# NETWORK INTRUSION DATASETS USED IN NETWORK SECURITY EDUCATION

Alia Yahia[1] and Eric Atwell[2]

[1]Sudan University of Science and Technology, Khartoum, Sudan
[2]University of Leeds, Leeds, UK

## *ABSTRACT*

*There is a gap between the network security graduate and the professional life. In this paper we discussed the different types of network intrusion dataset and then we highlighted the fact that any student can easily create a network intrusion dataset that is representative of the network they are in. Intrusions can be in form of anomaly or network signature; the students cannot grasp all types but they have to have the ability to detect malicious packets within his network.*

## 1. INTRODUCTION

The need to gap between theoretical and practical network security is a must as black hats are getting more aggressive by the day and the white hats must keep up with new attacks. Student learning network intrusion should be able to see attack signatures and learn the different techniques to detect them. The data used in generating these datasets; should it be live or simulated, should be labeled putting into account signature base or anomaly base detection to detect new attacks. The two mostly wide datasets in network intrusion research are the KDD CUP99 and DARPA 1998-1999 are still in use although they face strong criticism. Laskov highlighted the reasons why these datasets are used in spitecriticism generating new datasets with labeled packets, which is extremely time consuming and sometimes impossible.Live data does not cover all types of attacks available which hinders the training process.

## 2. DATASETS

### 2.1 DARPA

DARPA network dataset intrusion has been created for the purpose of evaluating the different Algorithms used to detect intrusions (Ciza et *al*., 2008). It has been criticized for its lack of ability to detect zero base attacks and the absence of false positives (Zuechet *al*., 2015). The data generated is made up of backward data that is intended to be completely free of attacks and attack data that is intended to consist entirely of attack scenarios. It is argued by McHugh (2000)that one can really simulate data that mimics the live data since it isunpredictable and does not follow a certain norm or pattern. The 300 attacks where synthesis in the data, but the order of the attacks was consistence through the 10 weeks of capture which can be argued as unrealistic. The purpose of the educational dataset is to help students understand intrusions and their detection and DARPA lacks this support due to the attack taxonomy used. (McHugh, 2000)

### 2.2 KDD

KDD99 dataset is the most widely used dataset when combining the domains of network intrusions with machine learning (Ozg¨ur& Erdem, 2016);despite the fact that there are redundant

DOI :10.5121/ijite.2018.7304

records in the dataset which makes the detection process skewed and unreliable. (Zuech et *al*., 2015).“Dataset contains 24 attack types in training and 14 more attack types in testing for total of 38 attacks. These 14 new attacks theoretically test IDS capability to generalize to unknown attacks. At the same time, it is hard for machine learning based IDS to detect these 14 new attacks” (Ozg¨ur& Erdem, 2016). Thus the main objective of the KDD is not achieved; which is to help in the evaluation of machine learning algorithms.

### 2.3 Indian River State College Dataset

The dataset generated is a hybrid between IRSC dataset representing real data and simulated network attacks generated by the network team. (Zuech et *al*., 2015) Full packet capture and Network Flow data capture are used. Full packet capture are achieved by capturing the packets using Snort. Data cleaning is performed by capturing the packets using both Snort and Wireshark, captured are compared and any missing packets merged. Network Flow data capture provide a higher level of abstraction since the volume of data it collected is less and it is summarized. (Zuech et *al*., 2015) Labeling simulated attack is straightforward because the network teach are the ones who generated the attacks. Labeling real attacks is done in a multistage process, stage 1 use Snort rules to detect intrusions and then pass them to the network flow component. Manual inspection is then performed. The advantage of this dataset it that it captures packets from a live network which reflects real traffic scenarios which is one of the problems of the DARPRA dataset.

### 2.4 Kyoto 2006+

Kyoto 2006+ is an evaluation dataset where the data is obtained from diverse honeypots from November 2006 to August 2006. One drawbacks is that the data is mostly intrusive which make the training method skewed. The dataset is labeled where 1(normal session) , -1 (known attacks) and -2( unknown attacks) . The dataset does not contain any attacks generated from or targeting windows machines which is a problem since the number 1 desktop operating system. The number of protocol types is limited to TCP , UDP and ICMP which does not reflect the different types of attacks found.

