International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 6 ǁ 2022 ǁ PP. 71-77

Intrusion Detection of Imbalanced Network Based On Machine Learning

Manjubashini B 1st, Vasanth R 2nd, Bharathkumar S 3rd, Ajithkumar K 4th, Chandhru S 5th

1st Assistant Professor, 2nd, 3rd, 4th, 5th UG Scholar (B.E), Department of Computer Science and Engineering, Mahendra Institute of Technology, Mahendhirapuri.

Abstract

The Internet, computer hackers are evolving quickly, and the state of cyber security isn't always optimistic. Machine Learning (ML) and Deep Learning (DL) methodologies for community intrusion detection evaluation and provides a brief educational summary of each ML/DL strategy. Papers from all approaches were indexed, read, and summarized only on the basis of their chronological overall thermal correlations. Because data is so important in ML/DL techniques, they present a few of the most commonly used common datasets in ML/DL, discuss the challenges of using ML/DL for cyber security, and provide advice for research possibilities. Within the research of Intrusion Detection methods, the KDD information set is a widely acknowledged benchmark. There is a lot of work going into the creation of intrusion detection approaches, and research into the information used for teaching and verifying out the detected version is also a big issue because better information quality can improve offline intrusion detection. This assignment evaluates the KDD data set with respect to four lessons: Basic, Content, Traffic, and Host, where all information properties can be classified using MODIFIED RANDOM FOREST (MRF). For Malware Detection, the evaluation is completed with attention to two important assessment metrics: Detection Rate (DR) and False Alarm Rate (FAR) (IDS). As a consequence of this empirical examination of the data set, the significance of each of the four qualities lessons on DR and FAR has been established, which may aid in improving the applicability of the data set to acquire the maximum DR with the least FAR.

Keywords: *Intrusion Detection, Imbalanced network, Detection of Fluctuate network, Malware detection using Intrusion Detection*

Date of Submission: 20-05-2022 Date of acceptance: 03-06-2022

I. INTRODUCTION

An interruption location structure is a customizing that assesses a single or a group of PCs for poisonous behaviors such as data theft, blue penciling, or corrupting framework exhibits. The majority of methods utilized in today's interference detection frameworks are unprepared to deal with the variable and intricate character of computerized attacks on Pc systems. Despite the fact that sophisticated flexible methods such as numerous AI frameworks can improve identification rates, reduce false alarm rates, and provide appropriate estimation and correspondence costs. Data mining can be used to achieve external model mining, request, collection, and more modest data streams than a usual data stream. The Interruption Detection System (IDS) is a product application that monitors the organization's or framework's activities and detects any malicious activity. The rapid growth and use of the internet has raised concerns about how to safeguard and transmit digital data in a secure manner. Nowadays, programmers use a variety of attacks to obtain crucial information. Identifying these assaults is aided by a variety of interruption location methodologies, techniques, and calculations. This interruption identification's main goal is to provide a comprehensive report on the meaning of interruption location, history, life cycle, different interruption discovery strategies, different types of assaults, various instruments and procedures, research needs, difficulties, and applications. A new (arising) subject is something individuals want to examine, remarking, or sending the data further to their companions. Customary methodologies for subject location have mostly been worried about the frequencies of (printed) words. Recognition and following of subjects have been concentrated on widely in the space of theme discovery and following (TDT) In this unique circumstance, the principle task is to either order another report into one of the known points (following) or to distinguish that it has a place with none of the known classes. In this way, worldly design of themes has been demonstrated and broken down through unique model choice, fleeting text mining, and factorial secret Markov models. This assault identification framework gives different layer safeguard to acquire the protectors valuable time before unrecoverable outcomes happen in the actual framework. The information utilized for showing the proposed recognition framework are from a constant ICS testbed. Five assaults, remembering individual for the center (MITM), forswearing of administration (DoS), information exfiltration, information altering, and misleading information infusion, are completed to mimic the results of digital assault and produce information for building information driven location models.

