International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 03 Mar. 2024, pp: 340-346 www.ijaem.net ISSN: 2395-5252

Detection of Dos Attack Using Machine Learning Approach

G.Tanmayi , S.Sree Raj, K.Sree Ram, D.Surya Prakash, G.Tanuja, Dr. J. Satish Babu
Computer Science Engineering , K L University
Vijayawada, Andhra Pradesh

Date of Submission: 11-03-2024 Date of Acceptance: 21-03-2024

ABSTRACT— As a global platform for linking people and businesses, the internet has become a crucial component of contemporary life. As a result of the growing However, with the increasing dependence on the internet comes a growing threat of cyber-attacks.Denial of service (DoS) attacks, one of the most common types of cyberattacks, have the ability to cause significant damage to businesses and organizations. Machine learning is a promising approach for detecting DoS attacks. A DoS attack may be indicated by aberrant patterns that machine learning systems are able to identify by analyzing vast volumes of network traffic data. In order to identify any deviations from the norm that can be a sign of an attack, the algorithm can be trained on past data to understand the features of typical network traffic. Random Forest is one of the most well-liked machine learning techniques for spotting DoS assaults. This algorithm can effectively detect anomalies in large datasets and can be trained to accurately classify network traffic as normal or malicious. Combining Random Forest with other machine learning algorithms, such as Naive Bayes, can further improve the accuracy and performance of the detection system. As the internet continues to grow and becomes more integrated into our lives, the threat of cyber-attacks will continue to increase. However, with the use of machine learning algorithms, we can improve our ability to detect and respond to these attacks, helping to protect businesses, organizations, and individuals from the devastating consequences of DoS attacks.

I. INTRODUCTION

Cybercriminals that use denial of service (DoS) attacks try to render a server, network, or website inoperable by flooding it with traffic. [4]. This type of attack is one of the most common and is a significant threat to businesses, organizations, and individuals who rely on the internet for their daily operations.DoS attacks can be executed in several methods, including delivering erroneous packets, overloading the network with data, or taking advantage of security holes in the network architecture. These attacks can cause significant damage, including loss of revenue, reputation, and customer trust. Detecting DoS attacks is crucial to preventing significant damage.Identifying patterns in network traffic that could point to an attack is a traditional approach of detecting denial-of-service (DoS) attacks. However, these methods can be time-consuming and may not be effective in detecting sophisticated attacks. A viable method for identifying DoS attacks is machine learning. Machine learning algorithms can scan vast volumes of network traffic data and identify patterns that might indicate an attack. The algorithm can become more accurate at detecting threats by learning to differentiate between legitimate and malicious traffic using past data for training [9]. There are several existing works on detecting DoS attacks using machine learning. Decision tree-based algorithms, such Random Forest and C4.5, for instance, have demonstrated efficacy in identifying denial-of-service (DoS) assaults because of their capacity to process enormous datasets and identify intricate patterns. In addition to neural networks, support vector machines, and K-Nearest Neighbors, machine learning techniques have been used to identify denial-of-service attacks.Among the most often used machine learning algorithms for DoS attack detection is Random Forest. This approach, known as ensemble learning, constructs several decision trees and aggregates their outcomes to enhance precision.Random Forest is effective at identifying complex patterns in large datasets and can provide high accuracy in detecting anomalies [13].

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 03 Mar. 2024, pp: 340-346 www.ijaem.net ISSN: 2395-5252

II. LITERATURE SURVEY:

In this paper, the author acknowledges the inevitability and unpredictability of DOS attacks, emphasizing their potentially irreversible impact. To mitigate this, the proposed classification and detection approach, using SVM and C 4.5 Supervised Algorithms for Learning with the NSL_KDD Dataset efficiently and accurately identifies DOS attacks in minimal time. [1]

In this paper,to efficiently detect application layer DoS assaults, the author presents Neural Network and Machine Learning techniques, including Random Forest and MLP. In terms of accuracy, the Random Forest algorithm performs better than MLP, according to the findings. However,note that only the Benign and DoS attack categories are currently included in the proposed system's classification of the CIC IDS 2017 dataset, leaving room for future work involving feature reduction and multiclassification of DoS attacks like Heartbleed, http flood and other. [3]

