Chapter 3

**Searching and analyzing open-source datasets for our cybersecurity use case**

# Summary of open-source datasets: -

We collected 7 open-source datasets from different sources, and they were analyzed to compare their characteristics, to gather information and finalize the dataset for the project.

Table 3.1.1 Summarizing analysis of all datasets

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sr No** | **Dataset name** | **Number of rows** | **Number of features** | **Number of duplicates** | **Null records: Y/N** | **Number of target features** | **Binary or Multi-class classification** | **Comments** |
| 1 | BETH | 763144 | 16 | 0 | N | 2 | Binary | 3 files: - training data validation data testingdata  This table hasrecords of training  data. |
| 2 | BrakTooth | 9002 | 5 | 1909 | N | 1 | Multi-class |  |
| 3 | Mil-STD- 1553 | 23000 | 52 | 0 | Y | 2 | Multi-class | 7 files, 1 is of benign data and 6 are of attacks  This table has records  of benign file. |
| 4 | ServerLogs | 172838 | 16 | 0 | N | 1 | Binary |  |
| 5 | UNR-IDD | 37411 | 34 | 1 | N | 2 | Binary, Multi-class |  |
| 6 | cic | 9167581 | 59 | 310 | N | 2 | Multi-class | The file is in  .parquet format. |
| 7 | KDD cup 1999 | 494021 | 42 | 348435 | N | 1 | Multi-class | Imported from sklearn  preloaded datasets. |

Table 3.1.2 Comparison of all datasets

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr No** | **Dataset name** | **Advantages** | **Disadvantages** |
| 1 | BETH | 1. Large number of records. 2. No duplicates and no null records. | 1. Does not allow to build model for multi-class classification. |
| 2 | BrakTooth | 1. No null values. 2. Moderate number of records 3. Allows to build model for multi-class classification. | 1. 21% records are duplicate. |
| 3 | Mil-STD- 1553 | 1. Large number of records with very high dimensionality. 2. Allows to build model for multi-class classification. | 1. Large number of null records. |
| 4 | ServerLogs | 1. Large number of records. 2. No duplicates or null records. | 1. Does not allow to build model for multi-class classification. |
| 5 | UNR-IDD | 1. Large number of records. 2. No duplicates or null records. 3. Allows to build model both binary and multi-class classification. |  |
| 6 | cic | 1. Large number of records. 2. Very high dimensionality. 3. Allow to build model for multi-class classification. |  |
| 7 | KDD cup 1999 | 1. Large number of records. 2. Very high dimensionality. 3. Allow to build model for multi-class classification. | 1. 70.5% records in the dataset are duplicate. |

From the initial analysis, we observed two datasets are most suitable for our project: -

* + 1. UNR-IDD
    2. cic

Thus, we further analyzed the two datasets to understand their properties.

