Release Note: Stable WiFi RF Datasets for Device Fingerprinting: Collected After Hardware Warm-up Period

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

Version 3, June 2024

1 Quick Dataset Download Links

This document presents four 15-Device WiFi 802.11B datasets that have been captured after the hardware warm-up period of each device (12 minutes after the activation of the devices). These datasets (download links given below) are described and used in the paper titled Distinguishable IQ Feature Representation for Domain-Adaptation Learning of WiFi Device Fingerprints in IEEE Transactions on Machine Learning in Communications and Networking.

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{elmaghbub2023dis,
author={Elmaghbub, Abdurrahman and Hamdaoui, Bechir},
journal={IEEE Transactions on Machine Learning in Communications and Networking},
title={Distinguishable IQ Feature Representation for Domain-Adaptation Learning of WiFi Device Fingerprints},
year={2024},
doi={10.1109/TMLCN.2024.3446743}}
```

The links to each of the tested setups are:

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

2 Dataset Description

These WiFi fingerprint datasets were collected at the NetSTAR lab at Oregon State University, as part of an NSF project [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 8, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 6]. The testbed used for collecting the datasets consists of 15 Pycom devices (both FiPy and LoPy4) 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 datasets contain 150GB of WiFi transmissions of 15 Pycom devices captured over 3 consecutive days in both wired and wireless connections and at 4 different locations for the different location scenarios. Utilizing GNURadio software, we configured USRP receivers to capture WiFi transmissions. Subsequently, we visualized their time and spectrum domains, applied preprocessing techniques, and stored the samples in files. Our investigation into hardware stabilization during warm-up periods informed this process [17]. As a result, all packets within these datasets were captured post-warm-up, hence we refer to them as stable datasets.

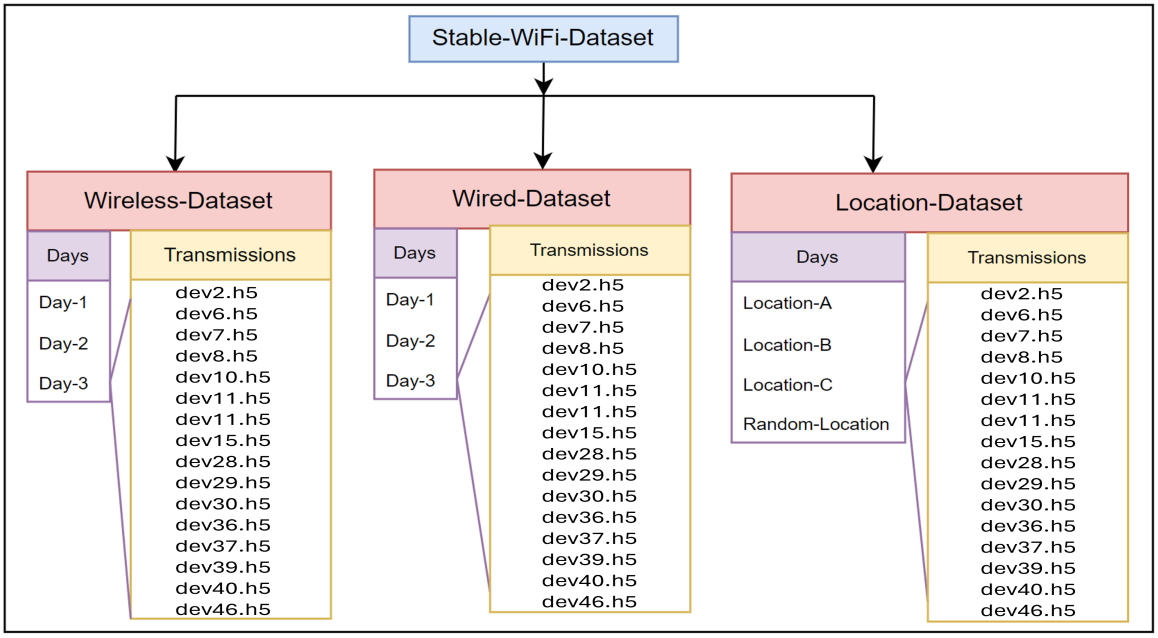

Figure 1: Structure and organization of the dataset: Wireless-Datasets, Wired-Datasets and Location-Datasets.

2.1 Description of the different scenarios

The datasets contain WiFi 802.11b transmissions from 15 Pycom devices captured in four different scenarios: Wired, Wireless, Different 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/device/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. Dataset link: Cross-Day Wired Scenario.
- **Scenario 2: Cross-Day Wireless Scenario:** Instead of wiring the transmitters to the USRP receiver, 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. This setup generated more than 5000 WiFi frames/device/day. Wireless-Dataset directory contains 3 subdirectories, each representing a different day and containing 15 HDF5 files (one for each of the 15 Pycom devices). Dataset link: Cross-Day Wireless Scenario.
