CIC-IoT 2022

Dataset-Description:

The following illustrates the main objectives of the CIC-IoT dataset project:

- Configure various IoT devices and analyse the behaviour exhibited.
- Conduct manual and semi-automated experiments of various categories.
- Further analyse the network traffic when the devices are idle for three minutes and when powered on for the first two minutes.
- Generating different scenarios and analysing the devices' behaviour in different situations.
- Conducting and capturing the network terrific of devices undercurrent and important attacks in IoT environment.

For collecting the data, we captured the network traffic of the IoT devices coming through the gateway using Wireshark and dump-cap in six different types of experiments. The former was used for manual experiments, while the latter was used for semi-automated ones.

The dataset prioritizes benign device behaviour, with extensive coverage of idle and active states to model typical usage patterns. While the original study mentions 60 devices, our analysis reveals that the idle and active state data pertains exclusively to 40 LAN/Wired or Wi-Fi devices. Zigbee and Z-Wave devices are excluded from these states but are included in isolated power and interaction stages for specialized analysis.

All the experiments can be organized as follows:

1. Power: In this experiment, we powered on all the devices in our lab individually and started a network traffic capture in isolation.

2. Idle: In this experiment, we captured the whole network traffic from late in the evening to early in the morning, which we call idle time. In this period, the whole lab was completely evacuated and there were no human interactions involved.

3. Interactions: In this experiment, all possible functionality on IoT devices has been extracted and the corresponding network activity and transmitted packets for each functionality/activity have been captured.

4. Scenarios: In these experiments, we conducted six different types of scenario experiments using a combination of devices as simulations of the network activity

inside a smart home.

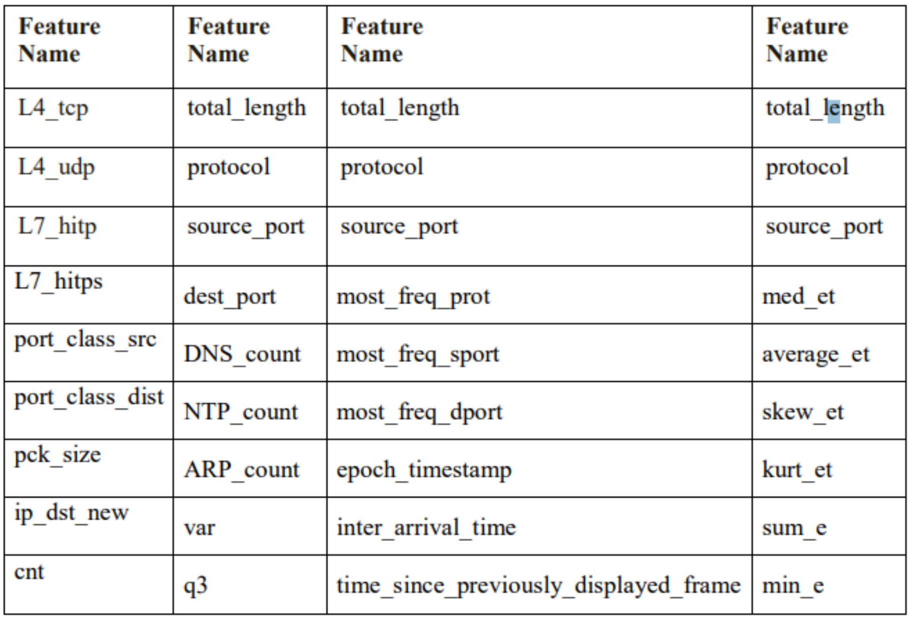
5. Active: In addition to the idle time, the whole network communications were also captured throughout the day. All fellow researchers during this period were allowed to enter the lab whenever they wanted.

6. Attacks: In this experiment, we performed two different attacks, Flood and RTSP-Brute Force, on some of our devices and captured their attack network traffic.

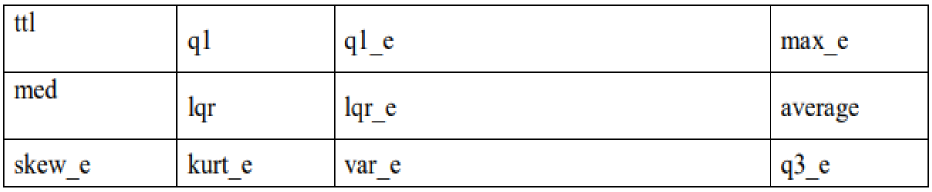
Feature Extraction:

A total of 48 features were extracted from the network traffic data, carefully selected for their relevance and discriminative power in device profiling and anomaly detection. These features encompass metrics such as packet size distributions, protocol-specific attributes, traffic timing patterns.

