

IoTDevID: A Behaviour-Based Fingerprinting Method for Device Identification in the IoT^{*}

Kahraman Kostas¹[0000–0002–4696–1857], Mike Just¹[0000–0002–9669–5067], and
Michael A. Lones¹[0000–0002–2745–9896]

Department of Computer Science, Heriot-Watt University, Edinburgh EH14 4AS, UK
{kk97, m.just, m.lones}@hw.ac.uk

Abstract. Device identification is one way to secure a network of IoT devices, whereby devices identified as suspicious can subsequently be isolated from a network. We introduce a novel fingerprinting method, *IoTDevID*, for device identification that uses machine learning to model the behaviour of IoT devices based on network packets. Our method uses an enhanced combination of features from previous work and includes an approach for dealing with unbalanced device data via data augmentation. We further demonstrate how to enhance device identification via a group-wise data aggregation. We provide a comparative evaluation of our method against two recent identification methods using three public IoT datasets which together contain data from over 100 devices. Through our evaluation we demonstrate improved performance over previous results with F1-scores above 99%, with considerable improvement gained from data aggregation.

Keywords: IoT security · IoT fingerprinting · Machine learning

1 Introduction

Internet of Things (IoT) is a collection of objects with embedded systems that can communicate with each other, wired or wirelessly. The “things” can be any physical items that we have used now or will use in the future [8]. IoT contributes greatly to human life in many critical areas such as [11] smart homes/cities, infrastructure, retail, healthcare, transportation, agriculture, military, petrochemical industry, and manufacturing. It is estimated that the number of IoT devices will reach 25 billion by 2021. By 2026, this number is predicted to reach 80 billion devices, and the total market will be about 1.1 trillion USD [10]. In other words, more than 150000 devices join the global IoT network every minute.

In parallel with this rapid development of devices and providers, security remains an important issue. Securing IoT devices can be challenging as a result of the heterogeneity of the devices [5, 8] and the difficulty of implementing traditional security solutions due to limited resources, such as processor, battery and bandwidth [18]. A device can have many sensors (temperature, humidity,

^{*} Supported by Republic of Turkey - Ministry of National Education

motion, light, etc.), and the channels it uses to communicate with other devices may need very different requirements. Thus, it is challenging to treat all IoT devices homogeneously. Although some research focuses on specialist areas such as home appliances [2, 15] or smart cities [6, 16], many devices have very different characteristics even under this classification. For example, baby monitors and smart kettles are both classified under home appliances.

Researchers have attempted to deal with this device heterogeneity by examining device behaviour, such as via the network packets with which they communicate. In this way, a device that is included in the network for the first time can be classified by conducting a behavioural analysis, based on the assumption that it will display similar behaviour with similar devices. In addition, this method is suitable for systems using encrypted protocols because it examines the packet properties and payload features, not the encrypted payload contents. It can be challenging to model device behaviour accurately as this method can depend on the variety and frequency of device communications.

In this paper, we introduce a new IoT device identification (fingerprint) method that models the behaviour of the network packets communicated by the devices. Our method, *IoTDevID*, and the comparative evaluation of our method, offers the following contributions. Firstly, our feature set improves on the two most notable existing fingerprint methods, IoT Sentinel [14] and IoT-Sense [3], allowing us to achieve more accurate device identification. Secondly, we incorporate data augmentation methods to address the very practical modelling challenges associated with unbalanced data across the set of devices. Thirdly, we use data aggregation techniques based on MAC addresses to perform a group-wise re-evaluation of our modelling results. We performed a comparative evaluation of our method with IoT Sentinel and IoTSense using three different IoT datasets containing more than 100 total devices and a half-dozen machine learning modelling algorithms. Based on our results, we discuss several considerations regarding identification and classification of IoT devices.

This paper is organised as follows: Section 2 reviews the related literature, Section 3 outlines the proposed fingerprinting method, *IoTDevID*, and describes the IoT datasets used in this study. Section 4 reports experimental results. The results are then discussed in Section 5, along with a discussion of limitations in Section 6, and Section 7 concludes.

2 Related Work

This section reviews previous studies that used fingerprinting methods to classify IoT devices. There are many studies in the literature on device identification using fingerprints, but their applicability to IoT devices is controversial, as these often focus on the physical layer or application layer where IoT has wide protocol variety [3]. Hence, we focus here on research that is based on network packet behaviour, including relevant information from each of the data link, network, and transport layers.

One of the first studies to use network packet features in a fingerprint method to identify IoT devices is *IoT Sentinel* [14]. This study uses network flow to identify vulnerable devices and isolate them from the user network. It involved the collection of benign data from 31 devices, with the data collected during the installation of the devices, and the installation process repeated 20 times for each device. The fingerprints specific to each device were created based on 23 features (see Fig. 1) extracted from each of the first 12 packets for each device, resulting in a fingerprint comprised of 276 values. These 12 packets do not exactly represent flow; they are sequential packets from the same MAC address. Each device connected to the network is identified by these fingerprints. If it is a vulnerable device, it is taken to the quarantine network and its connection to other devices, the local network, and the internet is restricted. This method is useful for identifying devices because the 23 features it creates are compressed/representative features: four of which are integers and the others binary-valued. In this study, 17 of 27 device types were detected with an identification accuracy of above 95%, and 10 with an accuracy of around 50% using Random Forest (RF).

IoTSense [4] uses selected features of *IoT Sentinel* based on their own design assessment. *IoTSense* chooses 17 protocol-based features of the *IoT Sentinel* study which reflect device behaviour and they also add three payload-related features (see Fig. 1). This feature list is applied to five packets for each device to produce a 100-member fingerprint, as an average of five packets were found to make up a session in this study. The packets are presumably aggregated according to common MAC address, though this is not explicitly stated in the paper. As a result of this study, per device recall of 93–100% and an average accuracy of 99% were achieved using Gradient Boosting (GB), Decision Tree (DT) and k-Nearest Neighbours (kNN). While some comparisons are made with the work of *IoT Sentinel*, the evaluation of *IoTSense* used a much smaller number of devices (i.e., 10 vs. 31). In addition, the *IoTSense* experiment set began with 14 devices, though only 10 devices were used for the evaluation as four devices did not produce sufficient data for the analysis approach used.

Meidan et al. [13] proposed two methods to distinguish IoT network traffic from non-IoT network traffic and to classify IoT devices. The experimental set up of this study consisted of nine IoT devices, two smartphones, one computer and one laptop. Four pairs of features derived from source and destination IP addresses, port numbers, SYN to FIN flags were utilized. Devices were classified as IoT and non-IoT with an average accuracy of 99.30% using GB and RF.

In another study [12], fingerprinting was used to classify traffic generated by IoT devices with Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) algorithms. In this context, more than 100 services and 266160 network flows were examined. In the classification stage, six features are focused on: the source port, destination port, number of bytes in packet payload, TCP window size (zero for UDP packets), interarrival time and direction of the packet (the authors stated that the IP addresses were not used for privacy reasons). IoT network traffic was classified with an average accuracy of 96.32%. However, the dataset is not shared and no information is provided about the device types.

Sivanathan et al. [17] used the data obtained from a reasonably large variety of 28 IoT devices (such as cameras, lights, plugs, motion sensors, appliances and health-monitors) during six months to classify IoT devices. Eight features were used to make this classification: flow volume, flow duration, average flow rate, device sleep time, server port numbers, Domain Name Server (DNS) queries, Network Time Protocol (NTP) queries and cipher suites. In this study, Naïve Bayes (NB) was used in the first step of the two-step classification system and RF was used in the second step. As a result, 28 IoT devices were classified with an accuracy of 99.88%. However, this study can be criticized for some elements of its feature set being too device specific (e.g., server port numbers, DNS queries, and cipher suites) and thereby not focusing on device behaviour.

3 Proposed Method

In this section we introduce our proposed fingerprint method for device identification, *IoTDevID*. In particular, we describe the feature set that we use for modeling network packet data and describe an approach for aggregating individual packets. We also introduce the datasets that we use for the comparative evaluation of our method. Note that we present our application of data augmentation to our dataset along with our results in Section 4.2.

3.1 Feature Selection

The IoTDevID¹ method was created using a feature set taken from the 23 features of IoT Sentinel and two of the payload features from IoTSense. Fig. 1 shows the resulting feature list and its relation to the other studies. We provide our rationale for using these features below.

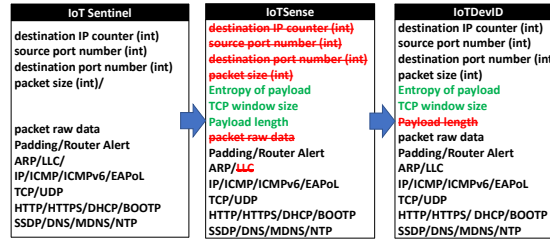


Fig. 1: The IoTDevID features list and its relation with the other studies.

Of the 23 features (see Fig. 1) used by IoT Sentinel [14], four are integers (packet size, destination IP counter, source port, destination port) and the remaining 19 are binary values. These binary values determine if a protocol was

¹ Source code available at: github.com/kahramankostas/IoTDevID

used in a packet (1 = in use, 0 = not in use). IoTSense [4] uses many protocol-based features of the IoT Sentinel study, suggesting that they could be useful features that reflect device behaviour (such as ARP, SSL, LLC, EAPOL, HTTP, MDNS and DNS, 17 features in total). On the other hand, they removed IP addresses and the target and source port numbers as they felt that these features are too specific and ineffective at distinguishing device behaviour. In addition to these features, three additional features from the payload are included in the fingerprint: entropy of payload, payload length and TCP window size. The features used are listed in Fig. 1.

We decided to retain the use of IP addresses and port number related features since a closer inspection revealed that IoT Sentinel does not use IP addresses and port numbers directly. Rather, it uses a counting process to summarise IP address usage, and assigns port numbers to classes. With this method, they avoid unnecessary and overly-specific information such as IP address and port number, whilst obtaining important information such as IP address and port number diversity. Further, IoTSense does not use the packet size feature, and defines a new feature named payload length. However, there is already a high correlation between payload length and packet size ($payloadlength = packetsize - headersizes$).

3.2 Aggregating Individual Packets

In our fingerprint approach, we use individual packets rather than network flow as there is no standard for the size of the network stream (number of packets, duration, etc.). Therefore, the use of individual packets will be beneficial in terms of speed and efficient use of limited resources [9]. However, it is very difficult to define a device using individual packets. To overcome this difficulty, other device identification studies [3, 14] have used MAC addresses to create pseudo network flows by aggregating packets, though the implementations of this approach are often not described in sufficient detail. Differing from these studies we use the MAC addresses to aggregate the ML results from individual packets, for which we provide a corresponding algorithm (see Algorithm 1).

This algorithm takes the MAC addresses of the devices on the network $M = \{m_1, m_2, \dots, m_n\}$ and the result of the machine learning (ML) algorithm $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n\}$ as input. It creates groups of size g from packets grouped according to MAC addresses. The results of packets originating from the same MAC address are collected under the same group, and then the algorithm re-evaluates these groups. During this re-evaluation process, the mode (the most frequent element) of each group is assigned to the entire group as a new value. The algorithm gives the modified results as outputs $\hat{Y}' = \{\hat{y}'_1, \hat{y}'_2, \dots, \hat{y}'_n\}$.

3.3 IoT Data Selection

We have found three different datasets that contain real device data, available for public use, that can be used in device identification. Their names, years of

Algorithm 1 Aggregation Algorithm

```

1:  $devices_{ij} \leftarrow []$  ▷ create an empty two-dimensional array
2:  $seen \leftarrow \emptyset$ 
3:  $g \leftarrow 12$  ▷ specifying the group size
4: for each  $m \in M$  do
5:   if  $m \notin seen$  then ▷ detect first-time  $m$  is seen
6:      $seen \leftarrow seen \cup m$ 
7:      $i \leftarrow |seen| - 1, j \leftarrow 0$ 
8:      $devices_{ij} \leftarrow m$  ▷ create new row in device and assign  $m$  as the first element
9:   for  $j \leftarrow 0, \text{length}(\hat{Y})$  do
10:    for  $i \leftarrow 0, |seen|$  do
11:      if  $device_{i0} == m_j$  then
12:         $device_i \leftarrow device_i \cup j$  ▷ assign the index of  $m_j$  to  $device_i$ 
13:    for  $i \leftarrow 0, |seen|$  do
14:       $C \leftarrow []$  ▷ will hold the indices of  $m$ s divided into chunks
15:      for  $j \leftarrow 1, \text{length}(device_i) - g, \text{Step} = g$  do
16:         $C \leftarrow C \cup device_{i[j:j+g]}$  ▷ divide  $device_i$  into chunks of  $g$ 
17:      for each  $c \in C$  do
18:         $g\_list \leftarrow []$ 
19:        for each  $j \in c$  do
20:           $g\_list \leftarrow g\_list \cup \hat{y}_j$  ▷ assign  $\hat{y}$  sharing same index with  $m$ 
21:         $mode \leftarrow \text{mode}(g\_list)$ 
22:        for  $j$  in  $c$  do
23:           $\hat{y}'_j \leftarrow mode$ 

```

creation, and the number of devices they contain are as follows: Aalto University [14], 2016, 27 devices²; UNSW-Sydney IEEE TMC [17], 2016, 31 devices; IoTFinder [2], 2018, 51 devices. The Aalto University dataset contains only the device installation data, but this installation process was repeated 20 times for each device to increase the data amount [14]. For other datasets, experiment setups that imitate the normal working environments of IoT devices were established. The Aalto University dataset has the advantage of having a sufficiently large set of labelled data that is grouped into a number of device types. Where appropriate, we present our results with a focus on this dataset for clarity.

4 Experiments and Analysis

In this section we perform an iterative, comparative evaluation of the IoTDevID method. First, we evaluate the performance of our method using several different machine learning (ML) algorithms and metrics. Second, we improve the balance of data per device in our dataset via data augmentation. Thirdly, we perform a group-wise aggregation of our data packets.

4.1 Initial Fingerprint Method Evaluation

Once fingerprints are extracted from IoT data, and a ML algorithm is used to provide discrimination. Based on earlier approaches (see Section 2) and surveys

² In the Aalto University dataset, D-LinkDoorSensor & D-LinkHomeHub and Hue-Bridge & HueSwitch device pairs have the same MAC address (a shared hub address). We labelled these device pairs as single devices. Therefore, 25 different labels were used instead of 27.

of related work [1, 8], we selected the following group of algorithms to use: RF, NB, kNN, GB, DT, and SVM (Support Vector Machine). We used the RandomizedSearchCV function from scikit-learn to find suitable hyperparameters for each algorithm. Using these six algorithms, we classified the 25 device types from the Aalto University dataset using the fingerprints produced by the three methods: IoT Sentinel, IoTSense, IoTDevID. We recreated the IoT Sentinel and IoTSense feature sets based on the descriptions in the associated publications. Since these three studies combine packets with different sizes and methods, we made the comparison only on individual packets. Table 1 shows the results for the Aalto University dataset.

Table 1: Comparison of ML algorithm results with average and standard deviation (SD) of 10-repeat 10-fold CV (Cross-Validation) on the Aalto University dataset. Time indicates testing duration. The overall best result for each metric is underlined.

ML	Method	Accuracy	Time (ms)	Precision	Recall	F1 score
DT	IoTDevID	0.77 \pm 0.00	2.69	0.70 \pm 0.01	0.65 \pm 0.01	0.67 \pm 0.01
	IoTSense	0.69 \pm 0.01	2.37	0.61 \pm 0.01	0.53 \pm 0.01	0.55 \pm 0.01
	IoT Sentinel	0.72 \pm 0.01	2.27	0.67 \pm 0.01	0.60 \pm 0.01	0.62 \pm 0.01
GB	IoTDevID	0.45 \pm 0.03	44.97	0.13 \pm 0.02	0.15 \pm 0.02	0.12 \pm 0.02
	IoTSense	0.37 \pm 0.03	42.86	0.07 \pm 0.03	0.11 \pm 0.02	0.07 \pm 0.02
	IoT Sentinel	0.43 \pm 0.02	38.45	0.10 \pm 0.02	0.13 \pm 0.01	0.11 \pm 0.02
kNN	IoTDevID	0.76 \pm 0.01	832.01	0.67 \pm 0.02	0.65 \pm 0.01	0.65 \pm 0.01
	IoTSense	0.66 \pm 0.01	1362.49	0.57 \pm 0.02	0.53 \pm 0.01	0.52 \pm 0.01
	IoT Sentinel	0.70 \pm 0.00	462.60	0.64 \pm 0.01	0.58 \pm 0.01	0.60 \pm 0.01
NB	IoTDevID	0.23 \pm 0.01	2.37	0.11 \pm 0.01	0.17 \pm 0.01	0.09 \pm 0.01
	IoTSense	0.06 \pm 0.00	2.84	0.07 \pm 0.02	0.11 \pm 0.01	0.04 \pm 0.00
	IoT Sentinel	0.36 \pm 0.01	2.48	0.24 \pm 0.01	0.25 \pm 0.01	0.22 \pm 0.00
RF	IoTDevID	0.78 \pm 0.00	184.88	0.70 \pm 0.01	0.66 \pm 0.01	0.68 \pm 0.01
	IoTSense	0.69 \pm 0.00	155.79	0.61 \pm 0.01	0.53 \pm 0.01	0.55 \pm 0.01
	IoT Sentinel	0.73 \pm 0.00	141.47	0.67 \pm 0.01	0.61 \pm 0.01	0.63 \pm 0.01
SVM	IoTDevID	0.75 \pm 0.01	29547.00	0.70 \pm 0.01	0.62 \pm 0.01	0.62 \pm 0.01
	IoTSense	0.68 \pm 0.01	28863	0.62 \pm 0.01	0.51 \pm 0.01	0.54 \pm 0.01
	IoT Sentinel	0.71 \pm 0.00	32092.00	0.67 \pm 0.01	0.58 \pm 0.00	0.62 \pm 0.01

It can be seen that respectively RF, DT, kNN and SVM algorithms have the highest accuracy for all three fingerprinting methods. The fastest algorithms are DT and NB, respectively. However, although fast, the accuracy level of the NB algorithm is very low. Although the kNN algorithm is satisfactory with a high accuracy rate, we can say that it is not practical to use due to its slowness. The GB algorithm has low accuracy and is also slow. The SVM algorithm is also not a reasonable option in terms of speed. Based on these observations, we use RF in the remainder of this work, since it is the most efficient algorithm in terms of accuracy, the most commonly used evaluation criteria. However, accuracy cannot be calculated for a class-based assessment. So, recall is used as an alternative evaluation method to accuracy for per-class evaluations. Overall, although the IoTDevID method is generally slower than the other two methods, it provides the most accurate results, with a difference of about 5% from its closest competitor.

4.2 Augmentation to Balance the Data

Table 2 shows the number of packets produced by the devices and their percentage in the Aalto University dataset. In this context, the dataset contains 107,652 packet/row data from a total of 25 type devices. However, the contribution of each device to this dataset is at different rates. For example, iKettle2 generates 0.12% of the dataset with 188 packets, while HueSwitch produces one third of the total data. This leads to an unbalanced distribution.

Table 2: Number of packets of IoT devices, their proportions over the entire dataset and recall values per device according to fingerprint method.

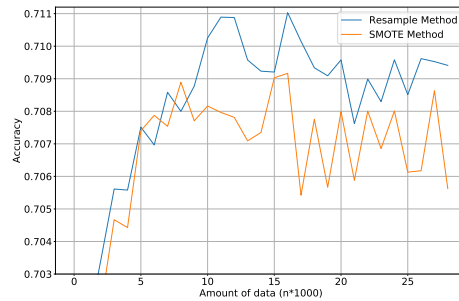
Device Name	Packet Statistics		Device Recall		
	Packets	Percent	IoT DevID	IoT Sense	IoT Sentinel
Aria	520	0.48%	0.763	0.731	0.591
D-LinkCam	6,358	5.91%	0.883	0.806	0.873
D-LinkDayCam	1,235	1.15%	0.794	0.682	0.791
D-LinkHomeHub	10,678	9.92%	0.783	0.829	0.664
D-LinkSensor	6,633	6.16%	0.419	0.409	0.337
D-LinkSiren	6,289	5.84%	0.380	0.222	0.280
D-LinkSwitch	6,614	6.14%	0.605	0.557	0.485
D-LinkWaterSensor	6,538	6.07%	0.447	0.105	0.393
EdimaxCam	896	0.83%	0.797	0.708	0.773
EdimaxPlug1101W	1,247	1.16%	0.541	0.444	0.369
EdimaxPlug2101W	1,131	1.05%	0.434	0.310	0.284
EdnetCam	390	0.36%	0.607	0.514	0.606
EdnetGateway	850	0.79%	0.676	0.540	0.553
HomeMaticPlug	639	0.59%	0.956	0.713	0.956
HueSwitch	32,581	30.27%	0.989	0.983	0.986
Lightify	4,384	4.07%	0.974	0.978	0.974
MAXGateway	634	0.59%	0.886	0.683	0.885
SmarterCoffee	190	0.18%	0.153	0.000	0.135
TP-LinkPlugHS100	729	0.68%	0.528	0.441	0.536
TP-LinkPlugHS110	698	0.65%	0.501	0.196	0.506
WeMoInsightSwitch	6,077	5.65%	0.797	0.493	0.713
WeMoLink	6,769	6.29%	0.859	0.731	0.830
WeMoSwitch	4,607	4.28%	0.829	0.451	0.784
Withings	777	0.72%	0.746	0.717	0.733
iKettle2	188	0.17%	0.095	0.000	0.087

Choosing the Right Evaluation Criteria One of the biggest problems that this imbalance will cause is misleading results. If the evaluation results per device are considered, it is seen that the performance of devices that represent a large proportion of the dataset is very high (see Table 2). For example, HueSwitch represents approximately one third of the total data, and its success rate is 99%. On the other hand, iKettle and SmarterCoffee devices make up 3.5% of the total data and their performance is very low (10% and 15%). This distribution causes the accuracy to appear much higher than it is. We used balanced accuracy and F1-score to deal with this problem. Balanced accuracy is the average of the recall of all classes. F1-score is the harmonic mean of recall and precision. In this context, we can update the values given in Table 1 with Table 3.

Table 3: Comparison of fingerprinting methods using RF

Fingerprint Method	Balanced Accuracy	F1-Score	Time (ms)
IoTDevID	0.658 \pm 0.008	0.673 \pm 0.008	184.884
IoTSense	0.530 \pm 0.007	0.552 \pm 0.007	155.795
IoT Sentinel	0.605 \pm 0.008	0.628 \pm 0.008	141.472

Data Augmentation Although the correction on evaluation criteria (see Table 3) allows us to see the performance more clearly, the unbalanced distribution in the dataset remains, and it has the potential to affect the performance of the ML algorithms being applied. Many methods can be used to eliminate this imbalance [7]. We used the resampling (Sk-resample) and synthetic data generation (SMOTE — Synthetic Minority Oversampling TEchnique) methods to make the distribution in our dataset more balanced. Resampling duplicates existing samples, and SMOTE generates new samples between randomly chosen neighbouring ones. These augmentation methods were used to observe the effect of data set size on test accuracy (see Fig. 2). An increase in accuracy is observed up to 10000 samples and, beyond this level, the results fluctuate. Therefore, 10000 samples per device was determined as the ideal training size. Test data size was decided as 3000. During the augmentation process, training and testing data were pre-parsed and isolated from each other. Table 4 shows the performance after data augmentation for IoTDevID. According to these results, the accuracy rate has increased by approximately 2% (see Table 4). The two data augmentation methods used do not differ significantly from each other in terms of accuracy, but resampling is faster.

**Fig. 2:** Effect of change in data quantity on accuracy.

Inclusion of Other IoT Data Below, we apply augmentation to the other two datasets in addition to the above application to the Aalto University dataset. For both datasets, 10000 fingerprints per device were used for training and 3000 for testing, using resampling where the amount of available data for a device

Table 4: Comparison with original and augmented data over 100 repeats.

Fingerprint Method	Accuracy	F1-Score	Time (ms)
IoTDevID	0.658 ± 0.008	0.673 ± 0.008	184.840
IoTDevID+Resampling	0.681 ± 0.001	0.696 ± 0.001	903.156
IoTDevID+SMOTE	0.681 ± 0.001	0.696 ± 0.001	988.785

was less than this. Data augmentation was required for seven of the 31 devices in UNSW-Sydney IEEE TMC, and three of the 51 devices in IoTFinder.

The average of the results obtained from Aalto University and other two datasets are presented in Table 5. The F1-scores per device are shown in Fig. 3a–c. In the overall results, the performance of the two newly included datasets are much higher than the Aalto University dataset. In device-based results, while most of the devices of these two datasets were identified with accuracies of 80% and above, less than half of the devices in the Aalto University dataset achieved this. The probable reason for this large difference is that the behaviour of some devices used in the Aalto data are very similar to each other and therefore cannot be easily discriminated (see Section 5 for further discussion).

Table 5: Results on the three datasets using data augmentation.

Dataset	Accuracy	F1-Score
Aalto University	0.681 ± 0.001	0.696 ± 0.001
UNSW-Sydney IEEE TMC	0.862 ± 0.000	0.882 ± 0.000
IoTFinder	0.892 ± 0.000	0.906 ± 0.000

4.3 Results of Aggregated Packets

So far in Section 4, we have classified devices using individual packets. In this section, we re-evaluate the results of the ML algorithm by using the MAC address, which is a unique identity of the device carried by each packet. For this process, we use the aggregating algorithm (see Algorithm 1) that we introduced in Section 3.2.

We repeated the experiments on the datasets, this time using the packet aggregating algorithm. We tested four values of group size ($\text{Group}/g = \{3, 6, 9, 12\}$). The results obtained from this process are shown in Table 6. It can be seen that packet aggregation significantly improves the accuracy, with an increasing benefit from using larger group sizes. Fig. 3d–f shows the per-device evaluation for $g = 12$. The only device not showing improvement is SmarterCoffee; see Section 5 for further discussion.

5 Discussion

In our study, we performed device identification by using individual packets that reflect the device behaviours via ML algorithms. Among the ML algorithms we

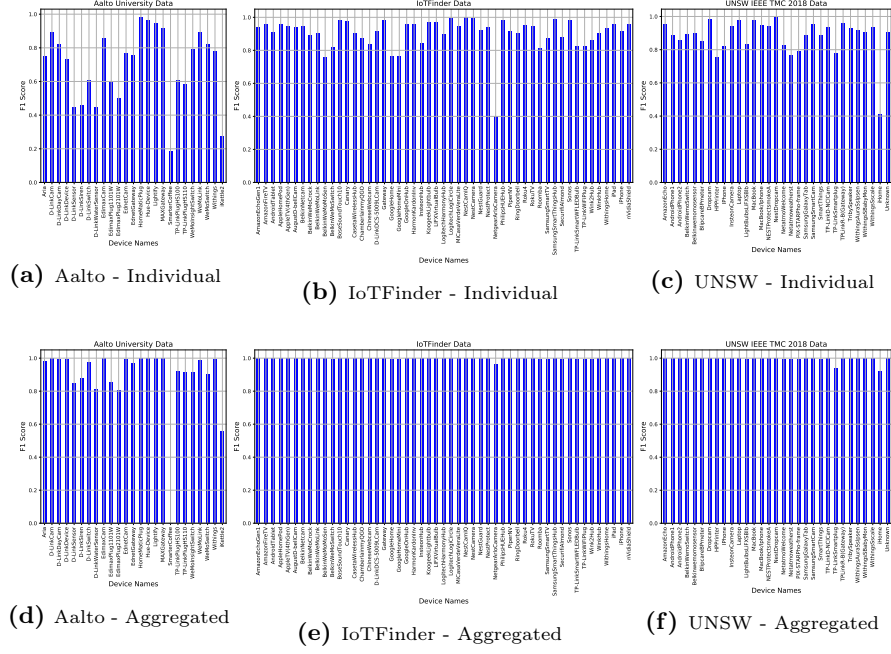


Fig. 3: Per-device F1-Scores with individual and aggregated method ($g = 12$).

Table 6: Accuracy and SD of aggregated packets for different groups using RF.

Dataset	Group			
	12	9	6	3
Aalto University	0.900 \pm 0.001	0.870 \pm 0.001	0.842 \pm 0.001	0.750 \pm 0.001
UNSW IEEE TMC	0.995 \pm 0.000	0.988 \pm 0.000	0.982 \pm 0.000	0.948 \pm 0.000
IoTFinder	0.998 \pm 0.000	0.986 \pm 0.000	0.983 \pm 0.000	0.956 \pm 0.000

used, RF has the highest accuracy rate in all of our experiments. In terms of accuracy, kNN and SVM, which are among the other algorithms that are close to RF, are disadvantageous in terms of the length of testing times. However, the testing time for DT, which displayed accuracy rates very close to RF, is quite notable. This time varied between 1/60 and 1/100 of the RF testing time in all the datasets. In this context, DT could be an alternative to RF in applications where speed is particularly important with satisfactory accuracy.

We applied data augmentation techniques to deal with the unbalanced data structure in the datasets we used. With this process, a 2% increase in overall accuracy was achieved in the Aalto University dataset, which is the most unbalanced dataset we used. However, this process did not contribute equally to every device. Out of 25 devices, there was a significantly (3% and more) increased recall for 10 devices, while decreases (3% and more) were observed in 6

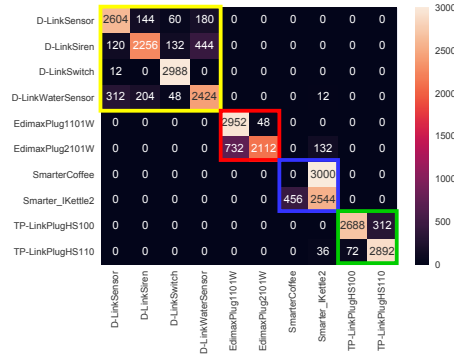


Fig. 4: The confusion matrix of low-performance devices. Groups 1: yellow, 2: red, 3: blue, 4: green

devices. For example, the recall of iKettle2, which is one of the devices with the least data, increased from 10% to 57%, while the D-LinkHomeHub device with the most data decreased from 78% to 69%. However, we did not see a general relationship between the initial amount of data and the change in performance.

As a result of data augmentation and aggregating, there has been a significant increase in accuracy for all the datasets we used. However, although we achieved performance above 99% in almost all of the datasets we worked with, our success rate in Aalto University data remained around 90%. When we focus on the results to determine the reason for this, we can see that low performance is concentrated in some device subgroups. A confusion matrix/heat-map containing only low-performance devices is given in Fig. 4. The point to be considered here is that these data have considerable similarities between the devices that make up these sub-groups: they are either similar purpose devices manufactured by the same companies (e.g., Group 1 and Group 3 in Fig. 4) or different models of the same device (e.g, Group 2 and Group 4 in Fig. 4).

It is possible that these devices use very similar hardware and software, so these devices exhibit similar behaviour. It does not seem possible to separate these devices according to their behaviour perfectly. Since they already use similar hardware and software, the vulnerabilities and prevention tend to be similar [14]. It may be a reasonable solution to put these devices under a single group to evaluate them. When we redo the evaluation by taking these device names under one device group, we see that the accuracy increases to 89.38% for individual packets, and 99.21% for aggregated packets (see Table 7).

In Table 7, the final results obtained from all datasets and the number of devices they contain are shared. The fact that the data set where we achieve the highest performance in both individual packets and the aggregated method is also the smallest data set (see final two rows of Table 7) raises the question of whether there is a relationship between the number of devices and successful device identification. Increasing the number of devices in the dataset will reduce the likelihood of correct classification of devices by chance. Therefore, it is usual to have an inversely proportional relationship between the number of devices in

this dataset and performance. However, the fact that the second most successful dataset is also the largest strongly suggests that the number of devices is not the main factor.

Table 7: Results using RF for individual and aggregated packets for all datasets.

Dataset Name	Number of Individual		Aggregated
	Devices	F1-Score	F1-Score
Aalto University	29	0.894	0.992
UNSW-Sydney IEEE TMC	28	0.862	0.995
IoTFinder	51	0.892	0.998

6 Limitations

In our study, we tried to use all the publicly available IoT device classification data we could find (see Section 3.3). However, IoT technologies are a rapidly developing field and new data is likely to be released even as we continue our work. So, there may be datasets which are not included. We kept the variety of ML algorithms used in our study as wide as possible (see Section 4.1). However, there are many ML subfields that have become popular in recent years. In our study, although we did some preliminary research in the field of deep learning, it seems that traditional ML techniques work better in this particular problem.

7 Conclusions

In our study, we developed a fingerprinting method that summarizes the network behaviour of IoT devices. Using this fingerprinting method and ML algorithms together, we performed IoT device classification. While doing this, we utilized data augmentation techniques and a wider variety of evaluation criteria to deal with unbalanced distribution, which is a common problem in IoT data.