# UNR-IDD dataset: -

Table 3.2.1 List of fields in UNR-IDD dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sr no** | **Field name** | **Description** | **Type of field** | **Comments** |
| 1 | Switch ID | 12 switches | High-order nominal |  |
| 2 | Port Number | 4 ports | Low-order nominal |  |
| 3 | Received Packets | Number of packets received by the port | Scale | Port statistics |
| 4 | Received Bytes | Number of bytes received by the port | Scale |
| 5 | Sent Bytes | Number of bytes sent | Scale |
| 6 | Sent Packets | Number of packets sent by the port | Scale |
| 7 | Port alive Duration (S) | The time port has been alive in seconds | Scale |
| 8 | Packets Rx Dropped | Number of packets dropped by the receiver | Scale |
| 9 | Packets Tx Dropped | Number of packets dropped by the sender | Scale |
| 10 | Packets Rx Errors | Number of transmit errors | Scale |
| 11 | Packets Tx Errors | Number of receive errors | Scale |
| 12 | Delta Received Packets | Number of packets received by the port | Scale | Delta port statistics |
| 13 | Delta Received Bytes | Number of bytes received by the port | Scale |
| 14 | Delta Sent Bytes | Number of packets sent by the port | Scale |
| 15 | Delta Sent Packets | Number of bytes sent | Scale |
| 16 | Delta Port alive Duration (S) | The time port has been alive in seconds | Scale |
| 17 | Delta Packets Rx Dropped | Number of packets dropped by the receiver | Scale |
| 18 | Delta Packets Tx Dropped | Number of packets dropped by the sender | Scale |
| 19 | Delta Packets Rx Errors | Number of transmit errors | Scale |
| 20 | Delta Packets Tx Errors | Number of receive errors | Scale |
| 21 | Connection Point | Network connection point expressed  as a pair of the network element identifier  and port number. | Scale | Flow entry and Flow table |
| 22 | Total Load/Rate | Obtain the current observed total load/rate (in bytes/s) on a link | Scale |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 23 | Total Load/Latest | Obtain the latest total load bytes counter viewed on that link. | Scale |  |
| 24 | Unknown Load/Rate | Obtain the current observed unknown-sized  load/rate (in bytes/s) on a link. | Scale |
| 25 | Unknown Load/Latest | Obtain the latest unknown- sized load bytes  counter viewed on that link. | Scale |
| 26 | Latest bytes counter |  | Scale |
| 27 | is\_valid | Indicates whether this load was built on valid  values. | Binary |
| 28 | Table ID | Returns the Table ID values. | Scale |
| 29 | Active Flow Entries | Returns the number of active flow entries in  this table. | Scale |
| 30 | Packets Looked Up | Returns the number of packets looked up in the table. | Scale |
| 31 | Packets Matched | Returns the number of packets that  successfully matched in the  table | Scale |
| 32 | Max Size | Returns the maximum size of this table. | Scale |
| 33 | Label | Normal: Normal network functionality  TCP-SYN: TCP-SYN Flood  PortScan: Port Scanning Overflow: Flow Table overflow Blackhole: Blackholde attack  Diversion: Traffic diversion attack | Multi-class classification | Target: Multi- class |
| 34 | Binary Label | Normal: Normal network functionality  Attack: Network intrusion | Binary classification | Target: Binary |

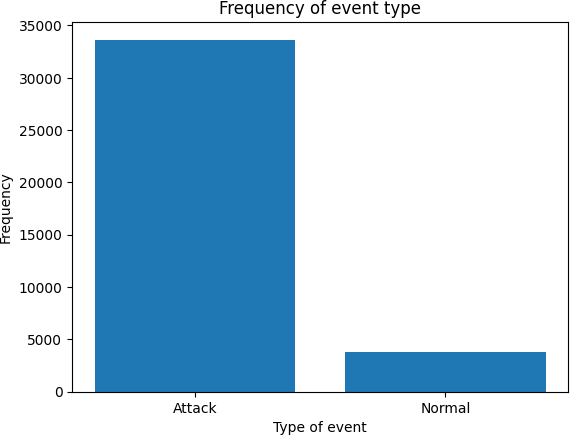


Figure 3.2.1 Bar chart of events in UNR-IDD dataset based on Binay label.

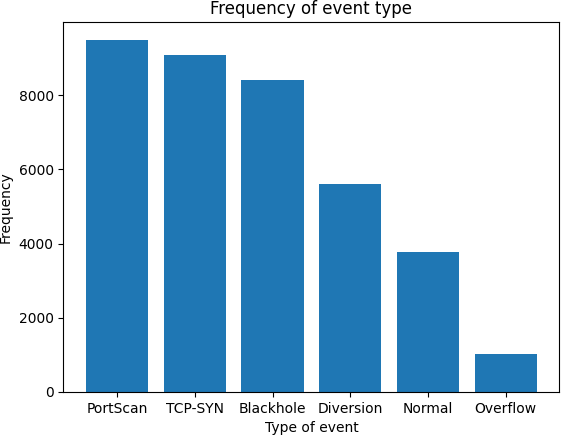


Figure 3.2.2 Bar chart of events in UNR-IDD dataset based on Label.

Observations from above analysis: -

* + 1. We have a high order nominal field: Switch ID with 12 distinct categories, thus, for which we need to identify if one-hot encoding is feasible or if any other better alternative method can be used for training the model.
    2. We have a low order nominal field: Port Number, with 4 distinct categories, thus, we can employ one-hot encoding to use the field while training the model.
    3. All other fields for training are numeric, and thus, we can normalize them for training the model.
    4. For target features, we have two fields: -
       1. Label: Multi-class classification
       2. Binary Label: Binary classification
    5. The dataset is highly imbalanced for target field: Binary Label. We have a greater number of records of type Attack and lesser number of records of type: Normal.
    6. The dataset is imbalanced for target field: Label. There are 3 types of attacks having the greatest number of events in the dataset: -

