IoTDevID: A Behavior-Based Device Identification Method for the IoT

Kahraman Kostas, Mike Just, and Michael A. Lones, Senior Member, IEEE

Abstract—Device identification (DI) is one way to secure a network of Internet of Things (IoT) devices, whereby devices identified as suspicious can subsequently be isolated from a network. In this study, we present a machine-learning-based method, IoTDevID, that recognizes devices through the characteristics of their network packets. As a result of using a rigorous feature analysis and selection process, our study offers a generalizable and realistic approach to modeling device behavior, achieving high predictive accuracy across two public data sets. The model’s underlying feature set is shown to be more predictive than existing feature sets used for DI and is shown to generalize to data unseen during the feature selection process. Unlike most existing approaches to IoT DI, IoTDevID is able to detect devices using non-IP and low-energy protocols.

Index Terms—Fingerprinting, Internet of Things (IoT) security, machine learning.

I. INTRODUCTION

The Internet of Things (IoT) is a collection of objects with embedded systems that can communicate with each other over a wired or wireless network. The “things” can be any physical items that we have now or will use in the future [1]. IoT contributes to human life in many critical areas, such as smart homes/cities, retail, healthcare, transportation, agriculture, military, and manufacturing [2]. By 2026, the number of IoT devices in the world is expected to reach 80 billion [3], with a total market of U.S. $1.1 trillion [4].

Securing IoT devices with traditional security solutions can be challenging due to limited device resources, such as processor, battery, and bandwidth [5]. Further challenges arise due to device heterogeneity [1]. For example, a device can have many sensors (temperature, humidity, motion, light, etc.), and the channels it uses to communicate with other devices may need very different requirements. Although some research focuses on specialist areas such as home appliances [6] or smart cities [7], many devices have very different characteristics even under this classification. For example, baby monitors and smart kettles are both classified under home appliances [6]. Further, IoT security has the potential to be neglected by users, primarily due to the lack of a familiar interface [6]. However, since today’s IoT devices are widespread and diverse, users cannot be expected to recognize and understand the security implications of all of them. Hence, automatic identification of devices in the network is vital in dealing with potential security problems. A successful device identification (DI) system can detect devices in a network, so necessary measures, such as updating, limiting, or isolating devices with security vulnerabilities, can be taken. In this way, potential security problems caused by IoT devices on the network can be prevented before they arise.

In this article, we introduce a new IoT DI method that models the behavior of the network packets communicated by the devices. With IoTDevID, devices are identified by their behavior, a critical first step to the subsequent detection of anomalous behavior. It classifies device behaviors at the individual packet level using generalizable features. In this way, it can detect all kinds of devices, while offering high detection success with its aggregation method. This work offers the following contributions and differences from previous work.

1) Our method does not strictly depend on identifying features, such as the MAC and IP address, in the feature extraction step. This gives it wider applicability than previous approaches that depended on this information, such as [8]–[10]. Notably, it allows the identification of devices that use non-IP and low-energy protocols, such as Bluetooth, ZigBee, or ZWave.

2) Our DI model uses a feature set that summarizes the network behavior of IoT devices. We carried out a systematic study of packet-level features, using various feature selection techniques to build a feature set that improves on previous approaches. Unlike previous studies [9], [11], [12], we were careful to identify and remove features (such as session IDs and port numbers) that do not generalize between data sets.

3) Our approach was developed using a rigorous experimental process. We took steps to prevent information leaking from the test set into training, and used a second, independent, data set to demonstrate the generality of our approach. Consequently, we believe that our results are more realistic and generalizable than much of the previously published work in this area.

As a consequence, our DI method is more suited to real-world IoT network environments than previous approaches, both because it supports a wider range of environments, but also because we have more confidence that the approach generalizes to environments that were unseen during development.
To promote transparency and repeatability, we have made our data set, scripts, and supplementary material public.\(^1\)

This article is organized as follows. Section II reviews related work. Section III describes the methods we use to model network packets, and the data sets used for evaluation. Section IV gives an overview of our model selection approach, Section V evaluates the selected model, and Section VI compares it against other DI approaches. Limitations are discussed in Section VII, and conclusions are presented in Section VIII.

### II. Related Work

This section reviews previous studies that used fingerprinting to classify IoT devices. Fingerprints are feature sets derived from network packets or statistics to reflect the behavior patterns of devices. There are many studies in the literature on DI using fingerprints, but their applicability to IoT devices is controversial, as these often focus on the physical layer or application layer where IoT has a wide protocol variety [10]. Hence, we focus here on research that is based on network packet behavior, including relevant information from each of the data link, network, and transport layers. Table I summarizes the previous studies in comparison to our study with the IoTDevID method (see Table VII for a comparison of results).