## 3. Problem Definition

Creating a network intrusion detection educational corpus using corpus linguistics methodology with a limited number of features and from real production network will led to a balance corpus that will contain different types of intrusions that can form the basis in which students can learn cyber security.

## 4. Methodology

Merging the corpus linguistics methodology with network security model will lead to a network dataset that can be used as a tool in teaching the different types of network attacks found in real network datasets. It will contain packets that are representative of real life scenarios.

### 4.1 Corpus Linguistics Methodology And Network Security Methodology

TCP/IP is the language of the Internet. Its main objective was to create a friendly communication between two heterogeneous entities. Over time it has been abused to allow intrusive communication. This has been possible by the vulnerabilities found in all aspects of the communication process from protocols,hardware, system software and application software.

“A corpus is a collection of machine-readable texts that have been produced in a natural communicative setting. They have been sampled to be representative and balanced with respect to particular factor.” (Evans, 2009). The network security model cycle is applied to build a corpus of intrusions .The cycle consist of collection,detection and then analysis.
Collection can be full content data, session data, statistical data , packet string data and alert data .

Detection according to intrusion detection system can be classified into vulnerability centric and threat centric (Ciza et *al*., 2008) Threat centric was adopted because it is base on collection and relies on utilizing all sources of attack .
Analysis can be packet analysis , network forensics , host forensics and malware analysis.

### 4.2 Collection

Before collecting the data we designed the corpus.There were a number of parameters that had to be addressed before creating the corpus (Evans, 2009). The size of the corpus(dataset) should represent the different type of intrusions so the students can learn from it. This will not be visible but at least there should be an example of a network attack for each of the protocol dialects represented in this paper by the major protocols found in each of the TCP/IP protocol stack.

### 4.3 Balance

We have collected the dataset from different networks at different time and days. Hopefully this will bring balance. The people (the computers. mobiles and processes) communicating using TCP/IP where diverse using different operating system and hardware architectures.

### 4.4 Representativeness

The packet capture were captured from inside the Sudan network , and our domain is to teach students inside Sudanese university the different type of network intrusions found inside Sudanese networks.

The capture was done using Wireshark ; which is an opens source packet sniffer.

### 4.5 Analysis And Detection

The second research question was can we label the packets as intrusive or not by using Wireshark as network intrusion dataset. We have adopted network base,signature base detection so the percentage of false negative becomes zero. This was not possible since no baseline of normal activities was created and network usage. So despite the fact that a signature can appear as an intrusion e.g the data value in the protocol hierarchy statistics; which means that Wireshark has located a dialect (protocol) it cannot detect , this can be an intrusion or a new application in the market and Wireshark has not built a protocol dissector for it.

### 4.6 Experiment

Intrusion detection starts with where to place the sensor that will capture the packets, in this situation, where to place our sensors was overlooked as it was one of the research question is it really of importance. We captured 11 datasets from different sites on different days, our first objective was to merge them but this was not possible because they used the same private address space which would haveled to duplicate and usually the attack is made of a stream of packets, if merge the stream will be invalid.

We adopted the five classes of attacks found DARPRA to see if they are present or not. One of the limitation DARPRA was that it did not represent real present attacks so if our datasets did not have at least one type of the major type of attack then it falls short of being fit as a network intrusion educational dataset.

**Classes of attacks are**

- Denial of service: looked for the SYN flood , ICMP flood , UDP flood (Degadzor et *al*., 2017) as a method of detection also Time To Live expiry attack. Found this attack in the 10 of 11 datasets.
- Remote to user and user to root where not present, here we searched for brute force attacks
- Surveillance attacks : found in all the sweep attacks of different types.

**Other attacks**

- Arp poisoning method of detection was the expert info tab where the warning duplicate IP should be present. There was now present.