II. LITERATURE SURVEY

Iman Sharaf Aldin et al., has proposed in this paper with exponential growth in the size of computer networks and developed applications, the significant increase of the potential damage that can be caused by launching attacks is becoming obvious. Meanwhile, Intrusion Detection Systems (IDSs) and Intrusion Prevention Systems (IPSs) are one of the most important defense tools against the sophisticated and ever-growing network attacks. Due to the lack of adequate dataset, anomaly-based approaches in intrusion detection systems are suffering from accurate deployment, analysis and evaluation. There exist a number of such datasets such as DARPA98, KDD99, ISC2012, and ADFA13 that have been used by the researchers to evaluate the performance of their proposed intrusion detection and intrusion prevention approaches. Based on our study of eleven available datasets since 1998, many such datasets are out of date and unreliable to use. Some of these datasets suffer from lack of traffic diversity and volumes, some of them do not cover the variety of attacks, while others anonym zed packet information and payload which cannot reflect the current trends, or they lack feature set and metadata. Having reliable, publicly available IDS evaluation datasets is one of the fundamental concerns of researchers and producers in this domain. In this paper, we have monitored the state-of-the-art in the IDS dataset generation and evaluation by analyzing the eleven publicly available datasets since 1998 which are limited because of the lack of the traffic diversity and volumes, anonymized packet information and payload, constraints on the variety of attacks, and lack of the feature set and metadata. On the evaluate section, we fist extract the 80 traffic features from the dataset and clarify the best short feature set to detect each attack family using Random Forest Regressor algorithm. Afterwards, we examine the performance and accuracy of the selected features with seven common machine learning algorithms.

Amirhossein Gharib et al., has proposed in this paper the growing number of security threats on the Internet and computer networks demands highly reliable security solutions. Meanwhile, Intrusion Detection (IDSs) and Intrusion Prevention Systems (IPSs) have an important role in the design and development of a robust network infrastructure that can defend computer networks by detecting and blocking a variety of attacks. Reliable benchmark datasets are critical to test and evaluate the performance of a detection system. There exist a number of such datasets, for example, DARPA98, KDD99, ISC2012, and ADFA13 that have been used by the researchers to evaluate the performance of their intrusion detection and prevention approaches. However, not enough research has focused on the evaluation and assessment of the datasets themselves. In this paper we present a comprehensive evaluation of the existing datasets using our proposed criteria, and propose an evaluation framework for IDS and IPS datasets. We have studied the exist datasets for the test and evaluation of IDSs, and presented a new framework to evaluate datasets with the following characteristics: Attack Diversity, Anonymity, Available Protocols, Complete Capture, Complete Interaction, Complete Network Configuration, Complete Traffic, Feature Set, Heterogeneity, Labeled Dataset, and Metadata. The proposed framework considers organization policy and conditions using a coefficient, W, which can be defined separately for each criterion.

Gerard Draper Gil et al., has proposed in this paper Traffic characterization is one of the major challenges in today’s security industry. The continuous evolution and generation of new applications and services, together with the expansion of encrypted communications makes it a difficult task. Virtual Private Networks (VPNs) are an example of encrypted communication services that are becoming popular, as a method for bypassing censorship as well as accessing services that are geographically locked. In this paper, we study the effectiveness of flow-based time-related features to detect VPN traffic and to characterize encrypted traffic into different categories, according to the type of traffic e.g., browsing, streaming, etc. We use two different well-known machine learning techniques (C4.5 and KNN) to test the accuracy of our features. Our results show high accuracy and performance, confirming that time-related features are good classifiers for encrypted traffic characterization. We have studied the efficiency of time related features to address the challenging problem of characterization of encrypted traffic and detection of VPN traffic. We have proposed a set of time-related features and two common machine learning algorithms, C4.5 and KNN, as classification techniques. Our results prove that our proposed set of time-related features are good classifiers, achieving accuracy levels above 80%. C4.5 and KNN had a similar performance in all experiments, although C4.5 has achieved better results. From the two scenarios proposed, characterization in 2 steps (scenario A) vs. characterization in one step (scenario B), the first one generated better result. In addition to our main objective, we have also found that our classifiers perform better when the flows are generated using shorter timeout values, which contradicts the common assumption of using 600s as timeout duration. As future work we plan to expand our work to other applications and types of encrypted traffic, and to further study the application of time-based features to characterize encrypted traffic.

