2021 IEEE International Systems Conference (SysCon) | 978-1-6654-4439-2/20/$31.00 ©2021 IEEE | DOI: 10.1109/SysCon48628.2021.9447099

Network Traffic Flow Based Machine Learning Technique for IoT Device Identification

Imtiaz Ullah, Qusay H. Mahmoud
Department of Electrical, Computer and Software Engineering
Ontario Tech University,
Oshawa, ON, L1G 0C5 Canada
{imtiaz.ullah, qusay.mahmoud}@ontariotechu.net

***Abstract*—Security and privacy issues are being raised as smart systems are integrated into our daily lives. New security issues have emerged with several new vendors that develop the Internet of Things (IoT) products. The contents and patterns of network traffic will expose vulnerable IoT devices to intruders. New methods of network assessment are needed to evaluate the type of network connected IoT devices. IoT device recognition would provide a comprehensive structure for the development of stable IoT networks. This paper chooses a machine learning technique to identify IoT devices linked to the network by analyzing network flow sent and received. To generate network traffic data, we have developed a dataset adapted from the IoT23 Pcap files to experiment with a smart home network. We have created a model to identify the IoT device based on network traffic analysis. We evaluate our proposed model via full features dataset, reduces features dataset, and flow-based features dataset. This paper focuses on using flow-based features to identify the IoT device connected to the network. Our proposed scheme results in 100% precision, precision, recall, and F score via a full features dataset, reduced features dataset, and flow-based features dataset. Through evaluations using our produced dataset, we demonstrate that the proposed model can accurately classify IoT devices.**

Keywords*—*Internet of Things, device identification, sensor profile, flow-based detection system, anomaly detection system.

I. Introduction

The growing popularity of Internet-connected advanced gadgets and equipment, known as the Internet of Things (IoT), offers new conveniences. The rapidly rising selection of smart user IoT products and usability promises to revolutionize how we connect to our living areas. The IoT provides many benefits in almost all facets of our lives. IoT becomes an essential part of our daily life, and it is expected that the number of IoT devices will be increased to 75 billion by the end of 2025 [1]. IoT provides new technology services like smart homes, better resource utilization like smart grid, enhanced industrial production, new building equipment for control, and assistance. IoT turns into existence as more and more people installing Internet-connected appliances and devices [2],[3]. With the adoption of these applications, security is becoming an emerging challenge. New companies with limited experience in security transported several new IoT devices with inherent security weaknesses. Many IoT devices also make it difficult for the average consumer to update their software. These factors make IoT devices more vulnerable to security attacks. Attackers have recently used specialized techniques to conduct massive DDoS attacks that exploit hundreds of thousands of infected IoT devices [4]. The intruder needed a limited amount of prerequisite technical knowledge about IoT devices or networks. The rapidly expanding and diversity of IoT devices also brings many exploits and security vulnerabilities. These IoT devices are plagued with many vulnerabilities, as many manufacturers sell smart devices without considering security [5].

The network traffic patterns and contents can expose user-sensitive information to intruders. A reliable and efficient communication interface between embedded devices and the Internet is important for an IoT-based platform. In smart networks, these findings in IoT systems raise significant privacy concerns. It is difficult to manage user security in the presence of exposed devices. In nearly every part of our lives, the IoT is spreading worldwide, offering different benefits. Unfortunately, IoT devices often bring many flaws and risks to computer security. In addition to its conventional deficiencies, the dangers and future global consequences of connecting IoT devices to the network are evident in all existing situations when IoT networks' inherent technological vulnerabilities are taken into account, the ease of hacker operations and their expected spillover around the world. The present analysis focuses on the complexities of IoT technologies for major corporations. IoT protection in businesses is linked to the conduct of the company, as well as its employees. A variety of corporate technologies can serve IoT systems that are self-deployed. Smart cameras boost security, smart sockets, smart light bulbs, and smart thermostats make it easy to save power, and so on. However, precaution needs to be taken to avoid expanding an organization's attack zone by these Internet-enabling gadgets.

The network traffic content, patterns, and metadata all expose confidential information about a user's online operation. Online communication has been generally limited to Internet browsing in the past. The always-power ON sensors in smart homes transmit information about a person's offline behaviors across the Internet. In several ways, this comprehensive data could be useful, like advertisement and market intelligence. However, privacy supporters contend that even though traffic content is inaccessible, metadata and traffic patterns will expose confidential details. Further research describing encrypted

978-1-6654-4439-2/21/$31.00 ©2021 IEEE

traffic privacy vulnerabilities and IoT technology is becoming more popular, but in this situation, metadata from IoT devices can help create future regulations. As devices are plugged-in to the IoT network or removed from an IoT network, it is essential to identify these devices' types to establish a cognitive benchmark. Due to many protocols, applications, and control interfaces around the IoT devices, identifying these devices is difficult. An IoT device can reply to their identification inquiries, which is a typical way to learn about the system remotely. But, by presenting false information about its name and type, an untrusted device will masquerade as another device. The IoT device profiling process is at an early stage due to its evolving presence in the IoT industry. IoT systems use various protocols at different points of their service, such as the ARP, SSL, LLC, EAPOL, HTTP, MDNS, and DNS subsets. The set of applications that the IoT system uses constitutes a strong predictor of system behavior. The application protocols, however, only provide a static view of the system's operations. To better understand the dynamic nature of the system's behavior, more study is needed.