**To evaluate the dataset captured we adopted a check list of (Bhuyan et *al*., 2015)**

- Dataset contains real world data, it does but the real world data lacks diversity which is essentialin teaching the different types of attacks and how they are launched.
- Complete and correct labeling should be performed on the dataset, when labeling the dataset one should be sure that the packet is intrusive or not. Signature-base was adopted because with it the packet is for certain intrusive or not. Using Wireshark and no network architecture led to us missing intrusive action and even the packet we label as intrusive there is a percentage that they were generated from a network device which was misconfigured.
- Sufficient trace size: The corpus collected was representing of the TCP/IP conversion on Sudan networks but not of the different types of attacks that have to be illustrated.
- Featureextraction:feature engineering reduces the amount of data chosen, improves accuracy by only choosing the fields you need to detect the intrusion you want. One of our research question was with the limited number of features and Wireshark can we detect network intrusions, the reason for this is that we didn't want to overwhelm the students with too many features and we wanted the dataset size not to increase,.
- Diverse attack scenarios: this was not achieved because the conversion captured did contain different types of attacks
- Ratio between normal and attack traffic was skewed due to the fact Wireshark drops packets, thus we used signature base analysis therefore zero base attacks went undetected.

## Step One

Capture the data from a live working network

## Step Two

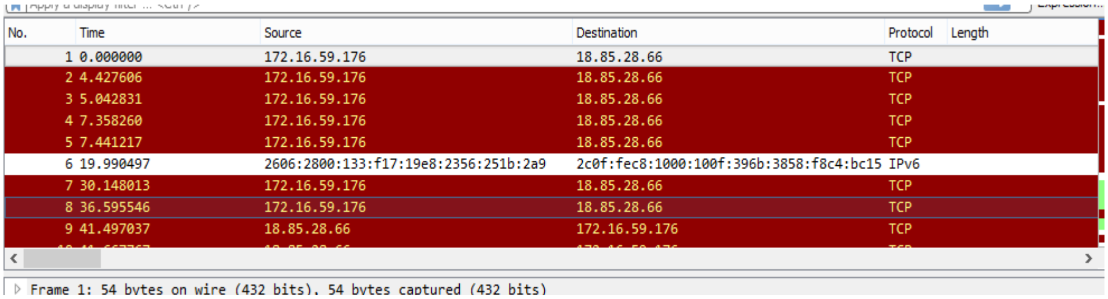
Export the packet using csv format

## Step Three

Run snort to detect Intrusion which will be used in labeling the data

## Step Four

Snortgenerates a log , open it in Wiresharkas shown in Wireshark Screen Shot below and then export it as CSV file



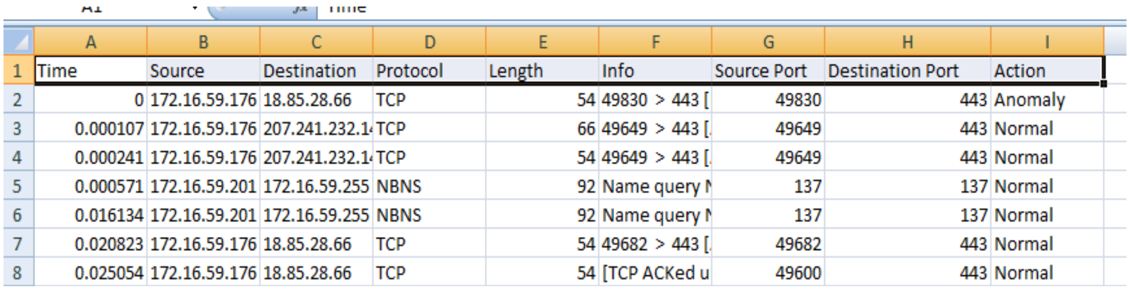
| No. | Time | Source | Destination | Protocol | Length |
|---|---|---|---|---|---|
| 1 | 0.000000 | 172.16.59.176 | 18.85.28.66 | TCP | |
| 2 | 4.427606 | 172.16.59.176 | 18.85.28.66 | TCP | |
| 3 | 5.042831 | 172.16.59.176 | 18.85.28.66 | TCP | |
| 4 | 7.358260 | 172.16.59.176 | 18.85.28.66 | TCP | |
| 5 | 7.441217 | 172.16.59.176 | 18.85.28.66 | TCP | |
| 6 | 19.990497 | 2606:2800:133:f17:19e8:2356:251b:2a9 | 2c0f:fec8:1000:100f:396b:3858:f8c4:bc15 | IPv6 | |
| 7 | 30.148013 | 172.16.59.176 | 18.85.28.66 | TCP | |
| 8 | 36.595546 | 172.16.59.176 | 18.85.28.66 | TCP | |
| 9 | 41.497037 | 18.85.28.66 | 172.16.59.176 | TCP | |

