Release Note: WiFi RF Datasets for Device Fingerprinting Using Deep Learning

Abdurrahman Elmaghbub and Bechir Hamdaoui
School of EECS, Oregon State University

Version 2, October 2023

1 Quick Dataset Download Links

Four WiFi 802.11B datasets (download links given below) are described and used in the paper titled EPS: Distinguishable IQ Data Representation for Domain-Adaptation Learning of Device Fingerprints.

The datasets can be downloaded and used for research, but we would like to request that any use that results in technical or other publications should include a citation to the following paper:

Copy and paste the bibtex below:

```
@article{elmaghbub2023eps,
title={EPS: Distinguishable IQ Data Representation for Domain-Adaptation Learning of Device Fingerprints},
author={Elmaghbub, Abdurrahman and Hamdaoui, Bechir},
journal={arXiv preprint arXiv:2308.04467},
year={2023}}
```

Click on the link corresponding to the setup you would like to download the dataset for:

- Scenario 1: Cross-Day Wired Scenario.
- Scenario 2: Cross-Day Wireless Scenario
- Scenario 3: Cross-Location Scenario
- Scenario 4: Random-Deployment Scenario

2 Brief Description

This WiFi fingerprint dataset has been collected at the NetSTAR lab at Oregon State University, as part of an NSF project in which we published several works in the effort of solving the RFFP problem [1], [2], [3], [4], [5], [6], [7], [8]. Our WiFi dataset contains WiFi 802.11b transmissions from 15 Pycom devices captured in four different setups/scenarios: Wired, Wireless, Differen Locations, and Random Deployment.

- **Scenario 1: Cross-Day Wired Scenario:** To rule out the impact of the wireless channel, we connected our transmitters directly to the USRP receiver via SMA cabling, and collected data over three days, generating more than 5000 WiFi frames per device every day. Wired-Dataset directory contains three subdirectories, each representing a different day. Within each day's subdirectory, there are 15 Hierarchical Data Format version 5 (HDF5) files corresponding to the 15 Pycom devices.
  Download Scenario 1 dataset: Cross-Day Wired Scenario.
- **Scenario 2: Cross-Day Wireless Scenario** Instead of wiring the transmitters to the USRP receiver as done in the Wired Setup, we placed them at a fixed location, 1m away from the USRP receiver which uses a VERT900 antenna to capture the signal. We repeated this experiment over three days to assess the

generalizability of the proposed technique over time. This setup generated more than 5000 WiFi frames per device every day. Wireless-Dataset directory contains three subdirectories, each representing a different day. Within each day's subdirectory, there are 15 Hierarchical Data Format version 5 (HDF5) files corresponding to the 15 Pycom devices.
Download Setup 2 dataset: Cross-Day Wireless Scenario.

- **Scenario 3: Cross-Location Scenario** The location from where the transmitter sends its data impacts the characteristics of the received signal, as signals transmitted from different distances/locations usually experience different channel conditions, which is considered in this work as another varying domain. For each transmitter, we then collected data at three different locations, A, B, and C, which are 1m, 2m, and 3m away from the USRP receiver, respectively. This was carried out in one day and generated more than 5000 WiFi frames per device at each location. Location-Dataset directory contains three subdirectories, each representing a location. Within each location's subdirectory, there are 15 Hierarchical Data Format version 5 (HDF5) files corresponding to the 15 Pycom devices.
  Download Scenario 3 dataset: Cross-Location Scenario.

- **Scenario 4: Random Deployment Scenario** we considered collecting datasets for two random-location scenarios on two different days, each consisting of an enrolment phase (data used for training) and a deployment phase (data used for testing). In both enrolment phases, all the transmitters transmitted from the same location, 1m away from the receiver, and in both deployment phases, the transmitters were located randomly within a radius of 3m away from the receiver. The enrollment datasets were collected in the morning while the random deployment datasets were collected on the night of that same day, generating more than 5000 WiFi frames per device for each dataset. The random-Location directory contains two subdirectories, representing the enrollment and deployment scenarios. Within each context's subdirectory, there are 15 Hierarchical Data Format version 5 (HDF5) files corresponding to the 15 Pycom devices.
  Download Scenario 4 dataset: Random Deployment Scenario.

3 Code Example

This is an example of using pytorch library to read the files from our dataset and build a dataset for training and testing deep learning models:

```
class HDF5Dataset(torch.utils.data.Dataset):
    def __init__(self, root_dir, window_size, transform=None):
        self.root_dir = root_dir
        self.window_size = window_size
        self.transform = transform
        self.samples = []
        for class_dir in os.listdir(root_dir):
            class_dir = os.path.join(root_dir, class_dir)
            #print(class_dir)
            for hdf5_file in os.listdir(class_dir):
                #print(hdf5_file)
                hdf5_file = os.path.join(class_dir, hdf5_file)
                with h5py.File(hdf5_file, 'r') as f:
                    data = np.concatenate(f['data'][:])
                    #data = data[0:52353600]
                for i in range(0, len(data) - 3*window_size, window_size):
                    self.samples.append((np.concatenate((np.array(data[i:i+window_size]),np.array(
                        data[i+2*window_size:i+2*window_size+window_size]))) , class_dir))
    def __len__(self):
        return len(self.samples)
    def __getitem__(self, idx):
        data, label = self.samples[idx]
        data = torch.tensor(data, dtype=torch.float32).view(2, 8390)
        if (self.transform):
            data = self.transform(data)
        label = os.path.basename(label)
        return data[:, 0:8192], int(label)
```