III. EXISTING METHOD

A new (emerging) topic is something people feel like discussing, commenting, or forwarding the information further to their friends. Conventional approaches for topic detection have mainly been concerned with the frequencies of (textual) words. Detection and tracking of topics have been studied extensively in the area of topic detection and tracking (TDT) In this context, the main task is to either classify a new document into one of the known topics (tracking) or to detect that it belongs to none of the known categories. Subsequently, temporal structure of topics has been modeled and analyzed through dynamic model selection, temporal text mining, and factorial hidden Markov models. This attack detection system provides multiple-layer defense in order to gain the defenders precious time before unrecoverable consequences occur in the physical system. The data used for demonstrating the proposed detection system are from a real-time ICS testbed. Five attacks, including man in the middle (MITM), denial of service (DoS), data exfiltration, data tampering, and false data injection, are carried out to simulate the consequences of cyber-attack and generate data for building data-driven detection models. Four classical classification models based on network data and host system data are studied, including k-nearest neighbor (KNN), decision tree, bootstrap aggregating (Bagging), and random forest, to provide a secondary line of defense of cyber-attack detection in the event that the intrusion prevention layer fails. Intrusion detection results suggest that KNN, Bagging, and random forest have low missed alarm and false alarm rates for MITM and DoS attacks, providing accurate and reliable detection of these cyber-attacks. This system auto-associative kernel regression (AAKR) model is studied to strengthen early attack detection. The result shows that this approach detects physically-impactful cyber-attacks before significant consequences occur. The proposed multiple-layer data driven cyber-attack detection system utilizing network, system,

IV. PROPOSED SYSTEM

This project proposed a new approach to detect the emergence of topics in social network streams. The basic idea of our approach is to focus on the social aspects of the post, not the textual content. This is reflected in the user's behavior. We have proposed a probabilistic model that captures both the number of mentions per post and the frequency of mentions. The overall flow of the proposal assumes that data from social network services arrives sequentially via the API. For each new post, we will use a sample within the past T time interval to train the reference model suggested by the corresponding user below. Assign an anomaly score to each post based on the probability distribution you have learned. Scores are then aggregated among users and sent to change point analysis. The approach is being explored to detect anomalies in large datasets using indicators focused on metaheuristic multi-start strategies and genetic algorithms. Rotation The methodology inherits some motivation from negative selection-based cognitive generation. The evaluation of this methodology is performed using the NSLKDD dataset, which is a modified version of the widely used KDD CUP99 dataset. To increase its adaptability and flexibility, the inspected parameter values are automatically selected according to the training data set used. It also reduces detection generation time by improving clustering.

