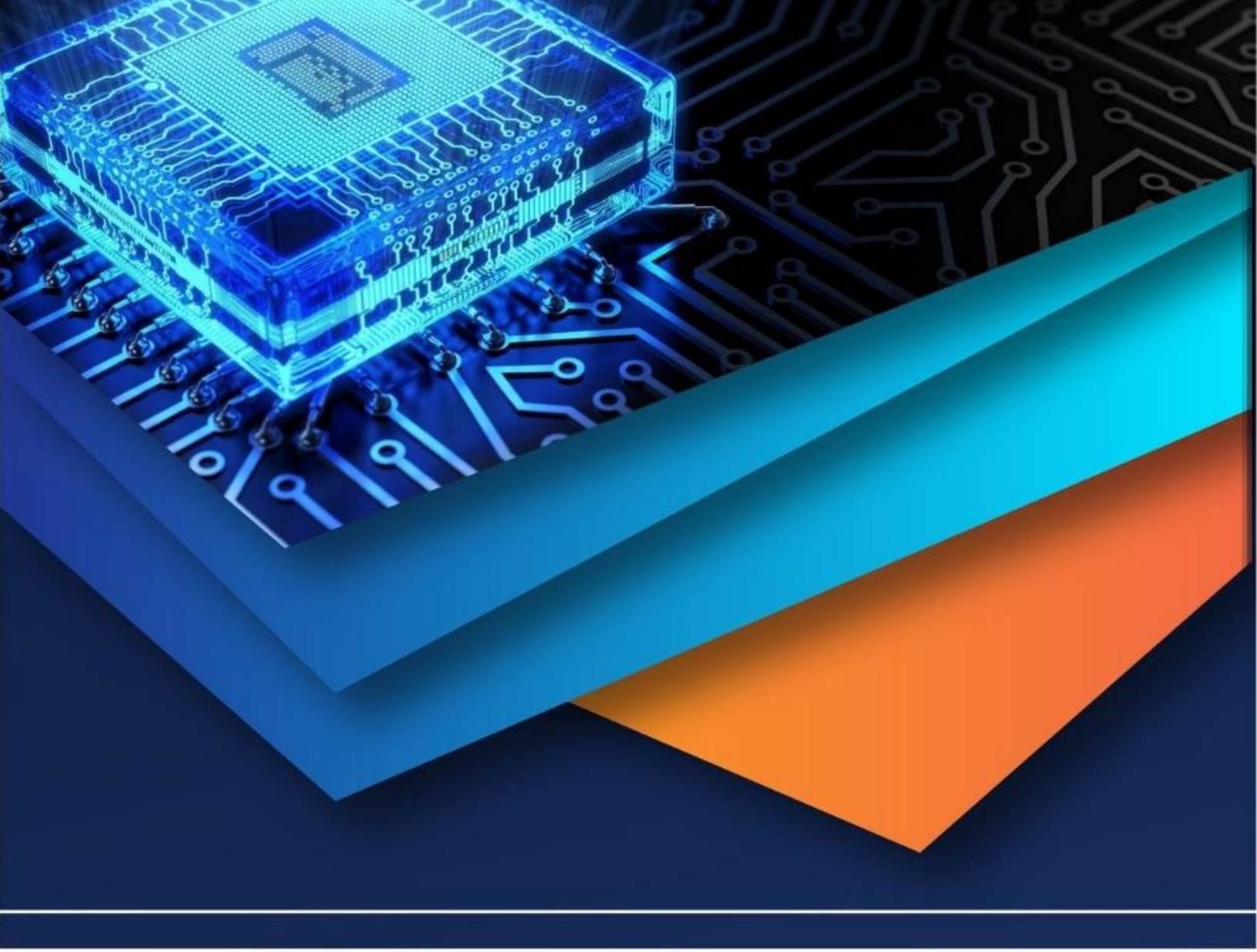
Fig 6: Dos attack detected

VII. CONCLUSION

As we see how the crime rate is increased, including cybercrime over the years and new attacks and viruses are being developed. Implementing an IDS to protect organizations and users from these attacks is necessary. By leveraging the power of machine learning algorithms, such as Support Vector Machines (SVM), Random forests, or Decision Trees, organizations can achieve improved detection accuracy, reduced false positives, and enhanced capabilities for handling complex and evolving attacks**.**

REFERENCES

[1] Monika D.Rokade, Yogesh Kumar Sharma," MLIDS: A Machine Learning Approach for Intrusion Detection for Real Time Network Dataset",2021
[2] Stephen D. Donald, Rabert V McMillen, David K Fard, and John C. McEachen," Therminator 2: a thermodynamics-based method for real-time patternless intrusion detection",
[3] Dr. Manish Kumar, Ashish Kumar Singh," Distributed Intrusion Detection System using Blockchain and Cloud Computing Infrastructure",2020
[4] S. Sivanantham, R. Abirami, R. Gowsalya," Comparing the Performance of Adaptive Boosted Classifiers in Anomaly-based Intrusion Detection System for Networks",2019
[5] Meng W, Tischhauser E W, Wang Q, et al. When Intrusion Detection Meets Blockchain Technology: A Review[J]. IEEE Access, 2018
[6] Zakiyabanu S. Malek, Bhushan Trivedi, Axita Shah," User behaviour Pattern -Signature based Intrusion Detection",2020
[7] Zhan Xin, Wang Xiaodong, Yuan Huabing," Research on Block Chain Network Intrusion Detection System",2019
[8] Ajay Shah, Sophine Clachar, Manfred Minimair, Davis Cook," Building Multiclass Classification Baselines for Anomaly-based Network Intrusion Detection Systems",2020
[9] Dong YuanTong," Research of Intrusion Detection Method Based on IL-FSVM",2019
[10] Sarwar Wasi, Dr.Sarmad Shams, Shahzad Nasim,"Intrusion Detection Using Deep Learning and Statistical Data Analysis",2019
[11] Dipali Mane, Chaitanya Chaudhari, Shivam Sashte, Mubin Shaikh, Saurabh Shitole.," A Machine Learning Approach For Intrusion Detection",2023



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY