*What do I focus on when writing about dataset?*

* *Origin*
* *How it generated*
* *Y*
* *X*
* *Data Distribution (Number of records)*

Dataset, Pre-processing and Performance Metrics

4.1 Datasets

The models are experimented on three datasets: KDD’99, CIC-IDS-2017 and ToN\_IoT dataset. All of them are publicly available and widely used in Network Intrusion Detection Domain.

4.1.1 KDD99 dataset

**The KDD99 dataset, a popular benchmark for intrusion detection systems[1], was born in 1999's KDD Cup, a competition to build network intrusion models.** **Fueling this competition was data from the 1998 DARPA program, where MIT Lincoln Labs mimicked a real Air Force LAN for nine weeks, peppering it with attacks and capturing the raw traffic.This data was then transformed into roughly five million connection records, each representing a sequence of data flowing between two IP addresses under a specific protocol.Importantly, each record is labeled as either normal or a specific type of attack, making it a valuable resource for training and evaluating intrusion detection systems.**

**The KDD99 dataset categorizes network connection features into three distinct groups [4]:**

* **Basic features: Encompasses all directly extractable attributes from a TCP/IP connection.**
* **Traffic features: These features, computed within a designated window interval, further analyze network activity for anomaly detection. They are further subdivided into same host features and same service features.**
* **Content features: Those features were extracted by using domain knowledge for detecting suspicious behaviors in data portions of the packages.**

**Details of all features can be found at [3].**

There are four main categories of attack and 24 attack types:

|  |  |
| --- | --- |
| Category | Attack |
| DOS: Denial-Of-Service | Back, Pod, Land, Teardrop, Smurf, Neptune |
| R2L: unauthorized access from a remote machine | ftp\_write, guess\_passwd, imap, multihop, phf, spy, warezclient, warezmaster |
| U2R: unauthorized access to local superuser privileges | perl, buffer\_overflow, loadmodule, rootkit, |
| Probing: surveillance and other probing | Ipsweep, nmap, portsweep |

In this experiment, I use file kddcup.data\_10\_percent.gz [2] for training and testing purpose. This is a 10% subset of the full dataset. This dataset contains 494021 records in totals, in which 280790 records recognized as smurf attack.

smurf. 280790

neptune. 107201

normal. 97278

back. 2203

satan. 1589

ipsweep. 1247

portsweep. 1040

warezclient. 1020

teardrop. 979

pod. 264

nmap. 231

guess\_passwd. 53

buffer\_overflow. 30

land. 21

warezmaster. 20

imap. 12

rootkit. 10

loadmodule. 9

ftp\_write. 8

multihop. 7

phf. 4

perl. 3

spy. 2

[1] A review of KDD99 dataset usage in intrusion detection and machine learning between 2010 and 2015

[2] <https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>

[3] <https://kdd.ics.uci.edu/databases/kddcup99/task.html>

[4] A detailed analysis of the KDD CUP 99 data set