- **Scenario 3: Cross-Location Scenario:** For each transmitter, we collected data at three different locations, A, B, and C, which are 1m, 2m, and 3m away from the USRP receiver, respectively. This is to allow the study of the impact of location. This experiment was carried out in one day and generated more than 5000 WiFi frames/device/location. Location-Dataset directory contains three subdirectories, each representing a location and having 15 HDF5 files (one for each device). Dataset link: Cross-Location Scenario.
- **Scenario 4: Random Deployment Scenario:** We collected datasets for testing random-location scenarios, with an enrolment phase (data for training) and a deployment phase (data for testing). In the enrolment phase, all the transmitters transmitted from the same location, 1m away from the receiver, and in the deployment phase, the transmitters were located randomly within 3m away from the receiver. The enrollment data was collected in the morning while the random deployment data was collected at night (of that same day). The Random-Location directory contains two subdirectories, representing the enrollment and deployment scenarios and each containing 15 HDF5 (one for each device). Dataset link: Random Deployment Scenario.

Refer to Fig. 1 for help with the organization and notation of the files.

2.2 Data Collection Description

We initiated the data-capturing process 12 minutes after the devices were activated, thereby ensuring an initial settling and warm-up period [17]. Each device was configured to operate over WiFi Channel 1 with a center frequency of 2412MHz and a bandwidth of 20MHz. The transmitters were programmed to transmit identical IEEE 802.11b frames with a duration of 559us back to back, separated by a small gap. We captured the first two minutes of transmissions using the USRP B210 at a sample rate of 45MSps. The captured signals were then digitally down-converted to the baseband and stored as IQ samples on our computer. To avoid any data dependency on the identity of the WiFi transmitter, all transmitters were configured to broadcast the same packets, which include the same spoofed MAC address and a payload of zero-bytes. Finally, we extracted the WiFi packets from the raw IQ sample files and stored them in HDF5 formatted files in the same order they were received. This method allowed us to maintain the integrity of the captured signals and ensured that they were accurately represented in the final dataset.

2.3 File format description

After storing the raw binary data, encompassing a continuous 2-minute capture session, we proceeded to extract the WiFi packets and archive them into HDF5 formatted files. We extracted packets by employing a power threshold and leveraging our knowledge of the packet length. This approach ensured precise capture of both the packet's inception and termination, while effectively filtering out noise and corrupted packets. Each HDF5 file comprises a dataset named 'data', structured with dimensions N x 50340. Here, N denotes the count of WiFi packets contained within the file. Meanwhile, the columns of this dataset correspond to time-domain In-phase (I) and Quadrature (Q) values of the packet. To specify, within each packet: The initial 25170 samples correspond to the I components of the packet, and the subsequent 25170 samples correspond to the Q components of the packet.

3 Code Example

This is an example of using Python to read a dataset from one of the files:

```
import h5py
#Change the file path
filePath = r'C:\Users\AbdurrahmanElmaghbub\Downloads\dev2.h5'
with h5py.File(filePath, 'r') as file:
    data = file['data'][:]              # 'data' is the name of the dataset.
# Now data has all the packets of this capture.
print(data.shape)  #(N, 50340)
```

References

[1] Abdurrahman Elmaghbub and Bechir Hamdaoui. Distinguishable IQ feature representation for domain-adaptation learning of WiFi device fingerprints. *IEEE Transactions on Machine Learning in Communications and Networking*, 2024.

[2] Abdurrahman Elmaghbub and Bechir Hamdaoui. Wireless device classification apparatus and method, March 26 2024. US Patent 11,943,003.

[3] 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.

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

[5] 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.

[6] Jared Gaskin, Bechir Hamdaoui, and Weng-Keen Wong. Tweak: Towards portable deep learning models for domain-agnostic lora device authentication. In *2022 IEEE conference on communications and network security (CNS)*, pages 1–9. IEEE, 2022.

[7] 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.

[8] Benjamin Johnson and Bechir Hamdaoui. On the domain generalizability of rf fingerprints through multifractal dimension representation. In *2023 IEEE Conference on Communications and Network Security (CNS)*, pages 1–9. IEEE, 2023.

[9] 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.

[10] 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.

[11] Jiaqi Bao, Bechir Hamdaoui, and Weng-Keen Wong. Iot device type identification using hybrid deep learning approach for increased iot security. In *2020 International Wireless Communications and Mobile Computing (IWCMC)*, pages 565–570. IEEE, 2020.

[12] Jun Chen, Weng-Keen Wong, and Bechir Hamdaoui. Unsupervised contrastive learning for robust rf device fingerprinting under time-domain shift. *arXiv preprint arXiv:2403.04036*, 2024.

[13] Jared Gaskin, Abdurrahman Elmaghbub, Bechir Hamdaoui, and Weng-Keen Wong. Deep learning model portability for domain-agnostic device fingerprinting. *IEEE Access*, 11:86801–86823, 2023.

[14] Abdurrahman Elmaghbub and Bechir Hamdaoui. A needle in a haystack: Distinguishable deep neural network features for domain-agnostic device fingerprinting. In *2023 IEEE Conference on Communications and Network Security (CNS)*, pages 1–9. IEEE, 2023.

[15] Luke Puppo, Weng-Keen Wong, Bechir Hamdaoui, and Abdurrahman Elmaghbub. Hinova: A novel open-set detection method for automating rf device authentication. In *2023 IEEE Symposium on Computers and Communications (ISCC)*, pages 1122–1128. IEEE, 2023.

[16] Abdurrahman Elmaghbub, Bechir Hamdaoui, and Weng-Keen Wong. Adl-id: Adversarial disentanglement learning for wireless device fingerprinting temporal domain adaptation. In *ICC 2023-IEEE International Conference on Communications*, pages 6199–6204. IEEE, 2023.

[17] Abdurrahman Elmaghbub and Bechir Hamdaoui. No blind spots: On the resiliency of device fingerprints to hardware warm-up through sequential transfer learning. In *Proceedings of the 17th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, pages 134–144, 2024.

[18] Bechir Hamdaoui, Nora Basha, and Kathiravetpillai Sivanesan. Deep learning-enabled zero-touch device identification: Mitigating the impact of channel variability through mimo diversity. *IEEE Communications Magazine*, 61(6):80–85, 2023.

[19] Nora Basha, Bechir Hamdaoui, Kathiravetpillai Sivanesan, and Mohsen Guizani. Channel-resilient deep-learning-driven device fingerprinting through multiple data streams. *IEEE Open Journal of the Communications Society*, 4:118, 2023.

[20] Bechir Hamdaoui and Abdurrahman Elmaghbub. Uncovering the portability limitation of deep learning-based wireless device fingerprints. *arXiv preprint arXiv:2211.07687*, 2022.

[21] Abdurrahman Elmaghbub and Bechir Hamdaoui. Eps: distinguishable iq data representation for domain-adaptation learning of device fingerprints. *arXiv preprint arXiv:2308.04467*, 2023.

[22] Luke Puppo, Weng-Keen Wong, Bechir Hamdaoui, Abdurrahman Elmaghbub, and Lucy Lin. On the extraction of rf fingerprints from lstm hidden-state values for robust open-set detection. *ITU Journal on Future and Evolving Technologies*, 5(1), 2024.

[23] 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.