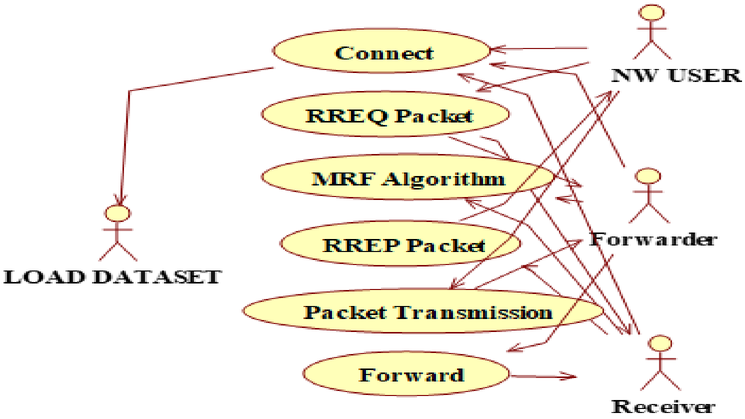


Fig 4.1 Use Case Diagram

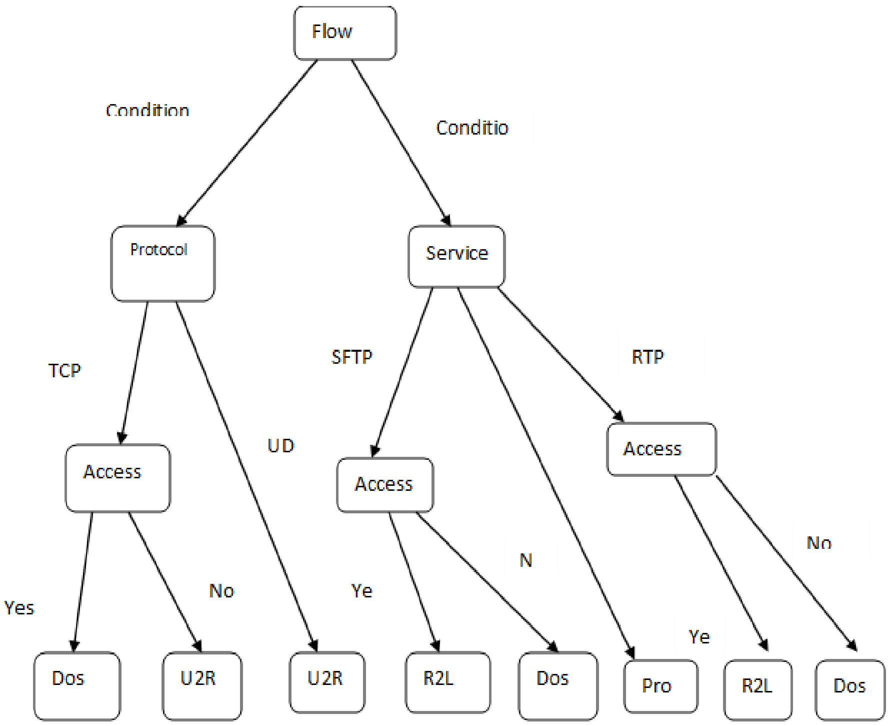
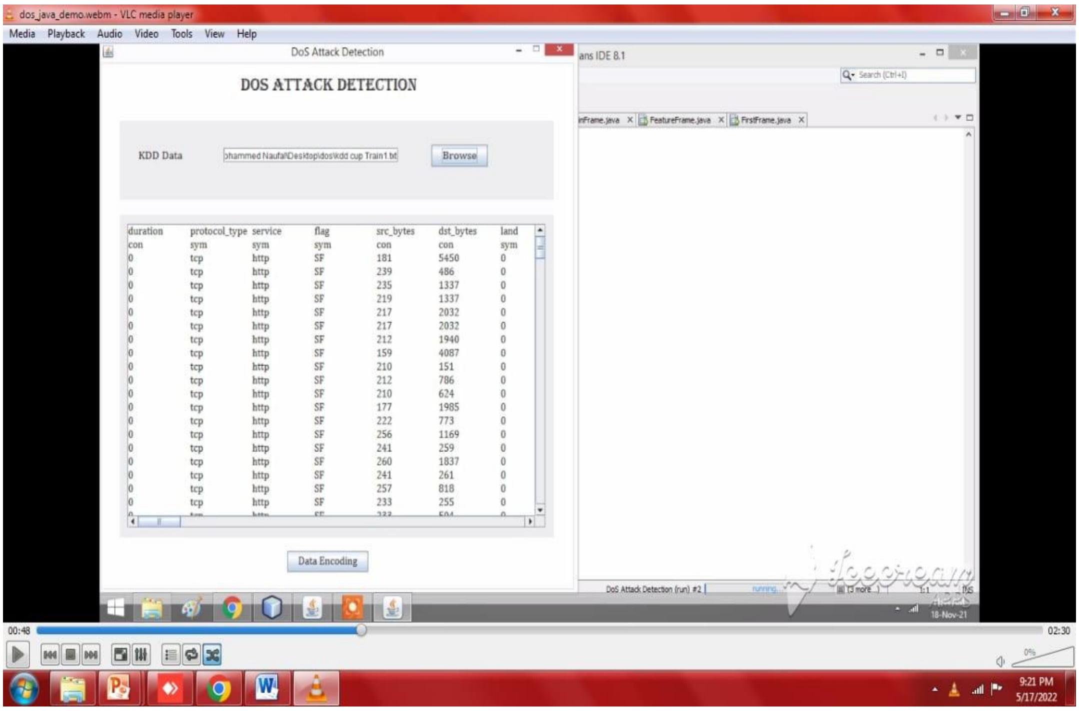