In this paper, Using DARF'A intrusion evaluation data, the author describes how Support Vector Machines (SVMs) are implemented for DOS pattern detections excel in scalability, training time, running time, and detection accuracy compared to additional methods of machine learning, such as neural networks. They consistently achieve high detection accuracy, surpassing 99%, even when considering feature ranking, suggesting the potential for customizable feature sets in IDS for DOS detection. [5]

In this paper, the author's overarching goal is to explore the data characteristics influencing naive Bayes' performance. They employ Monte Carlo simulations to systematically study classification accuracy across different problem classes and delve into the impact of distributed entropy of categorization mistakes. Surprisingly, they discover that naive Bayes performs poorest with features that are in between fully independent and functionally dependent features. The study highlights that accuracy is more related to the loss of class-related information under the naive Bayes model rather than the degree of feature dependencies, calling for further empirical and theoretical investigation into this relationship. [7]

In this paper, the author conducts a literature review on the use of deep learning (DL) and machine learning (ML) methods in addressing cyber security challenges. These methods, inspired by human brain learning, are increasingly utilized across various research domains to tackle evolving cyber threats. The review emphasizes recent advancements in DL and ML tools, platforms, and their effectiveness in providing security solutions for diverse categories of cyberattacks in today's internet-centric landscape. [8]

The author reviews recent research on the application of deep learning (DL) in this paper. with machine learning (ML)in network security, particularly focusing on intrusion detection methods. The review underscores the challenge of establishing a definitive best method due to the unique advantages and disadvantages of each approach. Additionally, it highlights the importance of quality datasets that solve the shortcomings of the current public datasets and are used to train these ML and DL models. [10]

In this paper, the author introduces a combined method for identifying lymph illnesses that uses Random Forest Classifier (RFC) and Genetic Algorithms (GA). The lymph diseases dataset's dimension is decreased using GA, and intelligent classification is accomplished using RFC. The system aims to capitalize on RFC's strengths, including superior generalization, rapid learning, and minimal parameter tuning. Comparative evaluations withthe suggested GA-RFC strategy is effective, as evidenced by the high accuracy, sensitivity, specificity, and AUC values that are obtained when other feature selection techniques are combined with RFC. The study suggests the potential application of this approach in other medical diagnosis scenarios and explores alternative classification algorithms with optimization techniques for furtherresearch. [11]

In this paper, the author's analysis reveals that combining Naïve Bayes with Random Forest yields the best results in their experiments. Conversely, when combined with KNN or KNN+NB, the error rates remain similar to those of KNN alone, indicating that KNN is dominant when integrated with any other classification method. [12]

This study examines how three alternative data types—text only, text plus numeric, and numeric only—affect the effectiveness of classifiers built with the Random Forest, k-NearestNeighbourr (KNN), and Naive Bayes (NB) algorithms. They investigate categorization issues in terms of parameter changes and mean accuracy across several dataset types from UCI.. Findings indicate that Random Forest and KNN classifiers perform similarly, with the numeric dataset producing the best results. Naïve Bayes demonstrates lower mean accuracy, potentially due to underlying dataset attribute dependencies, as it assumes attribute independence. The study underscores the significance of algorithm selection based on specific application requirements and suggests further exploration using parametric methods across multiple datasets. [15]

CHALLENGES IN DETECTION DoS ATTACK

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 03 Mar. 2024, pp: 340-346 www.ijaem.net ISSN: 2395-5252

Detecting a denial-of-service (DoS) attack may be difficult since attackers employ a variety of strategies to avoid detection, hide their identities, and distribute their attacks across multiple sources. Some of the common challenges associated with DoS attack detection are:

- Traffic volume variability: It can be difficult to distinguish between legitimate traffic and a DoS attack, especially when the attack traffic is distributed across multiple sources and varies in volume.
- Packet fragmentation: Attackers may use packet fragmentation to evade detection, by splitting up the attack traffic into smaller packets that are more difficult to detect.
- Spoofing: Attackers can use IP address spoofing to make it appear as if the attack traffic is coming from a legitimate source, making it difficult to distinguish between legitimate and malicious traffic.
- Slow-rate attacks: Some DoS attacks occur at a slow rate that may not be immediately noticeable, making it difficult to detect and respond to the attack in a timely manner.
- Encrypted traffic: Attackers may use encrypted traffic to evade detection, making it difficult to analyze and identify malicious traffic.
- False positives: DoS detection systems may generate false positives, alerting administrators to an attack that is not actually occurring, which can waste resources and divert attention from real attacks.
- Targeted attacks: Attackers may target specific applications or services on a network, making it difficult to detect attacks that are not affecting the entire network.

III. MACHINE LEARNING:

By examining patterns in network data and spotting anomalies that point to malicious behavior, machine learning can be used to detect cyberattacks. [8].

The field of machine learning is wide and includes a wide range of algorithms and methods. These are a few typical forms of machine learning.:

A. Supervised learning: Training a model on labeled data—where the inputs and outputs are known—requires supervised learning. Learning a function that can forecast the result for novel inputs is the aim. Time-series forecasting, regression, and classification are a few typical applications of supervised learning.

B. Unsupervised learning: This entails using unlabeled data—inputs without corresponding output labels—to train a model. Finding structures, patterns, or clusters in the data is the aim. A few typical applications of unsupervised learning are anomaly detection, dimensionality reduction, and clustering.

C. Semi-supervised learning: The Bayes theorem, a statistical theory, states the likelihood that an event will occur given prior knowledge of relevant conditions. with the intention of utilizing the unlabeled data to increase the model's accuracy. When access to labeled data is scarce or costly, semi-supervised learning may be helpful.

D. Reinforcement learning: To do this, a model must be trained to make decisions depending on input from the environment. Finding a policy that optimizes a reward signal, like a profit or score, is the aim. Common applications of reinforcement learning include control systems, robotics, and gaming.

E. Deep learning: This involves training deep neural networks with multiple layers of nonlinear transformations, using techniques such as backpropagation and gradient descent. Deep learning is particularly effective for applications like audio and picture recognition, natural language processing, and generative modeling.

F. Transfer learning: This involves reusing pre-trained models or features for a new task, to leverage the knowledge learned from previous tasks. Transfer learning can be useful when the new task has limited labeled data or is like previous tasks.

IV. NAÏVE BAYES THEOREM:

One well-liked machine learning approach for cybersecurity attack detection is Naive Bayes. A statistical theory known as the Bayes theorem expresses the probability of an event happening given prior knowledge of pertinent conditions. serves as the foundation for the algorithm. The algorithm works by learning the probabilities of different features in the network traffic data, including the protocol type, port number, source, and destination IP addresses, and so forth. It then uses these probabilities to calculate the likelihood of a given network traffic sample being either normal or malicious traffic.

Two scenarios yield the best results with Naïve Bayes: fully independent features and functionally dependent features. [7]. The network traffic data's various properties are assumed by the Naive Bayes algorithm to be independent of one another. This assumption is often unrealistic in practice, but it can still be effective for detecting certain types of attacks, such as spam emails or phishing attempts. The algorithm is also relatively simple and computationally efficient, which makes it a popular choice for real-time attack detection in cybersecurity.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 03 Mar. 2024, pp: 340-346 www.ijaem.net ISSN: 2395-5252

To use Naive Bayes for attack detection, you would need to train the algorithm using a dataset of network traffic samples labeled as either normal or malicious. The algorithm would then use this training data to learn the probabilities of different features in the network traffic data, which it would use to classify new network traffic samples as either normal or malicious. It is based on Bayes' theorem, which, given past knowledge of potential event-related factors, expresses the likelihood that an event will occur.

The Naïve Bayes formula is expressed as follows: [7]

$$P(A|B) = P(B|A) * P(A)/P(B)$$

The product of the prior probability of class cj and the probability of each feature given class cj yields the likelihood of an instance d belonging to class cj in this manner. divided by the probability of the instance d occurring. The prior probability represents the probability of observing class cj without any knowledge of the features, while the probability of each feature given class cj represents the probability of observing the specific value of each feature, given that the instance belongs to class cj. This formula still assumes independence between the features, but it explicitly includes the prior probability of class cj.

