Wireless Internet project
Wi-Fi encrypted traffic classification

Andrea Caravano

Academic Year 2023–24

1 Project requirements

1.1 Specifications

Implement a machine-learning classifier able to distinguish what kind of activity a user is performing with his/her smartphone/laptop by sniffing traffic in monitor mode.
The system should perform the following operations:

- Sniff traffic in monitor mode from a known MAC address
- Extract statistical features from the traffic every W seconds. The following traffic features can be extracted: number of packets up/down, average and variance of the packet size, average and variance of the inter-arrival packet times etc.
- Use a pre-trained machine-learning classifier of your choice to recognize the user activity among at least the following: idle, web browsing, YouTube streaming.
- Report the accuracy of the approach through a confusion matrix

2 Snippets

2.1 Traffic types

The final dataset is prepared alongside the target address (the one on which statistical indexes are extracted) and traffic types.

```
target_addr = '8a:c4:06:ee:1d:83' # target address on which we are applying the ML algorithm

dataset = [] # final datasets
traffic_types = ['youtube', 'speedtest', 'web', 'idle']
```

Chosen traffic types are:

- Idle: no relevant user activity

  Modern smartphones are still expected to exchange push/pull notifications frames and, eventually, reply to broadcast inquiries, if any
- Web browsing: news sources, emails and search engine queries
- YouTube video streaming: buffering is strongly used in the mobile context, so the packet flow is expected to be quite similar to the web browsing one
- Speedtest: several Ookla's Speedtests, the packet flow traits are expected to be quite unique

2.2 Data extraction

Valid packets are then kept and their fields stored to be finally gathered for statistical analysis.

```
# Processing of packets
for frame in cap:
  count += 1 # overall count
  success = False # malformed packets

  try:
    layers = frame.layers # pointer to layers
    timestamp = frame.sniff_timestamp # TS in ms
    sa = layers[2].sa # 802.11 frame: Source address
    da = layers[2].da # 802.11 frame: Destination address
    length = int(frame.length) # frame length
    success = True # packet is non malformed
  except:
    failed += 1 # processing failed: counting the packet as malformed
    success = False
```

As per project specifications, in both directions, the following statistical figures are used to form training and testing sets:

- Number of packets
- Average and variance of packet size
- Average and variance of inter-arrival packet times

The approach followed is the same applied for localization fingerprinting, from which the resulting outcomes are observed.