Fig 4.2 Flow Diagram

V.   KEY RESULTS

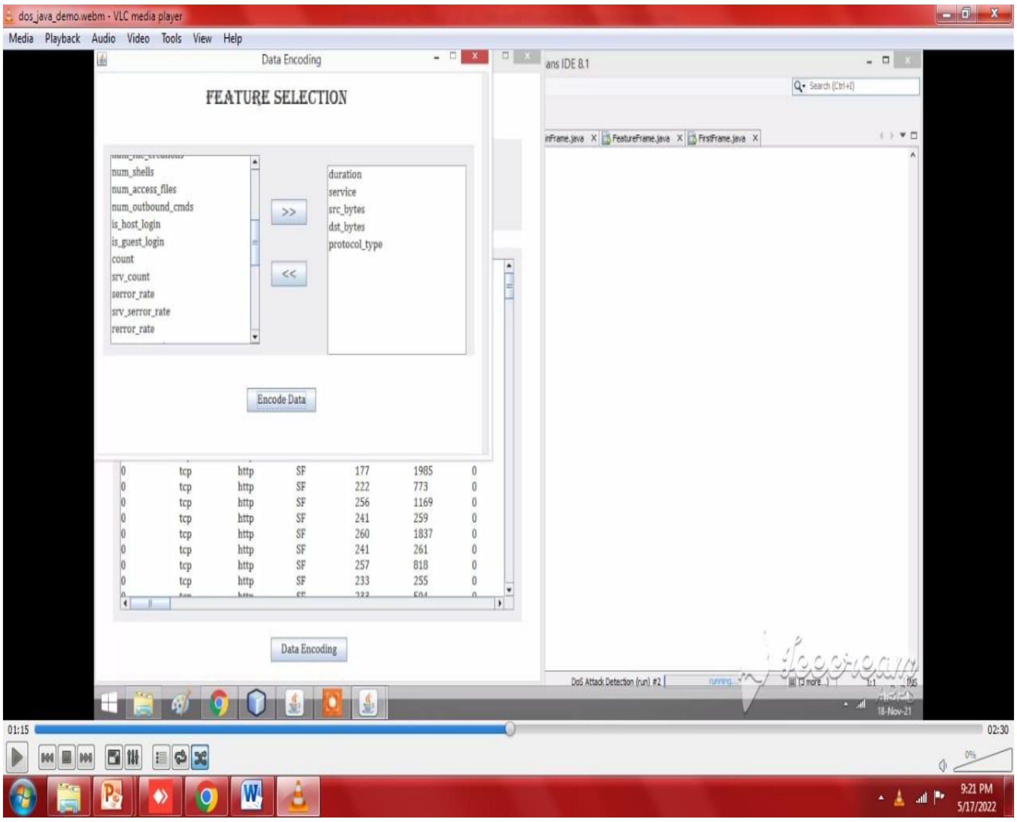


DOS ATTACK DETECTION

KDD Data ɔhammed Naufal\Desktop\dos\kdd cup Train1.txt

duration	protocol_type	service	flag	src_bytes	dst_bytes	land
con	sym	sym	sym	con	con	sym
0	tcp	http	SF	181	5450	0
0	tcp	http	SF	239	486	0
0	tcp	http	SF	235	1337	0
0	tcp	http	SF	219	1337	0
0	tcp	http	SF	217	2032	0
0	tcp	http	SF	217	2032	0
0	tcp	http	SF	212	1940	0
0	tcp	http	SF	159	4087	0
0	tcp	http	SF	210	151	0
0	tcp	http	SF	212	786	0
0	tcp	http	SF	210	624	0
0	tcp	http	SF	177	1985	0
0	tcp	http	SF	222	773	0
0	tcp	http	SF	256	1169	0
0	tcp	http	SF	241	259	0
0	tcp	http	SF	260	1837	0
0	tcp	http	SF	241	261	0
0	tcp	http	SF	257	818	0
0	tcp	http	SF	233	255	0

Fig 5.1 DOS attack Detection


dos_java_demo.webm - VLC media player
Media Playback Audio Video Tools View Help
Data Encoding
>>
<<
Encode Data
Data Encoding
01:15
02:30
9:21 PM
5/17/2022