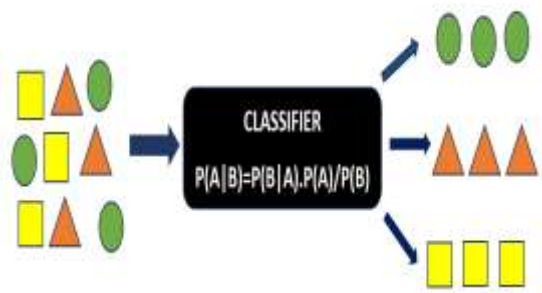

Figure 1: Naive Bayes Classifier

V. RANDOM FOREST ALGORITHM:

Known for its efficiency in high-dimensional classification and skewed problems, Random Forests are a popular ensemble learning technique in pattern recognition and machine learning. The Random Forests classifier (RFC) is recognized as one of the most successful algorithms in this field, with a proven track record of delivering accurate and reliable results.[11]

Using random selections of the training data and features, the approach creates several decision trees. Because each decision tree in the forest is trained using a different subset of the data, overfitting is less prevalent and the model's ability to generalize to new data is improved. [12]

The training phase involves building each decision tree in the forest by repeatedly splitting the data into subgroups based on the feature values. The method identifies which characteristic is best for separating the data at each node of the tree based on a criterion like information gain or Gini impurity. The product is a set of decision trees that are useful for forecasting fresh data.

The program combines the forecasts from each decision tree in the forest to create a prediction using a random forest model. The algorithm chooses the class from the decision trees that has the most votes for classification jobs. The algorithm determines the mean or median of the values that the decision trees predict for regression tasks.[12]

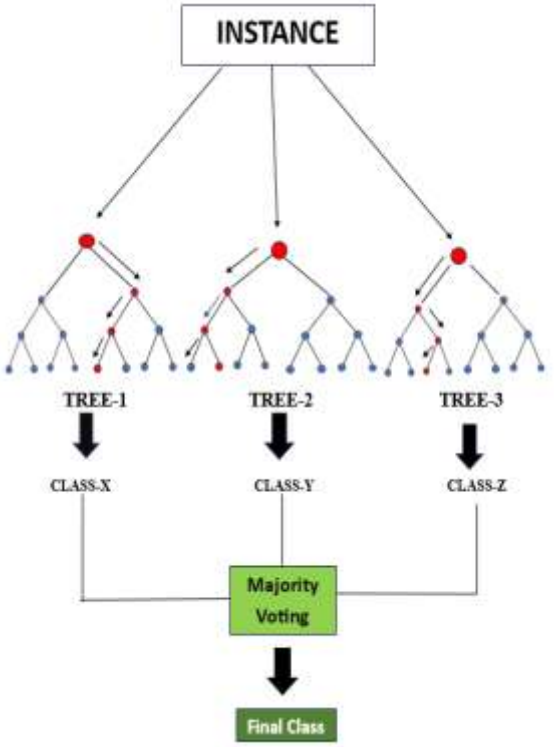

Figure 2: Random Forest Algorithm

COMPARATIVE STUDY:

In the field of network security and intrusion detection, the CIC IDS 2017 dataset is a commonly utilized benchmark. It offers an all-inclusive set of network traffic data that includes both regular traffic and other kinds of network attacks. The purpose of the dataset is to support researchers and practitioners in the creation and assessment of network security and intrusion detection systems (IDS).

An essential indicator for assessing the effectiveness of the Random Forest (RF) algorithm in the context of intrusion detection systems (IDS) is its

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 03 Mar. 2024, pp: 340-346 www.ijaem.net ISSN: 2395-5252

accuracy on the CIC IDS 2017 dataset. One popular benchmark dataset that includes network traffic data with different kinds of network assaults and regular traffic is the CIC IDS 2017 dataset.

A group of decision trees is used by the well-known machine learning method Random Forest to generate predictions. It is well-known for handling complicated and big datasets, which qualifies it for IDS applications.

In order to assess the precision of Random Forest on the CIC IDS 2017 dataset, researchers frequently employ techniques like cross-validation, which separate the dataset into training and testing subsets. The testing subset is used to compare the predicted and true labels in order to determine whether the model is accurate. The training subset is used to train the model.

VI. EXPERMENTAL RESULTS:

Through experiments in the CIC IDS 2017 dataset, Random-Forest provides superior results in identifying DoS attacks. A. Suggested Approach The suggested method for classifying DoS attacks consists of the following steps. [/3]

Step 1: The system takes as input the entire CIC IDS 2017 Wednesday dataset, replete with all attributes.