One of the first studies to use network packet features in a fingerprint method to identify IoT devices was IoT Sentinel [8]. This study used a behavior-based fingerprint to identify vulnerable devices and isolate them from the network, using data from Aalto University IoT device captures. The fingerprints specific to each device were created based on 23 features extracted from each of the first 12 packets for each device, resulting in a fingerprint with 276 values. These 12 packets do not exactly represent flow; they are sequential packets from the same MAC address. The features contain information about source and destination, packet properties, and protocols used. However, since it performs packet merging from MAC addresses, it cannot identify non-IP devices. (This situation, which we call the transfer problem, is explored in-depth in Section III-B.) Also, the IP address count feature used in this study is unlikely to generalize beyond the Aalto data set, since it refers to the number of devices each device communicates with, which is environment-dependent. 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%.

Two other notable studies also used data from the Aalto data set. In the first [9], a fingerprint consisting of 67 statistical features was created using information extracted from the headers of Ethernet, IP, UDP, and TCP packets of 20–21 consecutive packets. Unlike the others, this study used network statistics in addition to features obtained from the packets. If a device used in one network is moved to a different network, the features extracted from the network packet will not change, but the network statistics will change. Since the network statistics are based not only on devices but also on their interrelationships, it is unlikely that they will generalize well. This study classified the devices with 90.3% accuracy using various ML methods (see Table I). The second study [11] obtained 95% accuracy using 33 features selected by a genetic algorithm (GA) from 212 features extracted from the headers of individual packets. However, this study selectively presented its results with only part of the data set (25 out of 27 device classes) and it used features that are unlikely to generalize (e.g., IP ID, TCP acknowledgment, TCP sequence, and source and destination port numbers).

IoTSense [10] used selected features of IoT Sentinel based on their own design assessment, namely, 17 protocol-based features which reflect device behavior. They also added three payload-related features, notably payload length and entropy. This feature list was 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 merged according to a common MAC address, though this is not explicitly stated in this article. As a result of this study, per device recall of 93%–100% and an average accuracy of 99% were achieved. 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 versus 31). In addition, the IoTSense experiment set began with 14 devices, though only ten devices were used for the evaluation as four devices did not produce sufficient data for the analysis approach that was used.

Sivanathan et al. [12] 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 for the classification: 1) flow volume; 2) flow duration; 3) average flow rate; 4) device sleep time; 5) server port numbers; 6) domain name server (DNS) queries; 7) network time protocol (NTP) queries; and 8) cipher suites. As a result, 28 IoT devices were classified with an accuracy of 99.88%. However, four devices from the data set were not used, and some elements of the feature set were too specific, thereby not focusing on device behavior, e.g., port numbers, DNS queries, and cipher suites.

### III. Materials and Methods

#### A. System Model and Design Goals

IoTDevID analyzes the behavior of IoT devices in order to identify their brand and model. The resulting normal behavior

| Study       | Dataset | Devices | Algorithm(s)          | Key comparators                  |
|-------------|---------|---------|-----------------------|----------------------------------|
| [8]         | Aalto   | 27      | RF                    | Transfer problem                 |
| [9]         | Aalto   | 27      | RF, KK, GB, DT        | Transfer problem, overly specific features |
| [10]        | Private | 10      | GB, DT, NN            | Used partial dataset, overly specific features |
| [11]        | UNSW    | 28      | NB, RF                | Overly specific features         |
| Our Study   | UNSW    | 33      | NB, RF, SVM           | No trans. prob., overly specific features |

\(^1\)Materials available at: github.com/kahramankostas/IoTDevIDv2.
profiles can then be used to secure devices, e.g., by finding devices with known vulnerabilities. More details about the threat model can be found in Section VII. Our approach has similarities to the work reviewed in the previous section, in that we use ML methods to train device fingerprints from information stored in packet headers. However, we approach this in a more systematic way, using feature selection to identify fingerprints that generalize in a meaningful way from the training data, and with an emphasis on developing a model that can be deployed within real network environments.

B. IoT Data Selection

We used two public data sets to measure and compare the performance of our DI approach. Both of these contain real device data. Since our approach is designed for benign networks, the data sets do not contain attack data. First, the Aalto University IoT Devices Captures data set [8] (referred to as Aalto data set) has 31 devices and contains only the device installation data, but this installation process was repeated 20 times for each device to increase the amount of data [8]. While there are a total of 31 devices in this data set, four of these devices are in pairs (see the supplementary material SM-Table I). For example, there are two WeMoSwitches with different MAC/IP addresses. Our purpose is to detect according to device behavior, so these device pairs are considered as a single device, not as two separate devices. Therefore, although the data set contains 31 devices, it contains 27 classes. Another characteristic of the data set is that two devices using low-energy protocols connect to the gateway where the data is collected via other devices. Therefore, these two devices (D-LinkDoorSensor and HueSwitch) do not have identifying features such as their own MAC and IP address. They use the IP/MAC address of the devices they communicate through (D-LinkHomeHub and HueBridge, respectively). We refer to this as “the transfer problem” in this article (related device MAC addresses and labels are shared in SM-Table I). Notably, when using an individual packet-based approach (which is independent of MAC/IP addresses), our method is able to identify devices that are suffering from the transfer problem.

Second, UNSW IoT Traffic Traces [12] (referred to as UNSW data set) contains the day-to-day network logs of 28 IoT devices, each recorded over a period of 26 weeks. However, only a 60-day portion of this data is publicly available. This limited data does not include the four devices (Ring Door Bell, Hello Barbie, August Doorbell Camera, and Belkin Camera) from the work of Sivanathan et al. [12]. Therefore, we extracted data for these devices from the benign data of another study [13] by the same institution. The data set we use contains a total of 32 IoT devices and 7 non-IoT devices. We gathered non-IoT devices under one class, as the other study [12] working with this data set did, resulting in 33 total labels. Since the size of the UNSW data set is quite large, we used the smaller Aalto data set to formulate our modeling approach and the UNSW data set to measure its broader generality.

C. Individual, Aggregated, and Mixed Method Packets

Creating fingerprints by combining multiple packets is a common strategy [8]–[10], [12], where identifying features such as a MAC or IP address are used to combine multiple packets. These features uniquely identify the device, allowing us to know the source of the data, but do not provide any information about the device type or behavior. Assuming that packets from the same address come from the same device, larger fingerprints are often constructed by combining these packets. However, this assumption is not always correct. In cases where there is a transfer problem (e.g., multiple devices connecting through the same gateway), an IP/MAC address can represent more than one device. Consequently, the packet merging logic will not work, as devices with the same MAC address but different behavior will be considered as a single device. Further, there is no standard for the size of this merging process (number of packets, duration, etc.). Without merging, the use of individual packets is beneficial in terms of speed and efficient use of limited resources [14]. For these reasons, unlike many previous studies [8]–[10], [12], we chose to use individual packets, not merged ones as the basis for discrimination.

However, when viewed individually, some individual packets may be ambiguous and match the behavioral profile of more than one device. Therefore, the success rate from discrimination based on single packets is limited. We address this by introducing an alternative to individual packets, called aggregated method that groups packets based on identifier attribute (MAC address) and ML labels assigned at the individual packet level. In this way, we can increase the success rate by ignoring the mislabeled packets. The difference of this approach from previous work which merged packets is that the behavior analysis in our aggregation is performed before the merging process. In other words, we combine not the packets, but the labels assigned to these packets by the ML algorithm. Since this method does not work for MAC addresses suffering from the transfer problem, we have added an exception to our evaluation phase and followed a mixed method to deal with this issue. This reevaluation process works in parallel with the pseudocode (see SM-Algorithm I for pseudocode and corresponding line numbers mentioned below) as follows.

Our aggregation algorithm takes the device MAC addresses \( M = \{m_1, m_2, \ldots, m_n\} \), the result of the ML algorithm (initial labels) \( \hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n\} \), and group size \( g \) as input. The labels originating from the same MAC address are collected in the same array (lines 4–12). Each array is divided into small pieces of group number size (lines 13–16). Aggregated results are created by assigning the mode (the most frequent element) to the entire group as a new label \( \hat{Y}' = \{\hat{y}_1', \hat{y}_2', \ldots, \hat{y}_n'\} \) (lines 17–23). On the other hand, for creating the exception list of aggregation process, the MAC address corresponding to each label is added to the same array (lines 33–36). If a MAC address is the most frequent element in more than one array, it is added to the exception list (lines 37–41). As mixed result, individual labels \( \hat{y}_i \) are applied for the MAC addresses in the exception list, and aggregated labels \( \hat{y}_i' \) are applied for the remainder (lines 42–47).

However, while the aggregation algorithm works well for DI with benign data, care should be taken in networks with malicious data. Depending on the tolerance of the aggregation algorithm, malicious packets that impersonate an IP/MAC address and are close to the behavior of the benign packets
with the same IP/MAC address may be grouped together (see Section VII).