Additionally, we compared the fingerprint method with its counterparts and examined its performance with different learning algorithms. In this context, the IoTDevID feature set has been compared with its predecessors at the individual packet level and has been found to perform better by at least 5% in terms of accuracy. In addition, the IoTDevID method has been tested on three different datasets (Aalto University, UNSW-Sydney IEEE TMC, and IoTFinder) containing more than 100 devices in total and classified them at the individual packet level with an accuracy of 89.38%, 86.17%, and 89.22% respectively. Later, we increased these accuracy rates to 99.21%, 99.46% and 99.79%, respectively, using an algorithm that aggregates individual packets using MAC addresses.

References

1. Al-Garadi, M.A., Mohamed, A., Al-Ali, A., Du, X., Guizani, M.: A survey of machine and deep learning methods for internet of things (IoT) security. arXiv preprint arXiv:1807.11023 (2018)

2. Alrawi, O., Lever, C., Antonakakis, M., Monrose, F.: Sok: Security evaluation of home-based IoT deployments. In: Symp. on Security and Privacy. pp. 1362–1380. IEEE (2019)
3. Bezawada, B., Bachani, M., Peterson, J., Shirazi, H., Ray, I., Ray, I.: Behavioral fingerprinting of IoT devices. In: Proceedings of the 2018 Workshop on Attacks and Solutions in Hardware Security. pp. 41–50 (2018)
4. Bezawada, B., Bachani, M., Peterson, J., Shirazi, H., Ray, I., Ray, I.: IoTsense: Behavioral fingerprinting of iot devices. arXiv preprint arXiv:1804.03852 (2018)
5. Chaabouni, N., Mosbah, M., Zemmari, A., Sauvignac, C., Faruki, P.: Network intrusion detection for iot security based on learning techniques. *IEEE Communications Surveys & Tutorials* **21**(3), 2671–2701 (2019)
6. Esmalifalak, M., Liu, L., Nguyen, N., Zheng, R., Han, Z.: Detecting stealthy false data injection using machine learning in smart grid. *IEEE Systems Journal* **11**(3), 1644–1652 (2014)
7. Fernández, A., López, V., Galar, M., Del Jesus, M.J., Herrera, F.: Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-based systems* **42**, 97–110 (2013)
8. Hussain, F., Hussain, R., Hassan, S.A., Hossain, E.: Machine learning in IoT security: current solutions and future challenges. preprint arXiv:1904.05735 (2019)
9. Hwang, R.H., Peng, M.C., Nguyen, V.L., Chang, Y.L.: An lstm-based deep learning approach for classifying malicious traffic at the packet level. *Applied Sciences* **9**(16), 3414 (2019)
10. Insights, F.B.: Internet of things market size, growth | IoT industry report 2026 (2019), accessed: 2020-04-07
11. Kouicem, D.E., Bouabdallah, A., Lakhlef, H.: Internet of things security: A top-down survey. *Computer Networks* **141**, 199–221 (2018)
12. Lopez-Martin, M., Carro, B., Sanchez-Esguevillas, A., Lloret, J.: Network traffic classifier with convolutional and recurrent neural networks for internet of things. *IEEE Access* **5**, 18042–18050 (2017)
13. Meidan, Y., Bohadana, M., Shabtai, A., Guarnizo, J.D., Ochoa, M., Tippenhauer, N.O., Elovici, Y.: ProfillIoT: a machine learning approach for IoT device identification based on network traffic analysis. In: Proceedings of the Symp. on applied computing. pp. 506–509 (2017)
14. Miettinen, M., Marchal, S., Hafeez, I., Asokan, N., Sadeghi, A.R., Tarkoma, S.: IoT sentinel: Automated device-type identification for security enforcement in IoT. In: 37th Int. Conf. Distributed Computing Systems. pp. 2177–2184. IEEE (2017)
15. Nobakht, M., Sivaraman, V., Boreli, R.: A host-based intrusion detection and mitigation framework for smart home IoT using openflow. In: 2016 11th Int. Conf. on availability, reliability and security (ARES). pp. 147–156. IEEE (2016)
16. Ozay, M., Esnaola, I., Vural, F.T.Y., Kulkarni, S.R., Poor, H.V.: Machine learning methods for attack detection in the smart grid. *IEEE transactions on neural networks and learning systems* **27**(8), 1773–1786 (2015)
17. Sivanathan, A., Gharakheili, H.H., Loi, F., Radford, A., Wijenayake, C., Vishwanath, A., Sivaraman, V.: Classifying IoT devices in smart environments using network traffic characteristics. *IEEE-TMC* **18**(8), 1745–1759 (2018)
18. Zarpelão, B.B., Miani, R.S., Kawakani, C.T., de Alvarenga, S.C.: A survey of intrusion detection in internet of things. *Journal of Network and Computer Applications* **84**, 25–37 (2017)