## Dataset Setup

We will be using two datasets to simulate IoT

## UNSW-NB15

### UNSW-NB15 Dataset

The link to download this data set is : [https://unsw-my.sharepoint.com/personal/z5025758\_ad\_unsw\_edu\_au/\_layouts/15/onedUNSW-NB15 Dataset](https://unsw-my.sharepoint.com/personal/z5025758_ad_unsw_edu_au/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fz5025758%5Fad%5Funsw%5Fedu%5Fau%2FDocuments%2FUNSW%2DNB15%20dataset)

### Explanation of the Data Format

#### 2.1 Types of Attacks present in the data

The dataset has **9 types of cyberattacks**, including:

* Fuzzers (testing software for bugs)
* Analysis (e.g., data mining attacks)
* Backdoors (hidden ways to access systems)
* DoS (Denial of Service attacks)
* Exploits (using vulnerabilities to attack)
* Generic (general-purpose attacks)
* Reconnaissance (gathering information about systems)
* Shellcode (malicious code execution)
* Worms (self-replicating malware)
* It also has **normal network activities**, which are not attacks.

#### 2.2 Features present in the data

Tools like **Argus** and **Bro-IDS** (Intrusion Detection System) were used to analyze the network traffic. 49 features were extracted from this data

|  |
| --- |
| Source IP address |
| Source port number |
| Destination IP address |
| Destination port number |
| Transaction protocol |
| Indicates to the state and its dependent protocol, e.g. ACC, CLO, CON, ECO, ECR, FIN, INT, MAS, PAR, REQ, RST, TST, TXD, URH, URN, and (-) (if not used state) |
| Record total duration |
| Source to destination transaction bytes |
| Destination to source transaction bytes |
| Source to destination time to live value |
| Destination to source time to live value |
| Source packets retransmitted or dropped |
| Destination packets retransmitted or dropped |
| http, ftp, smtp, ssh, dns, ftp-data ,irc and (-) if not much used service |
| Source bits per second |
| Destination bits per second |
| Source to destination packet count |
| Destination to source packet count |
| Source TCP window advertisement value |
| Destination TCP window advertisement value |
| Source TCP base sequence number |
| Destination TCP base sequence number |
| Mean of the ?ow packet size transmitted by the src |
| Mean of the ?ow packet size transmitted by the dst |
| Represents the pipelined depth into the connection of http request/response transaction |
| Actual uncompressed content size of the data transferred from the server’s http service. |
| Source jitter (mSec) |
| Destination jitter (mSec) |
| record start time |
| record last time |
| Source interpacket arrival time (mSec) |
| Destination interpacket arrival time (mSec) |
| TCP connection setup round-trip time, the sum of ’synack’ and ’ackdat’. |
| TCP connection setup time, the time between the SYN and the SYN\_ACK packets. |
| TCP connection setup time, the time between the SYN\_ACK and the ACK packets. |
| If source (1) and destination (3)IP addresses equal and port numbers (2)(4) equal then, this variable takes value 1 else 0 |
| No. for each state (6) according to specific range of values for source/destination time to live (10) (11). |
| No. of flows that has methods such as Get and Post in http service. |
| If the ftp session is accessed by user and password then 1 else 0. |
| No of flows that has a command in ftp session. |
| No. of connections that contain the same service (14) and source address (1) in 100 connections according to the last time (26). |
| No. of connections that contain the same service (14) and destination address (3) in 100 connections according to the last time (26). |
| No. of connections of the same destination address (3) in 100 connections according to the last time (26). |
| No. of connections of the same source address (1) in 100 connections according to the last time (26). |
| No of connections of the same source address (1) and the destination port (4) in 100 connections according to the last time (26). |
| No of connections of the same destination address (3) and the source port (2) in 100 connections according to the last time (26). |
| No of connections of the same source (1) and the destination (3) address in in 100 connections according to the last time (26). |
| The name of each attack category. In this data set , nine categories e.g. Fuzzers, Analysis, Backdoors, DoS Exploits, Generic, Reconnaissance, Shellcode and Worms |
| Label : 0 for normal and 1 for attack records |

These 49 features are the header of the dataset given (**UNSW-NB15\_1.csv**, **UNSW-NB15\_2.csv**, **UNSW-NB15\_3.csv**, **UNSW-NB15\_4.csv). They have total of 2.5 million datapoints.**

#### 2.3 **Ground truth and events**:

* There's a file called **UNSW-NB15\_GT.csv** that provides the "ground truth" (labels indicating whether each record is normal or an attack).
* Another file, **UNSW-NB15\_LIST\_EVENTS**, lists specific events in the dataset.

#### 2.4 **Training and testing**:

The dataset is divided into:

* **Training set**: 175,341 records for training machine learning models.
* **Testing set**: 82,332 records for testing how well the models work.

### 3. Features used to train the model locally

the recommended list of features were used to train the neural network locally:

| **Feature** | **Description** |
| --- | --- |
| dur | Duration of the connection |
| sbytes | Source bytes transferred |
| dbytes | Destination bytes transferred |
| Sload | Source load |
| Dload | Destination load |
| Spkts | Source packets |
| Dpkts | Destination packets |
| smeansz | Source mean packet size |
| dmeansz | Destination mean packet size |
| sloss | Source packet loss |
| dloss | Destination packet loss |
| Sintpkt | Source inter-packet time |
| Dintpkt | Destination inter-packet time |
| swin | Source window size |
| dwin | Destination window size |
| tcprtt | TCP round-trip time |
| synack | SYN-ACK delay |
| ackdat | ACK delay |
| ct\_srv\_src | Connections to the same service from source |
| ct\_srv\_dst | Connections to the same service to destination |
| ct\_dst\_ltm | Connections to the same destination recently |
| ct\_src\_ltm | Connections to the same source recently |

### 4. Citation

If you are using this dataset then you need to cite this in the report mandatorily:

1. Moustafa, Nour, and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)."Military Communications and Information Systems Conference (MilCIS), 2015. IEEE, 2015.

2. Moustafa, Nour, and Jill Slay. "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set." Information Security Journal: A Global Perspective (2016): 1-14.