**D. Feature Extraction**

We performed feature extraction from the packet headers of pcap (packet capture) files. Before this, we separated the pcap files into test and training sets to completely isolate these sets. The Aalto data set has 20 sessions (pcap files) per device. We divided them into two parts: 16 parts training and four parts test data (80%:20%). The UNSW data set consists of a different pcap file for each day’s records. To isolate, we constructed the training and test data from data collected on different days (pcap files) (see SM-Table II for distribution). Section IV-A explains our feature selection process in detail.

In addition to the 111 features extracted from the network packet headers, we included payload entropy [10], protocol (from TCP-IP layers), and source and destination port class [8], which are all features found to be useful in previous studies. While payload entropy gives clues about the characteristic of the payload, the protocol and port class features provide summary information about source and destination ports. Source–destination port class features gather port numbers under 14 classes: No port, 0-Reserved, 53-DNS, 67-BOOTP server, 68-BOOTP client, 80-HTTP, 123-NTP, 443-HTTPS, 1900-SSDP, 5353-mDNS, 49153-ANTLR, 0:1023-well-known ports, 1023:49151-registered ports, and 49151:65535-dynamic (private) ports [8]. The port classification above is ordered so that, for example, a packet with port 53 is not also classified as a well-known port. Section IV-A explains further how port number classes are created.

MAC/IP addresses are source–destination-based identifying features. Although they uniquely identify the source and destination devices, they do not provide information about device behaviors. For example, two devices with the same behavior can have different MAC and IP addresses. Because they are not predictable and generalizable, we do not use them in our feature set (though they are used in our aggregation method).

**IV. MODEL SELECTION**

**A. Feature Selection**

With our initial features extracted, we used a feature-importance-based voting method via the xverse (pypi.org/project/xverse/) package to eliminate unnecessary features. This method calculates the importance scores of all features for each device using six different techniques and then uses voting to decide whether to include them (see Fig. 1 for features and their vote rates). The six scoring techniques used by this method are information value using the weight of evidence, variable importance using RF, recursive feature elimination, variable importance using extra trees classifier, chi-square best variables, and L1-based feature selection. We removed from our feature pool the 26 features that failed to receive votes on any device from any of these six techniques.

A number of the remaining features contain information that is potentially specific to a particular session, and therefore may not be useful for identifying devices more generally. For some features, this is clear, e.g., the initial values of the IP ID, TCP sequence, and acknowledgment numbers are randomly assigned, and the next values are consecutive numbers following this initial value; hence, they do not contain information that is useful for DI beyond their current session. For other features, such as TCP port numbers, their generality is less clear. So, to measure whether or not each of these features is beneficial, we excluded all of them from the feature set and trained an ML model (DT) to determine a baseline level of performance. In turn, we then added each of the features to the baseline set and observed whether this improved or impaired the performance.

The results are shown in Fig. 2. First, it is notable that the results are strongly dependent on whether the models are evaluated using cross-validation or using isolated training and test sets. This is presumably because, when using cross-validation, it is more likely that the training and test data will be sampled from the same session—meaning that session-specific patterns learned in the training set will also generalize to the test set. This may explain why session-specific features were chosen during feature selection and highlight the danger of using off-the-shelf feature selection algorithms on data of this kind. Using isolated training and test sets, by comparison, it is clear that session-specific features such as the IP ID cause overfitting and impair the generality of models.

On the other hand, Fig. 2 shows that port-based features can be useful for building generalizable models. For example,
consider the port numbers used by one of the devices in the Aalto data set which contains multiple device instances. Fig. 3 shows a word cloud built from the data from the two Edimax camera instances. Although a few port numbers (representing the protocols BOOTP, HTTP, SSDP, and mDNS) generalize between the devices, the remaining ports are presumably session-based and do not generalize. To address this, we do not use the raw feature values for port-based features, but rather map them to discrete values representing port classes and protocols (see Section III-D)—in effect, pruning the instance-specific information while preserving the behavioral information they provide.

Finally, after eliminating identifying and redundant features, we used a GA to decide on the most appropriate feature set from the remaining 52-member feature pool. The GA uses a wrapper method and thus tests the usefulness of feature sets in the context of a particular classifier (DT). In this way, we have created a feature subset that detects with higher performance, while decreasing the model complexity by reducing the number of features. Table II shows (highlighted in green) the final list of features selected by the GA.