This structured approach ensures a robust foundation for training machine learning models, enabling accurate device identification, behavioural analysis, and intrusion detection in IoT environments.



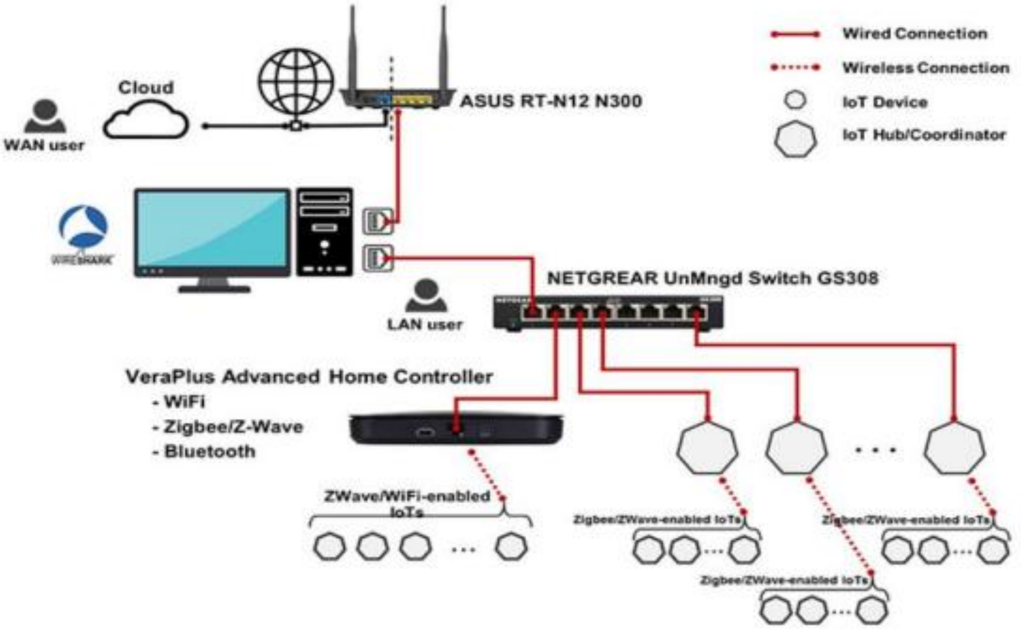
Feature Name	Feature Name	Feature Name	Feature Name
L4_tcp	total_length	total_length	total_length
L4_udp	protocol	protocol	protocol
L7_hitp	source_port	source_port	source_port
L7_hitps	dest_port	most_freq_prot	med_et
port_class_src	DNS_count	most_freq_sport	average_et
port_class_dist	NTP_count	most_freq_dport	skew_et
pck_size	ARP_count	epoch_timestamp	kurt_et
ip_dst_new	var	inter_arrival_time	sum_e
cnt	q3	time_since_previously_displayed_frame	min_e



ttl	q1	q1_e	max_e
med	lqr	lqr_e	average
skew_e	kurt_e	var_e	q3_e

Network configuration:

Our lab network configuration was configured with a 64-bit Window machine with two network interface cards - one is connected to the network gateway, and the other is connected to an unmanaged network switch. Hence, IoT devices that require an Ethernet connection are connected to this switch. Additionally, a smart automation hub, Vera Plus is also connected to the unmanaged switch, which creates our wireless IoT environment to serve IoT devices compatible with Wi-Fi, ZigBee, Z-Wave and Bluetooth.



Case study – device identification

After generating the dataset, we performed a case study on the idea of transferability – training datasets in our lab and transferring the trained model to another lab for testing. We conducted 20 different experiments based on the number of sampled devices from the United States lab.

Forty-eight features were extracted from both the training dataset from our lab and the testing dataset from the other lab. Three classes of device types were used in this experiment: Audio, Camera and Home Automation. However, no labels were required for the test dataset since that was what was to be predicted but the training dataset required labels.

After training, the model is transferred to the other lab for testing on each device to predict the class of the device in question. For example, if Amazon Echo Dot is tested on the trained model, the classifier should be able to predict this device as belonging to device type Audio. How this works is by counting the prediction of the classifier based on the features for each device type. The device type with the highest count is predicted as the class for the device in question.

Dataset Downloaded URL:

http://cicresearch.ca/IOTDataset/CIC_IOT_Dataset2022/Dataset/