Step 2: Weka, a well-known machine learning program, is utilized for simulation.

Step 3: The suggested system classifies traffic into benign and denial-of-service attacks using machine learning methods.

Step 4: The preprocessing phase involves using a specific percentage of data to train the algorithm.

Step 5: Finally, simulations of machine learning and neural network classifiers like RF and MLP are performed on the dataset to categorize it into DoS and benign attacks.

Sr. No	Training Records	Records Tested	Accuracy
1	40959(20%)	77744	98.3691%
2	61439(30%)	67979	98.8783%
3	81918(40%)	58181	98.5099%
4	102398(50%)	48535	98.8760%
5	122878(60%)	38844	98.8818%
6	143357(70%)	29152	98.8956%
7	163837(80%)	19429	98.8956%

TABLE-1: ACCURACY OF MLP ON CIC IDS 2017 DATASET

Sr. No	Training Records	Records Tested	Accuracy
1	40959(20%)	77744	99.9194%
2	61439(30%)	67979	99.9268%
3	81918(40%)	58181	99.9308%
4	102398(50%)	48535	99.9502%
5	122878(60%)	38844	99.9475%
6	143357(70%)	29152	99.9512%
7	163837(80%)	19429	99.9563%

TABLE-2:ACCURACYOF RF ON CIC IDS 2017 DATASET

PROPOSED METHODOLOGY:

Combining Random Forest with Naive Bayes can improve the detection of Attacks using denial of service (DoS) leveraging the strengths of both algorithms [12].

But compared to the MLP technique, the Random Forest approach offers greater accuracy. [2].

Random Forest is effective at identifying complex patterns in large datasets and can provide high accuracy in detecting anomalies. However, it may struggle to handle high-dimensional data, which can be a problem in some cybersecurity applications.

Naive Bayes, on the other hand, is particularly effective at handling high-dimensional data and is known for its simplicity, speed, and ability to handle missing data. It assumes independence among features, which can limit its effectiveness when dealing with complex relationships between variables.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 6, Issue 03 Mar. 2024, pp: 340-346 www.ijaem.net ISSN: 2395-5252

By combining the two algorithms, we can take advantage of the strengths of each approach. The Random Forest algorithm can be used to pre-select the most relevant features from the high-dimensional dataset, reducing the dimensionality and complexity of the problem. The selected features can then be fed into a Naive Bayes classifier, which can effectively handle the remaining features and classify the traffic as normal or malicious.

In this approach, Utilizing the Random Forest technique as a feature selector lowers computing cost and improves the Naive Bayes classifier's performance. This approach can lead to improved detection accuracy of DoS attacks, especially When working with data that is high-dimensional. Additionally, the combination of the two algorithms can provide robustness to the model by reducing the risk of overfitting and improving generalization performance on new data.

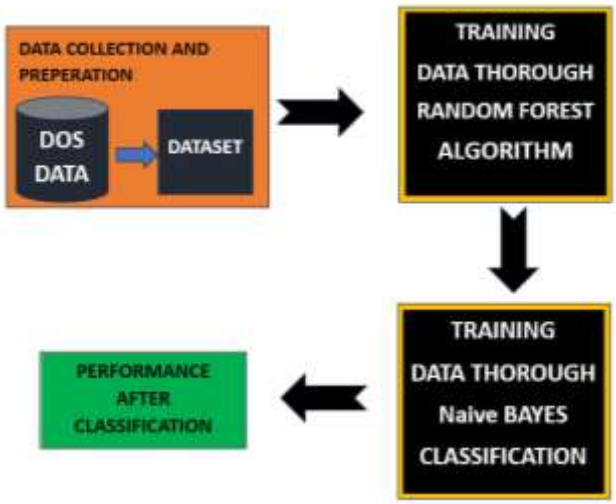

Figure 3: PROPOSED METHODOLOGY

VII. CONCLUSION AND FUTURE WORK:

In conclusion, DoS (Denial of Service) attacks pose a significant threat in the realm of cybersecurity, making their detection crucial for maintaining network integrity and availability. Machine learning methods have proven to be effective in identifying and mitigating such attacks. Among these methods, Random Forest has demonstrated promising capabilities in detecting DoS attacks.