### B. Algorithm Selection

Based on earlier approaches (see Section II) and surveys of related work [1], [15] which sought to train an ML algorithm to correctly predict a device’s type from extracted features, we considered the following six ML algorithms: RF, kNN, GB, DT, NB, and SVM. We used a random search with nested cross-validation (as implemented by scikit-learn: scikit-learn.org/stable/) to find suitable hyperparameters for each algorithm. Using these algorithms, we trained multiclass classifiers to discriminate the 27 classes from the Aalto data set, training each model 100 times to measure its stability. See Table III for the results.

It can be seen that RF, DT, and GB are the best performing algorithms, in terms of accuracy, they are not practical to use due to their slowness. The SVM algorithm is also not a reasonable option in terms of speed and accuracy. kNN, GB, and SVM are particularly disadvantageous due to their very large inference times. In a real-time device detection system that will operate in a millisecond-level processing environment such as network traffic, inference time in seconds are not acceptable. Based on these observations, we use DT in the remainder of this work, since it offers the best balance between speed and accuracy.

### V. Performance Evaluation

Based on our model selection, we present our results here for the three variations of our method: 1) individual; 2) aggregated; and 3) mixed. First, we looked at the relationship between group size and performance within the context of the aggregation algorithm. From Fig. 4, there is a positive relationship, showing that larger group sizes are generally more effective. However, many of the IoT devices communicate infrequently, so large group sizes would be impractical in some cases. For this reason, we selected a group size of 13, which is approximately the point where performance starts to plateau. Table IV shows the overall results for the two data sets. Using the individual packet approach, an F1 score of 73% was achieved in the Aalto data set and 83% in the UNSW data set. However, a significant increase in the ability to correctly
Fig. 4. Relationship between aggregation group size and performance. The benefit of increasing group size appears to plateau around $g = 13$.

### TABLE IV

| Method     | Dataset  | Accuracy | FI score | Test-t | Alg-t |
|------------|----------|----------|----------|--------|-------|
| Individual | Aalto    | 0.705±0.001 | 0.727±0.001 | 0.004  | 0.000 |
|            | UNSW     | 0.853±0.010 | 0.834±0.012 | 0.008  | 0.000 |
| Aggregated | Aalto    | 0.745±0.011 | 0.809±0.005 | 0.007  | 0.164 |
|            | UNSW     | 0.943±0.012 | 0.937±0.017 | 0.017  | 0.425 |
| Mixed      | Aalto    | 0.833±0.002 | 0.861±0.004 | 0.008  | 0.216 |
|            | UNSW     | 0.941±0.012 | 0.935±0.017 | 0.022  | 0.479 |

identify devices of both data sets was observed by using aggregation. The overall FI score increased to 81% on the Aalto data set and to approximately 94% on the UNSW data set. Since feature selection was done using only the Aalto data set, it is very encouraging to see such a high level of discrimination on the UNSW data set. This is a good indication that the selected feature set will generalize well to other IoT environments.

Table V shows the average discrimination performance of DT models at the device level for the Aalto data set. As Table V also shows, the data set is highly imbalanced, meaning that we need to use a metric that is relatively insensitive to class size imbalance in order to give a meaningful picture of the model’s performance at the device level—hence our use of the FI score. Other works have used accuracy for this, which is an inappropriate metric for imbalanced data sets.

When the device-based results of the Aalto data set are examined, we see that the aggregation algorithm contributes positively to almost all devices, while negatively affecting only four devices. This is because the pairs that make up these four devices suffer from the transfer problem (they share the same MAC address, see SM-Table I). To deal with this, we used our mixed method by adding an exception to our aggregation algorithm. Table IV shows that with the application of the mixed approach, the overall FI score, which was 81% in the Aalto data set, increased to 86%. Since there is no transfer problem in the UNSW data set, this approach does not impact its results significantly (see Table VI).

It is notable that the performance on the Aalto data set is significantly lower than the UNSW data set. This appears to be due to certain device subgroups. A confusion matrix containing only low-performance devices is given in Fig. 5. In this case, the devices in a subgroup have some similarities: they are either similar purpose devices manufactured by the same companies (e.g., yellow, blue, and orange groups in Fig. 5) or different models of the same device (e.g., red and green groups in Fig. 5). It does not seem possible to perfectly separate these devices according to their behavior, at least when observed at the network level. However, it is likely that these devices use very similar hardware and software, and so exhibit similar behavior, as well as similarities in vulnerabilities and their prevention [8]. Consequently, from a DI perspective, it is plausible to consider these devices under a single label. When doing this, the accuracy for the Aalto data set increases from 73% to 88% for the individual method, and from 86% to 97% for the mixed method.