2.3 Classification

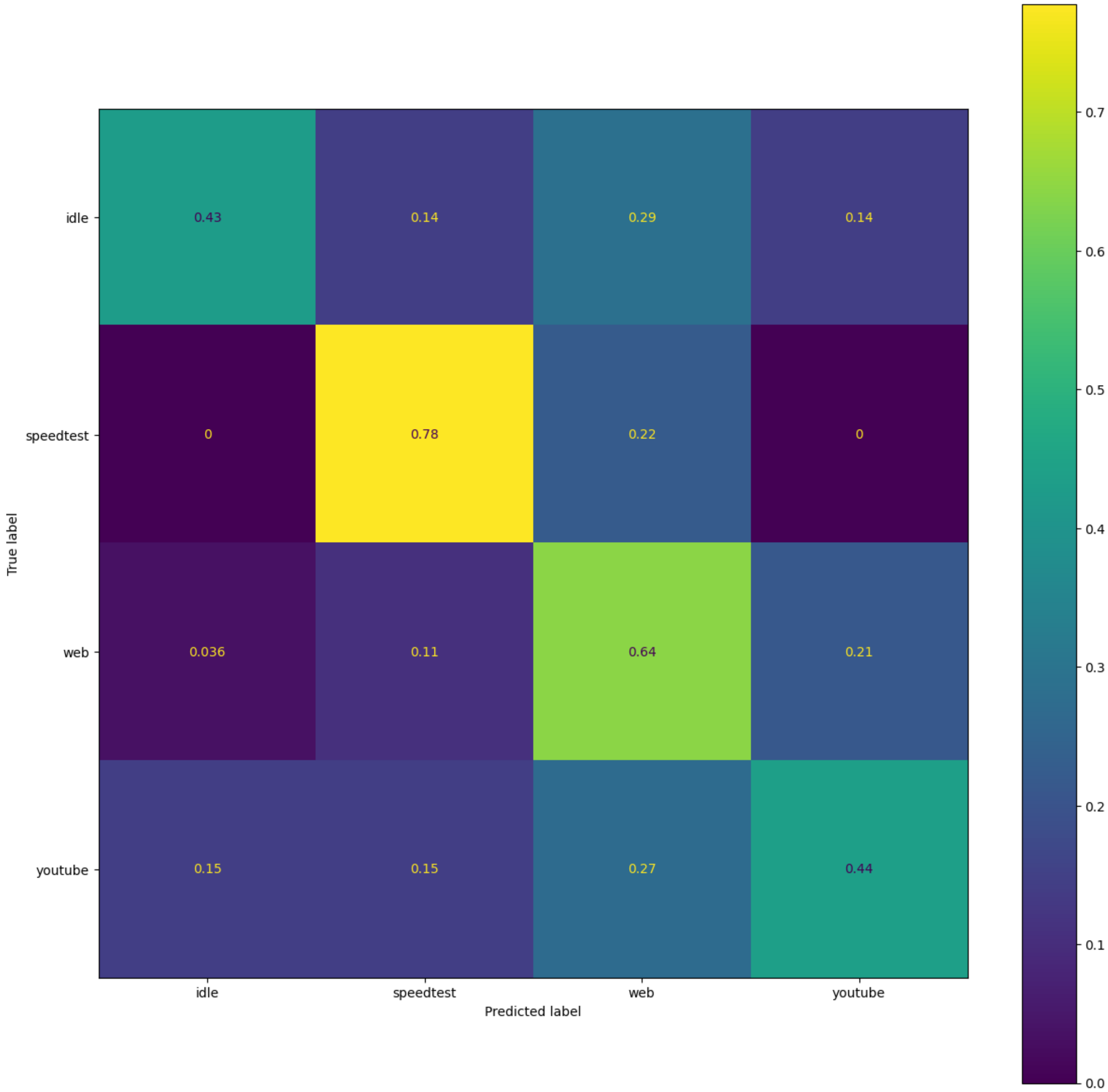
```
X_train, X_test, y_train, y_test = train_test_split(X_norm, Y, test_size=0.5) # 50%
↪  will be used for the training set, the rest for the testing set

# K-NeirestNeighbor (up to 4)
# as seen for localization fingerprinting
ACCURACY = []
for k in range(1,5):
  knn = KNeighborsClassifier(n_neighbors=k, weights='distance')
  knn.fit(X_train, y_train)
  knn_predict = knn.predict(X_test)
  accuracy = accuracy_score(y_test, knn_predict)
  ACCURACY.append(accuracy)

# we finally plot the confusion matrix
bestk = np.argmax(ACCURACY)+1
knn = KNeighborsClassifier(n_neighbors=bestk, weights='distance')
knn.fit(X_train, y_train)
knn_predict = knn.predict(X_test)
fig, ax = plt.subplots(figsize=(15, 15));
ConfusionMatrixDisplay.from_predictions(knn_predict, y_test, ax=ax, normalize='true');
```

3 Outcome

3.1 Confusion Matrix



3.2 Comment

The speedtest traffic classifier stands out as the most accurate, as anticipated.

Web traffic is the next one: its traffic figures are influenced by the adoption of buffering in the YouTube streaming case, resulting in mostly uniform characteristics.

YouTube streaming and idle traffic models have the highest imprecisions: this is also expected, since buffering hides the few bursts of traffic flowing, resulting in less distinguishable traffic traits.