1. PortScan
2. TCP-SYN
3. Blackhole

Foreseen challenges with the dataset: -

1. Imbalanced nature of target features.
2. High order nominal field: Switch ID.

# CIC dataset: -

Table 3.3.1 List of fields in CIC dataset

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr No** | **Field Name** | **Description** | **Type of field** |
| 1 | Flow Duration | The duration of the flow. | Scale |
| 2 | Total Fwd Packets | Total number of forward packets | Scale |
| 3 | Total Backward Packets | Total number of backward packets | Scale |
| 4 | Fwd Packets Length Total | Total length of forward packets | Scale |
| 5 | Bwd Packets Length Total | Total length of backward packets | Scale |
| 6 | Fwd Packet Length Max | Maximum length of forward packets | Scale |
| 7 | Fwd Packet Length Mean | Mean length of forward packets | Scale |
| 8 | Fwd Packet Length Std | Standard deviation length of forward packets | Scale |
| 9 | Bwd Packet Length Max | Maximum length of backward packets | Scale |
| 10 | Bwd Packet Length Mean | Mean length of backward packets | Scale |

|  |  |  |  |
| --- | --- | --- | --- |
| 11 | Bwd Packet Length Std | Standard deviation length of backward packets | Scale |
| 12 | Flow Bytes/s | Flow bytes per second | Scale |
| 13 | Flow Packets/s | Flow packets per second | Scale |
| 14 | Flow IAT Mean | Mean time between flows | Scale |
| 15 | Flow IAT Std | Standard deviation of time between flows | Scale |

|  |  |  |  |
| --- | --- | --- | --- |
| 16 | Flow IAT Max | Maximum time between flows | Scale |
| 17 | Flow IAT Min | Minimum time between flows | Scale |
| 18 | Fwd IAT Total | Total time between forward packets | Scale |
| 19 | Fwd IAT Mean | Mean time between forward packets | Scale |
| 20 | Fwd IAT Std | Standard deviation of time between forward packets | Scale |
| 21 | Fwd IAT Max | Maximum time between forward packets | Scale |
| 22 | Fwd IAT Min | Minimum time between forward packets | Scale |
| 23 | Bwd IAT Total | Total time between backward packets | Scale |
| 24 | Bwd IAT Mean | Mean time between backward packets | Scale |
| 25 | Bwd IAT Std | Standard deviation of time between backward packets | Scale |
| 26 | Bwd IAT Max | Maximum time between backward packets | Scale |
| 27 | Bwd IAT Min | Minimum time between backward packets | Scale |
| 28 | Fwd PSH Flags | Forward packets with PUSH flags | Scale |
| 29 | Fwd Header Length | Length of header in forward packets | Scale |
| 30 | Bwd Header Length | Length of header in backward packets | Scale |
| 31 | Fwd Packets/s | Forward packets per second | Scale |
| 32 | Bwd Packets/s | Backward packets per second | Scale |
| 33 | Packet Length Max | Maximum length of packets | Scale |
| 34 | Packet Length Mean | Mean length of packets | Scale |
| 35 | Packet Length Std | Standard deviation length of packets | Scale |
| 36 | Packet Length Variance | Variance of length of packets | Scale |
| 37 | SYN Flag Count | Number of SYN flags | Scale |
| 38 | URG Flag Count | Number of URG flags | Scale |
| 39 | Avg Packet Size | Average packet size | Scale |
| 40 | Avg Fwd Segment Size | Average forward segment size | Scale |
| 41 | Avg Bwd Segment Size | Average backward segment size | Scale |
| 42 | Subflow Fwd Packets | Subflow forward packets | Scale |
| 43 | Subflow Fwd Bytes | Subflow forward bytes | Scale |
| 44 | Subflow Bwd Packets | Subflow backward packets | Scale |
| 45 | Subflow Bwd Bytes | Subflow backward bytes | Scale |
| 46 | Init Fwd Win Bytes | Initial forward window size | Scale |
| 47 | Init Bwd Win Bytes | Initial backward window size | Scale |
| 48 | Fwd Act Data Packets | Forward packets with actual data | Scale |

|  |  |  |  |
| --- | --- | --- | --- |
| 49 | Fwd Seg Size Min | Minimum segment size in forward packets | Scale |
| 50 | Active Mean | Mean active time | Scale |
| 51 | Active Std | Standard deviation of active time | Scale |
| 52 | Active Max | Maximum active time | Scale |
| 53 | Active Min | Minimum active time | Scale |
| 54 | Idle Mean | Mean idle time | Scale |
| 55 | Idle Std | Standard deviation of idle time | Scale |
| 56 | Idle Max | Maximum idle time | Scale |
| 57 | Idle Min | Minimum idle time | Scale |
| 58 | Label | The intrusion type | Targetclass: Multi- class classification |
| 59 | ClassLabel | Subtype of intrusion | Targetclass: Multi- class classification |