### VI. COMPARISON WITH PREVIOUS WORK

In Table VII, we compare the overall results of our study with previously published results. While, at first glance, our study does not appear to yield higher numeric results than existing work, there are a number of reasons why this is not a fair comparison. First, many approaches use overly specific features that are unlikely to generalize to unseen data sets. As our analysis in Section IV-A showed, this may be a result of information leaking from the test set into the training set, something that is easy to do unintentionally. Second, there is the widespread use of accuracy figures with imbalanced data sets; since IoT device data sets inherently suffer from unbalanced distributions, using accuracy as the sole evaluation
TABLE VI

DEVICE PROPORTIONS OVER THE ENTIRE DATA SET, F1 SCORES PER DEVICE ACCORDING TO DIFFERENT APPROACHES (UNSW DATA SET, AVERAGE 100 REPETITIONS)

| Device name          | Packet statistics |          |          |          |          |
|----------------------|-------------------|----------|----------|----------|----------|
|                      | Packets | Percent | Individual | Aggregated | Mixed   |
| Amazon Echo          | 10000   | 2.831   | 0.882     | 0.985     | 0.983    |
| AugustDoorbellCam    | 10000   | 2.831   | 0.701     | 0.903     | 0.954    |
| AwairAirQualityMon   | 10000   | 2.831   | 0.952     | 0.993     | 0.984    |
| Belkin Camera        | 10000   | 2.831   | 0.890     | 1.000     | 1.000    |
| BelkinWeMoSwitch     | 10000   | 2.831   | 0.803     | 0.988     | 0.988    |
| BelkinWeMoSensor     | 10000   | 2.831   | 0.772     | 0.987     | 0.988    |
| BicicareBPMeter      | 119     | 0.034   | 0.851     | 1.000     | 1.000    |
| Canary Camera        | 10000   | 2.831   | 0.761     | 0.958     | 0.911    |
| Dropcam              | 10000   | 2.831   | 0.998     | 1.000     | 1.000    |
| GoogleChromecast     | 10000   | 2.831   | 0.448     | 0.347     | 0.341    |
| HP Printer           | 10000   | 2.831   | 0.468     | 0.710     | 0.707    |
| HelloBarbie          | 164     | 0.046   | 0.677     | 1.000     | 1.000    |
| InsteonCam           | 10000   | 2.831   | 0.953     | 0.997     | 0.997    |
| LightiLFXSmartBulb   | 10000   | 2.831   | 0.979     | 1.000     | 1.000    |
| NESTPSSmokeAlarm     | 7063    | 1.999   | 0.935     | 1.000     | 1.000    |
| Nest Dropcam         | 10000   | 2.831   | 0.966     | 1.000     | 1.000    |
| NetatmoWelcome       | 10000   | 2.831   | 0.922     | 0.998     | 0.994    |
| NetatmoWeatherSt     | 10000   | 2.831   | 0.896     | 1.000     | 1.000    |
| Non-IoT              | 67295   | 19.05   | 0.872     | 0.929     | 0.929    |
| PDX-STAR-Pframe      | 10000   | 2.831   | 0.587     | 0.928     | 0.885    |
| PhilipsHlightbulb    | 10000   | 2.831   | 0.971     | 1.000     | 1.000    |
| RingDoorBell         | 10000   | 2.831   | 0.856     | 0.999     | 0.995    |
| Samsung&Cam          | 10000   | 2.831   | 0.678     | 0.892     | 0.896    |
| SmartThings          | 10000   | 2.831   | 0.953     | 1.000     | 1.000    |
| TP-LinkCloudCam      | 10000   | 2.831   | 0.979     | 1.000     | 1.000    |
| TP-Link Smart plug   | 10000   | 2.831   | 0.875     | 1.000     | 1.000    |
| TP-LinkRouter        | 10000   | 2.831   | 0.908     | 0.999     | 0.999    |
| TribySpeaker         | 10000   | 2.831   | 0.908     | 0.999     | 0.999    |
| WithingsSipSensor    | 10000   | 2.831   | 0.437     | 0.404     | 0.411    |
| WithingsBabyMon      | 10000   | 2.831   | 0.984     | 1.000     | 1.000    |
| WithingsSScale       | 8279    | 2.344   | 0.973     | 1.000     | 1.000    |
| iHome                | 10000   | 2.831   | 0.958     | 1.000     | 1.000    |
| unknownIoT           | 340     | 0.096   | 0.676     | 0.894     | 0.891    |

Fig. 5. Confusion matrix of devices with poor discrimination. The colored groups show this is mostly due to the difficulty of discriminating between similar devices from the same manufacturer.

criterion can be highly misleading. Third, there is inconsistency amongst the data sets used, with some studies reporting results for only partial data sets. We summarize these points in Section II and Table I. In our work, by comparison, we were careful to filter out data set-specific features, prevent information leakage, use appropriate metrics, and report full results. We also explicitly used a second data set to measure the generality of the approach. Consequently, we believe that our results are unbiased and more likely to indicate likely success on unseen data than those reported in previous studies.