The rest of the paper proceeds as follows: Next, we discuss the related work in Section II, with the Internet of Things (IoT) discussed in Section III. In Section IV, the proposed technique is discussed. The analysis of the results is presented in Section V, with discussion in Section VI. Finally, Section VII concludes the paper and offers ideas for future work.

II. Related Work

The growing need for IoT devices brings new challenges to the monitoring of the network's operations. To discover which gadgets are linked to the network, new network classification technologies are required. Statical assessment methods may be utilized to identify unique trends that can classify IoT devices. A consistent framework for device type recognition is needed to execute an IoT operating policy systematically. Professional adversaries can create an IoT device's MAC address, so MAC address is insufficient to identify IoT devices. In addition, while MAC addresses can be used to identify the provider of a user's device, there is no accepted standard for defining a system name based on the MAC address [6].

A method to identify compromised IoT devices is proposed by Nguyen et al. [7]. Their model uses the temporal frequency of the stream of traffic produced by IoT devices. First, for individual devices, normal profiles are produced then they used a recurrent neural network to find variance from the normal actions predicted to identify compromised IoT devices. Authentication-based approaches have also been studied as a form of IoT system classification and identity management. This methodology was evaluated in [8], in which IoT whitelisting classification for Industrial Automation Control Systems was introduced. However, as described by the authors, these devices are usually built in such a way to understand communication interactions so that the whole objective of an organization becomes tractable. On the other side, the large-scale system is much more complicated, where new IoT product types or brands are frequently incorporated. Techniques focused on authentication are also susceptible to scale-based battles. Also, it is not feasible to enable all vendors to follow standard encryption requirements. It is also not possible to create a standard for global public-key networks.

A variety of fingerprinting methods are used for legal purposes, such as forensics and intrusion prevention, and harmful applications, such as attack identification and user profiling. Franklin et al. [9] define a passive fingerprinting approach to classify various implementations of 802.11 wireless system drivers on clients. In a specific application implementation, they analyze the successful channel scanning techniques via statistical relationship. This method effectively determines the type of device driver, but this technique cannot be used to identify the type of an IoT device. For example, over many device types, a vendor may reuse the same device driver implementation. GTID was defined in [10] by Radhakrishnan et al. for device type recognition on common-purpose gadgets such as smartphones, tablet PCs, and laptops. They used packet inter-arrival times of the different packets to retrieve the necessary attributes unique to a specific application. However, in terms of traffic production, IoT devices are typically quite conservative and do not produce much traffic. It would take non-trivial changes to the original series of algorithms to extend these methods to IoT networks.

IoT Sentinel, a system for device fingerprinting and securing IoT networks, was defined in [11] by Miettinen et al. When an IoT device initially registers to the network, their framework identity the IoT device using machine learning techniques. Their work does not, however, examine a device's behavior. Siby et al. [12] define IoTScanner, a design that examines network traffic flow at the data link layer. They used a frame header to analyze the network traffic through a particular examination time frame. The drawback of their strategy is that similar devices will be identified as different devices if the traffic capture windows are different during system traffic. Kawai et al. [13] use traffic characteristics and ML algorithms to classify IoT devices. Their methodology is entirely different from other research because they used only two types of traffic attributes, i.e., packet size and IAT. Some methods provide other kinds of traffic characteristics to improve the precision of identification. An approach focused on a neuro-fuzzy framework has been introduced by Rizzi et al. [14] to classify early-stage device traffic. A composite attribute was introduced by Dainotti et al. [15] to provide early phase identification of flow with higher precision. Their experiments indicate that an effective composition relationship is significant where there are early recognition limitations, i.e., when necessary to obtain stream identification for interactive network resources at an early level.

III. Internet of Things

The Internet of Things combines objects from multiple domains and the Internet. IoT offers a wide variety of applications that integrate intelligent objects and the Internet. Smart objects incorporate a system's physical assets to make effective use of resources. By 2025, projections predict that more than 75 billion linked devices utilizing the Internet of Things (IoT) would be in operation, which will be a roughly three-fold rise of IoT devices installed in 2019, as shown in Figure 1[1].