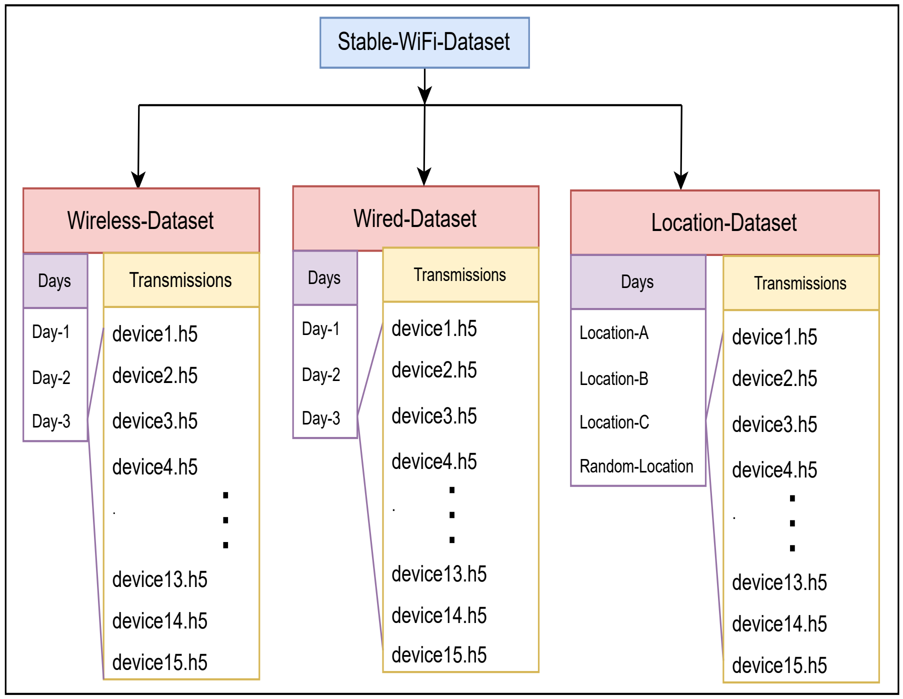

Figure 1: File/link structure and organization of the dataset. Notice that Stable-WiFi-Dataset directory is under rf-datasets directory.

4 Detailed Description

The testbed consists of 15 Pycom devices and an Ettus USRP B210 receiver, operating at a center frequency of 2.412GHz, used for recording the received signals sampled at 45MS/s. Our WiFi dataset contains 150GB of WiFi transmissions of 15 Pycom devices captured over 3 consecutive days in both wired and wireless connections and on 4 different locations. We used the GNURadio software to set up and configure the USRP receivers to capture WiFi transmissions, plot their time and spectrum domains, implement some preprocessing techniques, and store the samples in their files.

References

[1] Abdurrahman Elmaghbub, Bechir Hamdaoui, and Arun Natarajan. Widescan: Exploiting out-of-band distortion for device classification using deep learning. In *GLOBECOM 2020-2020 IEEE Global Communications Conference*, pages 1–6. IEEE, 2020.

[2] Abdurrahman Elmaghbub and Bechir Hamdaoui. Comprehensive RF dataset collection and release: A deep learning-based device fingerprinting use case. In *2021 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2021.

[3] Abdurrahman Elmaghbub and Bechir Hamdaoui. LoRa device fingerprinting in the wild: Disclosing RF data-driven fingerprint sensitivity to deployment variability. *IEEE Access*, 2021.

[4] Nora Basha, Bechir Hamdaoui, and Kathiravetpillai Sivanesan. Leveraging mimo transmit diversity for channel-agnostic device identification. In *ICC 2022-IEEE International Conference on Communications*, pages 2254–2259. IEEE, 2022.

[5] Jared Gaskin, Bechir Hamdaoui, and Weng-Keen Wong. Tweak: Towards portable deep learning models for domain-agnostic lora device authentication. *arXiv preprint arXiv:2209.00786*, 2022.

[6] Bechir Hamdaoui and Abdurrahman Elmaghbub. Deep-learning-based device fingerprinting for increased lora-iot security: Sensitivity to network deployment changes. *IEEE network*, 36(3):204–210, 2022.

[7] Bechir Hamdaoui, Abdurrahman Elmaghbub, and Siefeddine Mejri. Deep neural network feature designs for rf data-driven wireless device classification. *IEEE Network*, 35(3):191–197, 2020.

[8] Jun Chen, Weng-Keen Wong, Bechir Hamdaoui, Abdurrahman Elmaghbub, Kathiravetpillai Sivanesan, Richard Dorrance, and Lily L Yang. An analysis of complex-valued cnns for rf data-driven wireless device classification. *arXiv preprint arXiv:2202.09777*, 2022.