A key contribution of our work is the determination of a feature set that provides good discrimination across data sets and is therefore likely to provide a good basis for DI more generally. To give more insight into this, and to offer a more meaningful comparison against previous approaches, we compared this feature set against the feature sets used in IoTSense [10] and IoT Sentinel [8]. Note we did not include the other methods listed in Table VII in this comparison, either because they were flow-based [9], [12] or their feature set was not shared [10]. We applied all three of our approaches (individual, aggregated, and mixed) using all three feature sets and both data sets. Table VIII shows that, in all cases, our feature set resulted in significantly better device discrimination than the previously published feature sets. It is also notable that the performance metrics we observed when using the feature set from IoTSense are significantly worse than that method’s published figure of 99% accuracy, which again highlights the problem of basing comparisons on published figures alone.
VII. LIMITATIONS

In our study, we used all the IoT DI data sets that were publicly available at the time (see Section III-B). However, IoT technologies are a rapidly developing field and new data is likely to have been released even as we continued our work. We encourage other researchers to use other data sets to further test the generalizability of our results and we provide scripts to make this possible. We used a broad selection of commonly used ML algorithms (see Section IV-B), though this was not exhaustive, and there may be other algorithms that might perform better. However, we have done some preliminary research using deep neural networks (not reported here), and so far our results suggest that more traditional ML techniques work better on this particular problem.

Although IoTDevID identifies devices well, it does not provide solutions for detecting vulnerable devices and taking necessary precautions about these devices. A software-defined networking (SDN)-based network management solution may be useful for this process. Furthermore, IoTDevID is currently intended only for use in benign networks, meaning that there is still a need for an intrusion detection system to protect against attacks. A further consideration is how a DI model would be deployed, trained, and maintained in a particular network. While we focus on the performance of multiclass classifiers, i.e., a single model for discriminating between all devices on the network, this may not be optimal in situations where devices are frequently added and removed. A more practical architecture in this case may be an ensemble of single-class classifiers, since this would not require retraining of the entire model. However, our initial results (not reported here) suggest there may be a tradeoff against performance, with multiclass models offering better discrimination.

VIII. CONCLUSION

In this study, we present an ML-based method for identifying IoT devices. Identifying the IoT devices with a network is important, both for finding and removing rogue devices, and for understanding the security vulnerabilities of a network more generally. A number of previous studies have attempted to use ML in this process. However, we have identified various factors that may limit the generality of their findings, including the use of session-based identifying features, and the use of inappropriate metrics. In this study, we have attempted to go about this process in a rigorous and transparent manner, with the aim of developing a robust method of DI that reliably generalizes beyond the data on which it was trained.

A key contribution of this work is the use of multistage feature selection to determine a set of generalizable packet-level features that can be used for building robust models. Notably, we compared this feature set against those used in previous studies and showed that it supports the training of significantly better ML models. We also demonstrate that the feature set generalizes well to a data set that was unseen during feature selection, giving us confidence that it provides a meaningful basis for the broader discrimination of IoT devices.

We also address the problem of detecting non-IP devices. A limitation of earlier approaches is that they use IP or MAC addresses to merge packets prior to applying an ML model, which means they are not applicable in situations where a device does not have an IP or MAC address. Instead, we apply ML at the packet level, and then use an aggregation algorithm that takes into account both the packet-level classification and, if available, the IP or MAC address. If the latter is not available, it is still possible to carry out identification based upon individual packets alone.

From an ML perspective, we found decision trees to offer the best tradeoff between predictive performance and inference time, the latter being an important consideration for deploying a model to monitor network traffic in real time.

In future work, we plan to develop an IDS that will work with IoTDevID and detect attacks against the network, and an SDN-based network management system that evaluates both DI and intrusion detection outputs. Thus, the aim is to have an IoT security system that avoids the risk of potential vulnerabilities, prevents attacks, and can be used in a real-world setting.