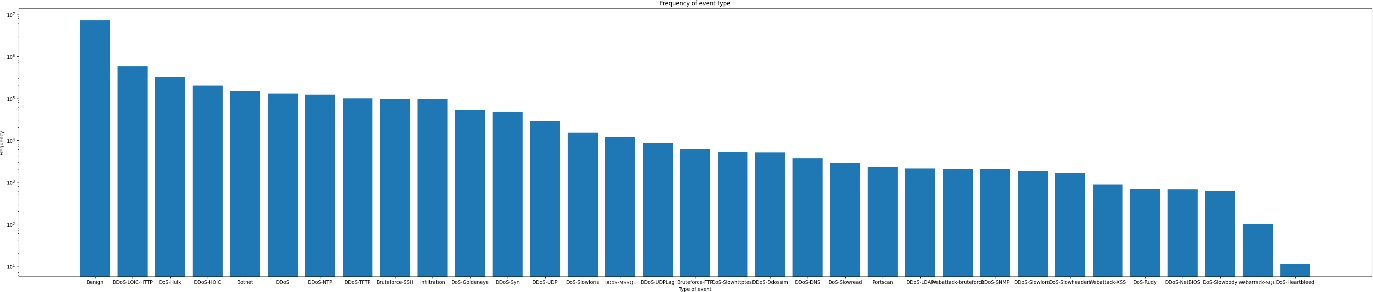


Figure 3.3.1 Bar chart of events in CIC dataset based on Label

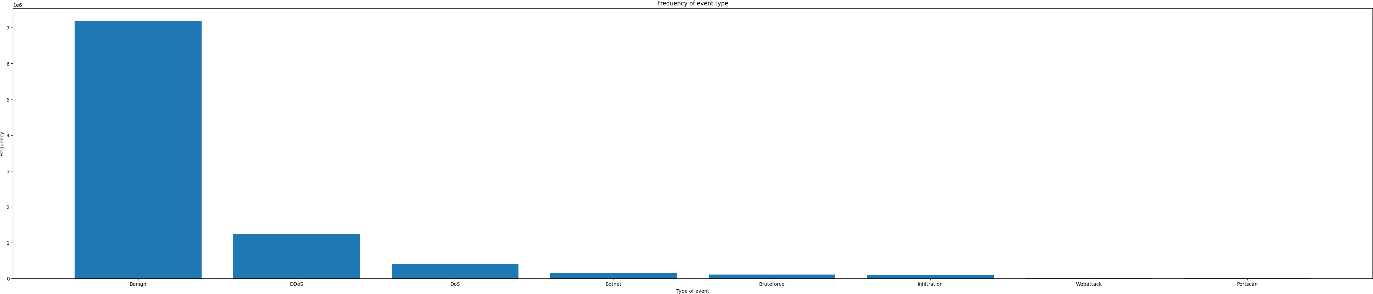


Figure 3.3.2 Bar chart of events in CIC dataset based on ClassLabel

Observations from above analysis: -

* + 1. We have high dimensional and large volume dataset.
    2. All independent features are of type scale. Thus, we need to normalize them prior training the model.
    3. The target features: Label and ClassLabel are highly imbalanced, the greatest number of events are of type Benign.
    4. In target feature: Label, there are 7186189 records of type Benign ~ 78% of the total records.
    5. Thus, for binary classification, we need to create a new field to differentiate between Benign and Malicious events.

Foreseen challenges with the dataset: -

1. Imbalanced nature of target features.

# Selection of dataset: -

We will build the project using CIC dataset. Reason for selection: -

* + 1. All independent features are of type scale. Thus, we can use normalization operation to handle the data.
    2. There are no independent features of type: ordinal or nominal, thus, we will not have more than the original number of features to analyze.
    3. We will be able to handle high dimensionality of the dataset using optimization approaches during feature selection.