Frame 1: 54 bytes on wire (432 bits), 54 bytes captured (432 bits)

**Screenshot of Snort log file opened in Wireshark**

## Step Five

Export it as CSV file and open in Microsoft Excel . Add a new column name it action label the data as attack . Add other packets from step two and label them as normal, as illustrated in the table below.



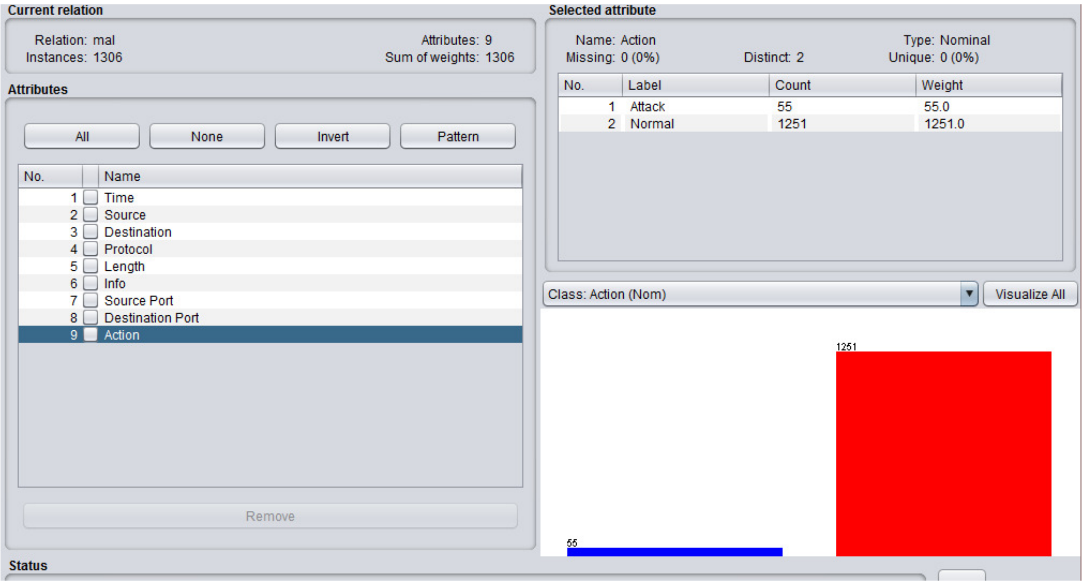
| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Time | Source | Destination | Protocol | Length | Info | Source Port | Destination Port | Action |
| 2 | 0 | 172.16.59.176 | 18.85.28.66 | TCP | 54 | 49830 > 443 [ | 49830 | 443 | Anomaly |
| 3 | 0.000107 | 172.16.59.176 | 207.241.232.1 | TCP | 66 | 49649 > 443 [ | 49649 | 443 | Normal |
| 4 | 0.000241 | 172.16.59.176 | 207.241.232.1 | TCP | 54 | 49649 > 443 [ | 49649 | 443 | Normal |
| 5 | 0.000571 | 172.16.59.201 | 172.16.59.255 | NBNS | 92 | Name query N | 137 | 137 | Normal |
| 6 | 0.016134 | 172.16.59.201 | 172.16.59.255 | NBNS | 92 | Name query N | 137 | 137 | Normal |
| 7 | 0.020823 | 172.16.59.176 | 18.85.28.66 | TCP | 54 | 49682 > 443 [ | 49682 | 443 | Normal |
| 8 | 0.025054 | 172.16.59.176 | 18.85.28.66 | TCP | 54 | [TCP ACKed u | 49600 | 443 | Normal |

**Table of labeled CSV file opened in Excel Spreadsheet**

## Step Six

Run several classifiers to choose the best one. Since the data is labeled using supervised learning ,and we choose the classifiers that were easy to explain to our students and were not black boxes.

