**Collecting and Preparing the dataset**

**Output:-Work dataset (modified KDD format) for one ip address, a text file containing the ip addresses**

**Collection and Processing of the dataset:-**

Step 1: The connection data was collected using **tcpdump**

Step 2: The collected data was then converted into the standard gureKDD dataset format which in turn was converted to KDDCup format.

**Detailed information about conversion of Tcpdump data to KDD Format:**

* The Sniff traffic data is collected with tcpdump.
* This data is passed to a .bro code, which computes the attributes, which defines the connections, intrinsic attributes and content attributes.
* The bro code converts this data and stores it in a list file.
* The list file contains the following attributes:

num\_conn, startTimet, orig\_pt, resp\_pt, orig\_ht, resp\_ht, duration, protocol, resp\_pt, flag, src\_bytes, dst\_bytes, land, wrong\_fragment, urg, hot, num\_failed\_logins, logged\_in, num\_compromised, root\_shell, su\_attempted, num\_root, num\_file\_creations, num\_shells, num\_access\_files, num\_outbound\_cmds, is\_hot\_login, is\_guest\_login.

* This list is sorted based on connection identifier number (num\_conn) which orders the connections by starting time.
* Finally, the sorted list is sent as an argument to a C Program .out file, which creates a list file containing the required gureKDDCup99 attributes, which is then converted to the KDDCup format by removing unnecessary attributes.

num\_conn, duration, protocol, src\_bytes, dst\_bytes, land, wrong\_fragment, urg, hot, num\_failed\_logins, logged\_in, num\_compromised, root\_shell, su\_attempted, num\_root, num\_file\_creations, num\_shells, num\_access\_files, num\_outbound\_cmds, is\_hot\_login, is\_guest\_login, count\_sec, srv\_count\_sec, serror\_rate\_sec, srv\_serror\_rate\_sec, rerror\_rate\_sec, srv\_error\_rate\_sec, same\_srv\_rate\_sec, diff\_srv\_rate\_sec, srv\_diff\_host\_rate\_sec, count\_100, srv\_count\_100, same\_srv\_rate\_100, diff\_srv\_rate\_100, same\_src\_port\_rate\_100, srv\_diff\_host\_rate\_100, serror\_rate\_100, srv\_serror\_rate\_100, rerror\_rate\_100, srv\_rerror\_rate\_100.

This collective tcpdump dataset is **split** into constituent datasets each having their **own destination ip address.**

For example if we wanted to analyse ip address **‘ip1’** then we extract all the rows from the collective tcpdump dataset which have destination ip address to be ‘ip1’ and form a new dataset containing all these extracted rows.

All these new modified datasets are then sent to **the prediction module(for preparing the forecasted data) for each individual ip address.**

A **separate text file containing the list of ip addresses** is also prepared from the tcpdump data. This textfile is then passed onto the **blocking module** for further processing.