Considering the potential for further improvement, combining Random Forest with the Naive Bayes theorem holds the promise of achieving even better results in DoS detection. The Naive Bayes theorem is a probabilistic classification technique known for its simplicity and efficiency. By integrating its probabilistic reasoning with the ensemble approach of Random Forest, it is anticipated that the combined model will enhance accuracy and provide more robust detection capabilities.

To validate this hypothesis, we plan to conduct comprehensive experiments and evaluate the performance of the combined Random Forest and Naive Bayes approach on detecting DoS attacks. The results of these experiments will be documented in our upcoming research paper.

By exploring the synergy between Random Forest and Naive Bayes, our goals are to improve DoS detection techniques and offer insightful information on how well coupled machine learning techniques work. Ultimately, this research endeavor seeks to enhance network security measures and bolster defense mechanisms against DoS attacks, ensuring a safer digital environment for organizations and individuals alike.

REFERENCE:

[1]. Shinde, Poonam Jagannath, and Madhumita Chatterjee. "A novel approach for classification and detection of dos attacks." 2018 International Conference on Smart City and Emerging Technology (ICSCET). IEEE, 2018. [1]

[2]. Habib, Ahsan, and Debashish Roy. "Steps to defend against DoS attacks." 2009 12th International Conference on Computers and Information Technology. IEEE, 2009. [2]

[3]. Wankhede, Shreekh, and Deepak Kshirsagar. "DoS attack detection using machine learning and neural network." 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA). IEEE, 2018. [3]

[4]. Biju, Jibi Mariam, Neethu Gopal, and Anju J. Prakash. "Cyber-attacks and its different types." International Research Journal of Engineering and Technology 6.3 (2019): 4849-4852. [4]

[5]. Mukkamala, Srinivas, and Andrew H. Sung. "Detecting denial of service attacks using support vector machines." The 12th IEEE International Conference on Fuzzy Systems, 2003. FUZZ'03. Vol. 2. IEEE, 2003. [5]

[6]. Liang, Lulu, et al. "A denial-of-service attack method for an iot system." 2016 8th international conference on Information Technology in Medicine and Education (ITME). IEEE, 2016. [6]

[7]. Rish, Irina. "An empirical study of the naive Bayes classifier." IJCAI 2001 workshop on empirical methods in artificial intelligence. Vol. 3. No. 22. 2001. [7]

[8]. Geetha, R., and T. Thilagam. "A review on the effectiveness of machine learning and deep learning algorithms for cyber security." Archives of Computational Methods in Engineering 28 (2021): 2861-2879. [8]

[9]. Shaukat, Kamran, et al. "Cyber threat detection using machine learning techniques: A performance evaluation perspective." 2020 international conference on cyber warfare and security (ICCWS). IEEE, 2020. [9]

[10]. Xin, Yang, et al. "Machine learning and deep learning methods for cybersecurity." Ieee access 6 (2018): 35365-35381. [10]

[11]. Azar, Ahmad Taher, et al. "A random forest classifier for lymph diseases." Computer methods and programs in biomedicine 113.2 (2014): 465-473. [11]

[12]. Devi, R. Gayathri, and P. Sumanjani. "Improved classification techniques by combining KNN and Random Forest with Naive Bayesian classifier." 2015 IEEE international conference on engineering and technology (ICETECH). IEEE, 2015. [12]

[13]. Breiman, Leo. "Random forests." Machine learning 45 (2001): 5-32 [13]

[14]. Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorba-ni, Intrusion Detection Evaluation Dataset (CICIDS2017), Ca-nadian Instituteof Cybersecurity, https://www.unb.ca/cic/datasets/ids-2017.html [14]

[15]. Singh, Asmita, Malka N. Halgamuge, and Rajasekaran Lakshmiganthan. "Impact of different data types on classifier performance of random forest, naive bayes, and k-nearest neighbor's algorithms." International Journal of Advanced Computer Science and Applications 8.12 (2017). [15]

[16]. Rajanikanth Aluvalu and M. A. Jabbar. "RFAODE: A novel ensemble intrusion detection system." Computer Science Proceedings 115 (2 017): 226-234. [16]