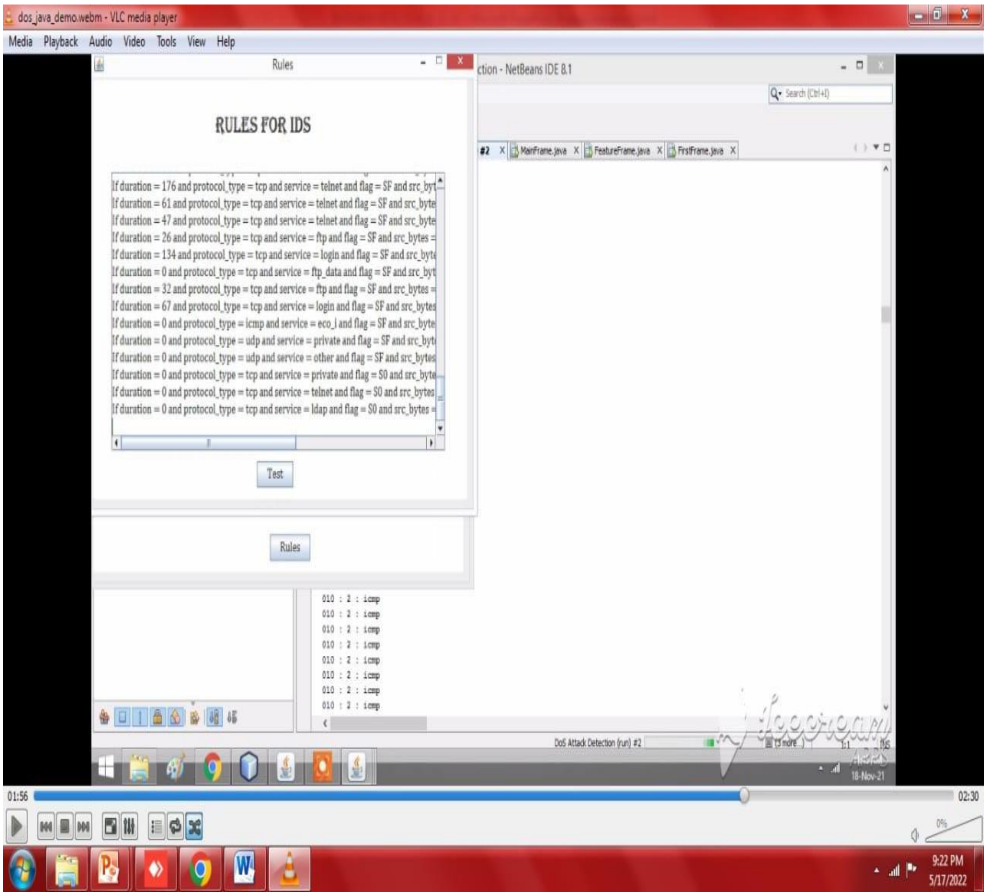
FEATURE SELECTION

```
num_shells
num_access_files
num_outbound_cmds
is_host_login
is_guest_login
count
srv_count
serror_rate
srv_serror_rate
rerror_rate
```

```
duration
service
src_bytes
dst_bytes
protocol_type
```

```
0    tcp    http    SF    177    1985    0
0    tcp    http    SF    222    773     0
0    tcp    http    SF    256    1169    0
0    tcp    http    SF    241    259     0
0    tcp    http    SF    260    1837    0
0    tcp    http    SF    241    261     0
0    tcp    http    SF    257    818     0
0    tcp    http    SF    233    255     0
```