REFERENCES

[1] F. Hussain, R. Hussain, S. A. Hassan, and E. Hossain, “Machine learning in IoT security: Current solutions and future challenges,” IEEE Commun. Surveys Tuts., vol. 22, no. 3, pp. 1686–1721, 3rd Quart., 2020.
[2] D. E. Kouicem, A. Bouabdallah, and H. Lakhlef, “Internet of Things security: A top-down survey,” Comput. Netw., vol. 141, pp. 199–221, Aug. 2018.
[3] M. Rosen, “Driving the Digital Agenda Requires Strategic Architecture.” [Online]. Available: https://idc-cema.com/dwn/SF_177701 (Accessed: Apr. 4, 2022).
[4] “IoT Market Size, Growth IoT Industry Report 2026.” [Online]. Available: https://www.fortunebusinessinsights.com/industry-reports/iot-market-100307 (Accessed: Apr. 4, 2022).
[5] B. B. Zarpelao, R. S. Miani, C. T. Kawakani, and S. C. de Alvarenga, “A survey of intrusion detection in Internet of Things,” J. Netw. Comput. Appl., vol. 84, pp. 25–37, Apr. 2017.
[6] O. Alrawi, C. Lever, M. Antonakakis, and F. Monrose, “SoK: Security evaluation of home-based IoT deployments,” in Proc. Sympos. Security Privacy, 2019, pp. 1362–1380.
[7] M. Ozay, I. Esnaola, F. T. Y. Vural, S. R. Kulkarni, and H. V. Poor, “Machine learning methods for attack detection in the smart grid,” IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 8, pp. 1773–1786, Aug. 2016.
[8] M. Miettinen, S. Marchal, I. Hafeez, N. Asokan, A.-R. Sadeghi, and S. Tarkoma, “IoT SENTINEL: Automated device-type identification for security enforcement in IoT,” in Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst., 2017, pp. 2177–2184.
[9] S. A. Hamad, W. E. Zhang, Q. Z. Sheng, and S. Nepal, “IoT device identification via network-flow based fingerprinting and learning,” in Proc. 18th IEEE TrustCom, 2019, pp. 103–111.
[10] B. Bezawada, M. Bachani, J. Peterson, H. Shirazi, I. Ray, and I. Ray, “Behavioral fingerprinting of IoT devices,” in Proc. Workshop Attacks Solutions Hardw. Sec., 2018, pp. 41–50.
[11] A. Aksoy and M. H. Gunes, “Automated IoT device identification using network traffic,” in Proc. IEEE ICC, 2019, pp. 1–7.
[12] A. Sivathanan et al., “Classifying IoT devices in smart environments using network traffic characteristics,” IEEE Trans. Mobile Comput., vol. 18, no. 8, pp. 1745–1759, Aug. 2019.
[13] A. Hamza, H. H. Gharakheili, T. A. Benson, and V. Sivaraman, “Detecting volumetric attacks on IoT devices via SDN-based monitoring of MUD activity,” in Proc. ACM Symp. SDN Res., 2019, pp. 36–48.
[14] R.-H. Hwang, M.-C. Peng, V.-L. Nguyen, and Y.-L. Chang, “An LSTM-based deep learning approach for classifying malicious traffic at the packet level,” Appl. Sci., vol. 9, no. 16, p. 3414, 2019.
[15] M. A. Al-Garadi, A. Mohamed, A. Al-Ali, X. Du, I. Ali, and M. Guizani, “A survey of machine and deep learning methods for Internet of Things (IoT) security,” IEEE Commun. Surveys Tuts., vol. 22, no. 3, pp. 1646–1685, 3rd Quart., 2020.
Kahraman Kostas received the M.Sc. degree in computer networks and security from the University of Essex, Colchester, U.K., in 2018. He is currently pursuing the Ph.D. degree in computer science with Heriot-Watt University, Edinburgh, U.K.

His research focuses on the security of computer networks and Internet of Things. You can find more information at https://kahramankostas.github.io/.

Mike Just received the Ph.D. degree in computer science from Carleton University, Ottawa, ON, Canada, in 1998.

He is currently an Associate Professor with Heriot-Watt University, Edinburgh, U.K. He is primarily interested in computer security, and in applying human–computer interaction and machine learning techniques to solve computer security problems. You can find more information at https://justmikejust.wordpress.com/.

Michael A. Lones (Senior Member, IEEE) received the M.Eng. and Ph.D. degrees from the University of York, York, U.K., in 1999 and 2003, respectively.

He is an Associate Professor of Computer Science with Heriot-Watt University, Edinburgh, U.K. He carries out research in the areas of machine learning and optimization, where he has a particular interest in biologically inspired approaches. Application areas of his work include medicine, robotics, and security. You can find more information at http://www.macs.hw.ac.uk/~ml355.