When we opened the data in Wekawe found there was an imbalance in data as shown in the screenshotbelow. We first worked with the imbalance data and then used cost sensitive classifier to try to solve this imbalance.



Current relation

Relation: mal
Instances: 1306
Attributes: 9
Sum of weights: 1306

Attributes

| No. | Name |
|---|---|
| 1 | Time |
| 2 | Source |
| 3 | Destination |
| 4 | Protocol |
| 5 | Length |
| 6 | Info |
| 7 | Source Port |
| 8 | Destination Port |
| 9 | Action |

Selected attribute

Name: Action
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

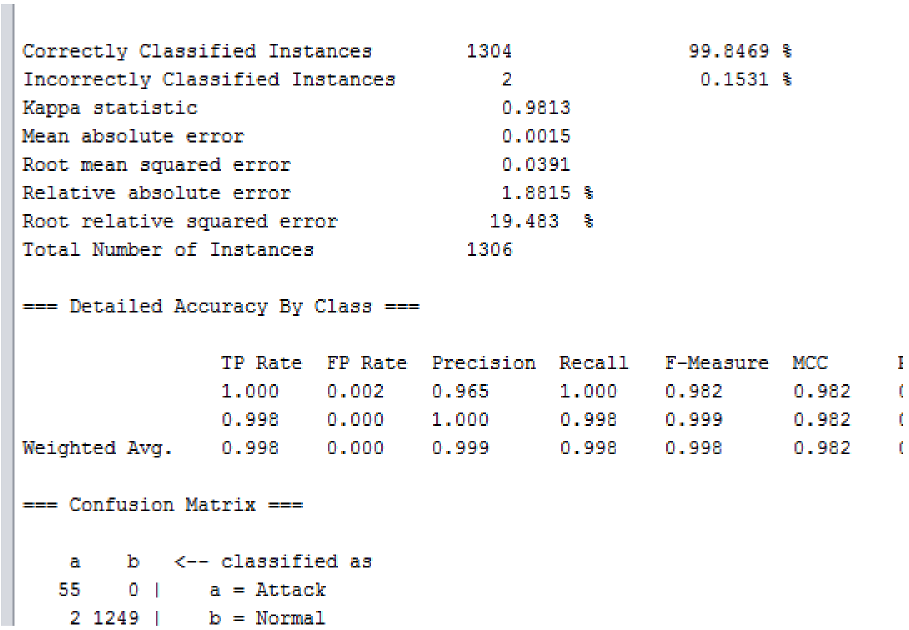
| No. | Label | Count | Weight |
|---|---|---|---|
| 1 | Attack | 55 | 55.0 |
| 2 | Normal | 1251 | 1251.0 |

Class: Action (Nom)

**Screenshot demonstrates the imbalance in the data as shown in Weka**

## Step Seven

Applied decision tree classifier  asillustrated in Screenshot below the detection rate was very high.