Fig 5.2 Feature Selection

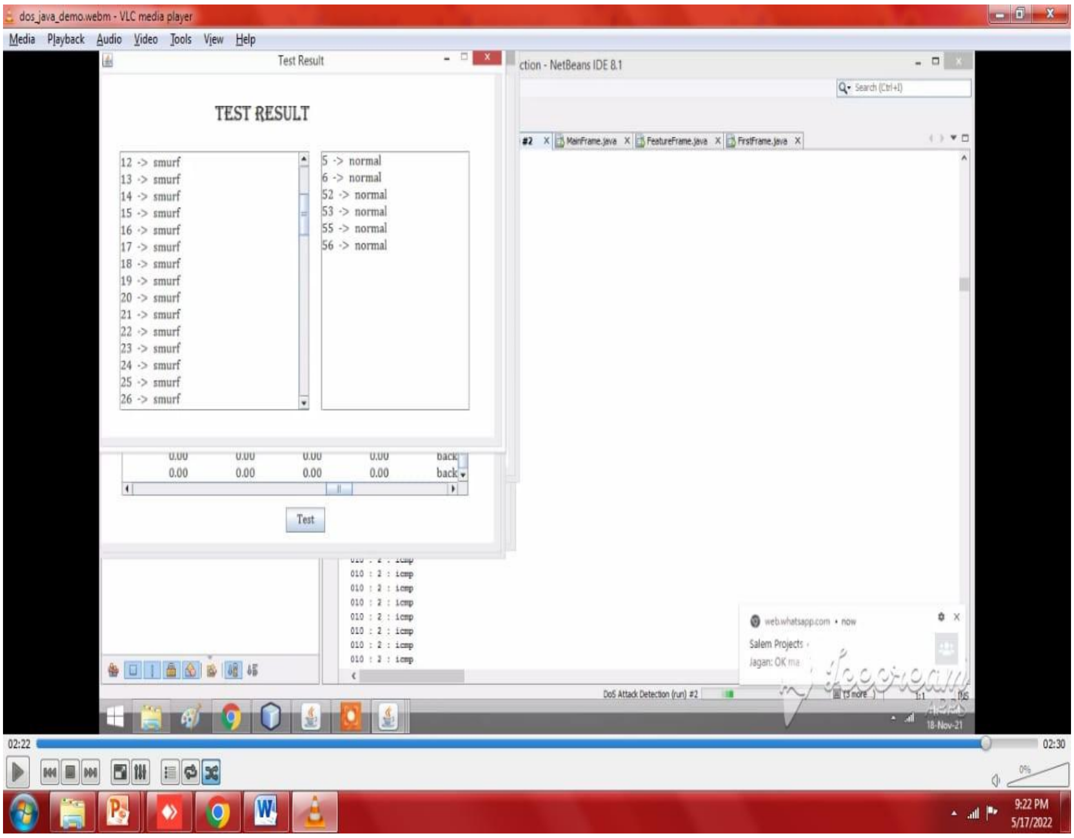

dos_java_demo.webm - VLC media player
Media Playback Audio Video Tools View Help
Rules
Test
Rules
01:56
02:30
9:22 PM
5/17/2022

RULES FOR IDS

```
If duration = 176 and protocol_type = tcp and service = telnet and flag = SF and src_byt
If duration = 61 and protocol_type = tcp and service = telnet and flag = SF and src_byte
If duration = 47 and protocol_type = tcp and service = telnet and flag = SF and src_byte
If duration = 26 and protocol_type = tcp and service = ftp and flag = SF and src_bytes =
If duration = 134 and protocol_type = tcp and service = login and flag = SF and src_byte
If duration = 0 and protocol_type = tcp and service = ftp_data and flag = SF and src_byt
If duration = 32 and protocol_type = tcp and service = ftp and flag = SF and src_bytes =
If duration = 67 and protocol_type = tcp and service = login and flag = SF and src_bytes
If duration = 0 and protocol_type = icmp and service = eco_i and flag = SF and src_byte
If duration = 0 and protocol_type = udp and service = private and flag = SF and src_byt
If duration = 0 and protocol_type = udp and service = other and flag = SF and src_bytes
If duration = 0 and protocol_type = tcp and service = private and flag = S0 and src_byte
If duration = 0 and protocol_type = tcp and service = telnet and flag = S0 and src_bytes
If duration = 0 and protocol_type = tcp and service = ldap and flag = S0 and src_bytes =
```

```
010 : 2 : icmp
010 : 2 : icmp
010 : 2 : icmp
010 : 2 : icmp
010 : 2 : icmp
010 : 2 : icmp
010 : 2 : icmp
010 : 2 : icmp
```

Fig 5.3 Rules for IDS



TEST RESULT

```
12 -> smurf        5 -> normal
13 -> smurf        6 -> normal
14 -> smurf        52 -> normal
15 -> smurf        53 -> normal
16 -> smurf        55 -> normal
17 -> smurf        56 -> normal
18 -> smurf
19 -> smurf
20 -> smurf
21 -> smurf
22 -> smurf
23 -> smurf
24 -> smurf
25 -> smurf
26 -> smurf
```

```
0.00    0.00    0.00    0.00    back
0.00    0.00    0.00    0.00    back
```

```
010 : 2 : icmp
010 : 2 : icmp
010 : 2 : icmp
010 : 2 : icmp
010 : 2 : icmp
010 : 2 : icmp
010 : 2 : icmp
010 : 2 : icmp
```

Fig 5.4 Test Results

VI. CONCLUSION AND FUTURE ENHANCEMENTS

In this project, we have proposed a new approach to detect the emergence of topics in a social network stream. The basic idea of our approach is to focus on the social aspect of the posts reflected in the mentioning behavior of users instead of the textual contents. We have combined the proposed mention model with the MRF change-point detection algorithm. The signature-based detection gives higher detection accuracy and lower false positive rate but it detects only known attack but anomaly detection is able to detect unknown attack but with higher false positive rate. The Intrusion Detection System plays a very significant role in identifying attacks in network. There are various techniques used in IDS like signature-based system, anomaly-based system. But Signature based system can detect only known attack, unable to detect unknown attack but anomaly-based system is able to detect attack which is unknown. Here Anomaly based system with integrated approach using multi-start metaheuristic method is defined. The various detection techniques introduced but till the main issue is regarding detection accuracy and false positive rate. The various types of attacks are also described and also terms regarding Intrusion detection system are also described

REFERENCES

[1]. Sharafaldin, I, Lashkari,A.H and Ghorbani, A.A, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization", fourth International Conference on Information Systems Security and Privacy (ICISSP), Purtogal, (2018).

[2]. Gharib, A., Sharafaldin, I., Lashkari, A.H. also, Ghorbani, A.A., "An Evaluation Framework for Intrusion Detection Dataset". 2016 IEEE International Conference Information Science and Security (ICISS), pp. 1-6, (2016)

[3]. Gil, G.D., Lashkari, A.H., Mamun, M. also, Ghorbani, A.A., "Portrayal of encoded and VPN traffic utilizing time-related highlights. In Proceedings of the second International Conference on Information Systems Security and Privacy, pp. 407-414, (2016).

[4]. Moustafa, N. also, Slay, J., "The assessment of Network Anomaly Detection Systems: Statistical investigation of the UNSW-NB15 informational collection and the correlation with the KDD99 dataset". Data Security Journal: A Global Perspective, 25(1-3), pp.18-31, (2016).

[5]. Moustafa, N. also, Slay, J., "UNSW-NB15: a far reaching informational collection for network interruption recognition frameworks (UNSW-NB15 network informational collection). IEEE Military Communications and Information Systems Conference (MilCIS), pp. 1-6, (2015).

[6]. Pongle, Pavan, and Gurunath Chavan. "An overview: Attacks on RPL and 6LoWPAN in IoT." IEEE International Conference on Pervasive Computing, (2015).

[7]. Oh, Doohwan, Deokho Kim, and Won Woo R, "A malevolent example recognition motor for inserted security frameworks in the Internet of Things." Sensors, pp, 24188-24211, (2014).

[8]. Mangrulkar, N.S., Patil, A.R.B. also, Pande, A.S., "Organization Attacks and Their Detection Mechanisms: A Review". Worldwide Journal of Computer Applications, 90(9), (2014).

[9]. Kasinathan, P., Pastrone, C., Spirito, M. A., and Vinkovits, M. "Denialof-Service recognition in 6LoWPAN based Internet of Things." In IEEE ninth International Conference on Wireless and Mobile Computing, Networking and Communications, pp. 600-607, (2013).

[10]. Kanda, Y., Fontugne, R., Fukuda, K. also, Sugawara, T., "Respect: Anomaly recognition strategy utilizing entropy-based PCA with three-venture portrays". PC Communications, 36(5), pp.575-588, (2013).