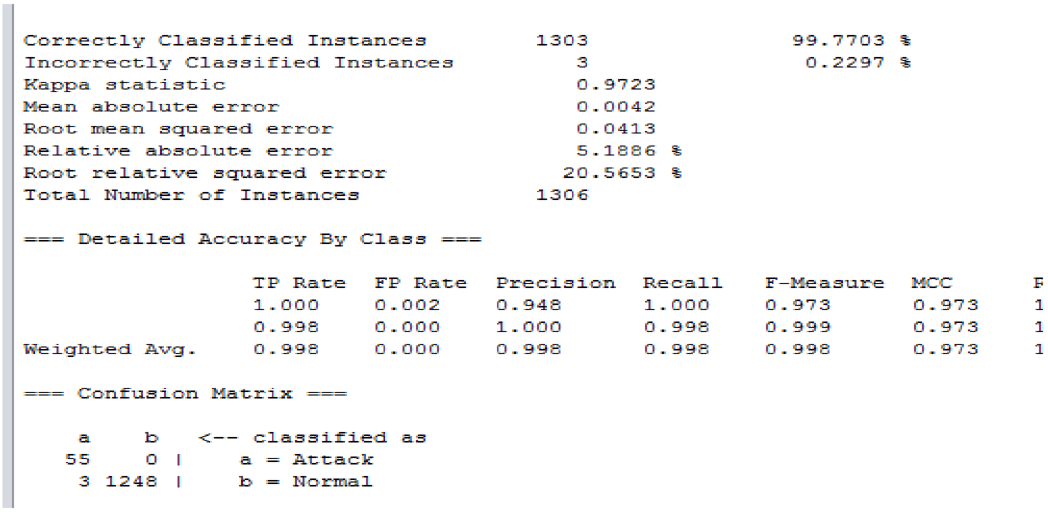


```
Correctly Classified Instances        1304               99.8469 %
Incorrectly Classified Instances         2                0.1531 %
Kappa statistic                          0.9813
Mean absolute error                      0.0015
Root mean squared error                  0.0391
Relative absolute error                  1.8815 %
Root relative squared error             19.483  %
Total Number of Instances             1306

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      I
                 1.000    0.002    0.965      1.000    0.982      0.982    (
                 0.998    0.000    1.000      0.998    0.999      0.982    (
Weighted Avg.    0.998    0.000    0.999      0.998    0.998      0.982    (

=== Confusion Matrix ===

    a    b   <-- classified as
   55    0 |    a = Attack
    2 1249 |    b = Normal
```

**Screenshot of Decision tree classifier in Weka**

Decided to try other classifiers

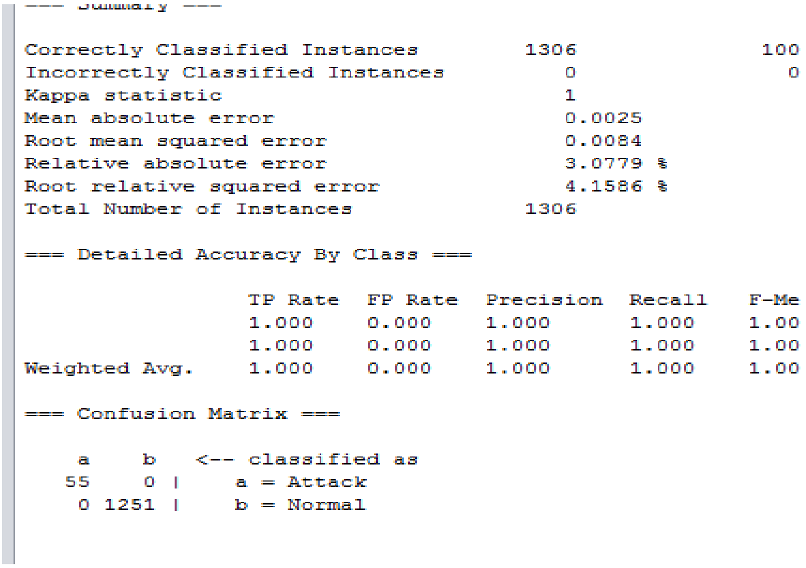
Naive Bayes



```
Correctly Classified Instances        1303               99.7703 %
Incorrectly Classified Instances         3                0.2297 %
Kappa statistic                          0.9723
Mean absolute error                      0.0042
Root mean squared error                  0.0413
Relative absolute error                  5.1886 %
Root relative squared error             20.5653 %
Total Number of Instances             1306

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      F
                 1.000    0.002    0.948      1.000    0.973      0.973    1
                 0.998    0.000    1.000      0.998    0.999      0.973    1
Weighted Avg.    0.998    0.000    0.998      0.998    0.998      0.973    1

=== Confusion Matrix ===

    a    b   <-- classified as
   55    0 |    a = Attack
    3 1248 |    b = Normal
```

**Screenshot of  Naive Bayes classifier**

Decision table



```
=== Summary ===

Correctly Classified Instances        1306               100
Incorrectly Classified Instances         0                 0
Kappa statistic                          1
Mean absolute error                      0.0025
Root mean squared error                  0.0084
Relative absolute error                  3.0779 %
Root relative squared error              4.1586 %
Total Number of Instances             1306

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Me
                 1.000    0.000    1.000      1.000    1.00
                 1.000    0.000    1.000      1.000    1.00
Weighted Avg.    1.000    0.000    1.000      1.000    1.00

=== Confusion Matrix ===

    a    b   <-- classified as
   55    0 |    a = Attack
    0 1251 |    b = Normal
```

**Screenshot of decision table classifier**

The above detection percentage are not logical.

## Step Eight

Applied the cost sensitive to balance the data but even after changing the weights the detection rate was the same.

## 5. CONCLUSION

Collecting the dataset not putting into account where the sensors are located ;before or after the firewall and not knowing the network map and not creating a baseline of normal behavior makes it sometimes impossible to differentiate between network related issues and network intrusions.It has been demonstrated that when apply the network dataset evaluation criteria it was obvious that the dataset created did not pass the test.Network security prerequisite in networks is a must. It has been demonstrated without knowledge of networks and protocol analysis it will be hard to detect intrusions and sometimes impossible. For educational dataset according to our finding it is better to generate packets using a simulated network environment to represent the different types of attack classes. The dataset should contain training dataset this should represent different types of intrusions . If the training dataset is not balance then the result will be skewed.

## REFERENCES

[1] Bhuyan, Monowar H., Bhattacharyya, Dhruba K., and Kalita, JugalK.. (2015). Towards Generating Real-life Datasets for Network Intrusion Detection. International Journal of Network Security, Vol.17, No.6, PP.675-693.

[2] Ciza, Thomas, Vishwas,Sharma , N. Balakrishan. (2008). Usefulness of Darpra Dataset for Intrusion Detection system Evaluation. Data Mining, Intrusion Detection, Information Assurance, and Data Networks Security 2008, 69730G .doi: 10.1117/12.777341

[3] Degadzor, Mohammed, M. A., Effrim, A. F., B. F., &Appiah, K. A. (2017). Brute Force Attack Detection And Prevention On A Network Using Wireshark Analysis. International Journal of Engineering Sciences & Research Technology, 6(6), 26-37. doi:10.5281/zenodo.802797

[4] Engels, Andre. (2015). Wikimedia Traffic Analysis Report – OS breakdown per Country. Wikimedia. 9:11. https://stats.wikimedia.org/wikimedia/squids/SquidReportCountryOs.htm

[5] Evans, David. (2009). Introduction to Corpus building and investigation for the humanities, An on-line information pack about corpus investigation techniques for the Humanities. Available: https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/Intro/Unit1.pdf

[6] Khan, NiloferShoaib and Lilhore,Umesh. (2016). Review of various intrusion detection methods for training data sets. International Journal of Modern Trends in Engineering and Research (IJMTER) Volume 03, Issue 12, ISSN (Online):2349–9745; ISSN (Print):2393-8161

[7] McHugh, John.(2000).The 1998 Lincoln Laboratory IDS Evaluation. RAID 2000, LNCS 1907, pp. 145-161.

[8] Ozg¨ur,Atilla&Erdem,Hamit.(2016).A Review of KDD99 Dataset Usage in Intrusion Detection and Machine Learning between 2010 and 2015. Open Access. https://doi.org/10.7287/peerj.preprints.1954v1

[9] Sangster, Benjamin , O'Connor, T. J. , Cook, Thomas ,Fanelli, Robert , Dean, Erik, Adams, William J. , Morrell, Chris, Conti, Gregory. (2009). Toward instrumenting network warfare competitions to generate labeled datasets. Proceedings of the 2nd conference on Cyber security experimentation and test.p.9-9. Montreal, Canada

[10] Zuech, Richard, Khoshgoftaar, Taghi M., Seliya,Naeem,.Najafabadi, Maryam M, and Kemp, Clifford.(2015). A New Intrusion Detection Benchmarking System.Florida Artificial Intelligence Research Society Conference;The Twenty-Eighth International Flairs Conference.Association for the Advancement of Artificial Intelligence (www.aaai.org)
https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS15/paper/view/10368/10314