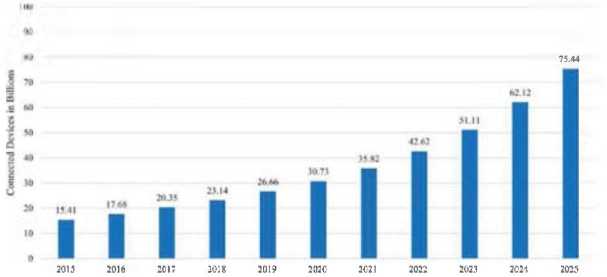

Fig.1. Globally deployed Internet of Things (IoT) related gadgets from 2015 to 2025

IoT provides new technology-based services like smart homes, better resource utilization like smart grid, enhanced industrial production, new building equipment for control, and assistance. Smart infrastructure intelligently responds to a change in its environment to improve performance, including user demands and other infrastructure [16]. Figure 2 shows a generic architecture of the Internet of Things.

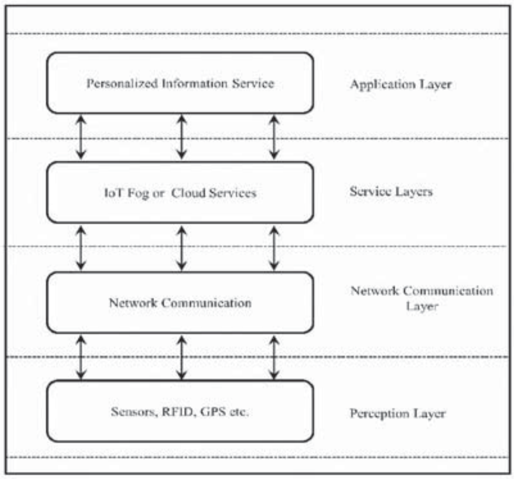

Fig.2. IoT Network Security Architecture

IoT sensors embedded in a smart infrastructure connected to the Internet allow real-time data acquisition and data analysis. A smart infrastructure provides suggestions for making better decisions in different domains such as public safety, public transport, energy, etc. A smart infrastructure required consistency, robustness, and flexibility, e.g., transportation at a reasonable cost, inexpensive and reliable electricity, high-quality water, etc. IoT networks empowered the end-user to make a real-time decision to provide an analytical service for healthcare, water management, energy distribution, and transport management. Managing these resources and making them accessible at the correct place and at the correct time is the major challenge in developing smart infrastructure.

An essential characteristic of smart infrastructure is to provide a service consistently and robustly. An Individual IoT device in a smart infrastructure has small states, but when these separate IoT devices are combined, they become an enormous number of interconnecting states. IoT networks required sensing and actuating devices; therefore, the cost becomes a significant constraint in designing and developing an IoT-based smart infrastructure. Other constraints of IoT devices include computational power, storage, and batteries. IoT connects more physical objects to the Internet to create a large attack surface for attackers; thus, a smart public infrastructure can be easily attacked by intruders. Therefore cybersecurity has become a public safety concern in recent years. As technologies for information management allow processes and services, IoT networks may offer several systems opportunities. The primary aim of IoT is to provide modern means for day-to-day interactions to satisfy social needs. IoT's main challenges are security, availability, cost, effectiveness, data management, scalability. The security of IoT networks has been a critical issue in the deployment of smart infrastructure.

In any network security strategy, device identification and access control are likely to be one of the most critical and challenging issues. The Internet of Things has posed potential vulnerabilities in device detection. However, the technical sophistication of authentication algorithms and key management-related scalability concerns make almost all authentication protocols based on cryptography ineffective for IoT networks. Detection of sensors is a complex task, but some vendors are aware of the unsolved problem. As devices are plugged in or removed from the IoT network, it is essential to identify these devices to establish a behavioral model. Data mining and machine learning techniques play an imperative role in developing and enhancing IoT device identification. Predictions are the main objective of data mining and machine learning. Classification in supervised data mining allows an identifications model to produce a given result from the input. The purpose of classification is to create a model from labeled entities that can be categorized explicitly as required. The characteristics of network traffic approximate the actions of IoT applications. These features are used to train an IoT-specific machine learning algorithm model and can be used to identify similar IoT devices.

IV. Architecture

The research into IoT device type identification is at the initial stage due to the IoT industry's evolving complexity. The growing requirement for IoT devices generates several problems for the network to sustain network operations. This section presents a technique to design and develop a framework to identify IoT devices in IoT networks when a new IoT device has been added, or an IoT device has been compromised or provides false information. There is a need for new network exploration techniques to identify the IoT devices connected to the network. In this sense, it is possible to use data analysis methods to identify distinctive configurations that can distinguish device types. IoT devices perform particular tasks, unlike desktop computers, making them merely more predictable. Network traffic analysis is proposed to detect IoT devices with high precision and low false alarms. The proposed methodology will safeguard the IoT networks' operations against different attacks by checking and analyzing IoT devices' operations. Figure 3 shows our proposed framework for sensor profiles in the IoT network. The testbed comprises a mixture of IoT modules and interconnecting structures. The IoT devices include sensors to receive/send data from/to the physical world. The IoT device identification technique is divided into five steps, as shown in Figure 3.

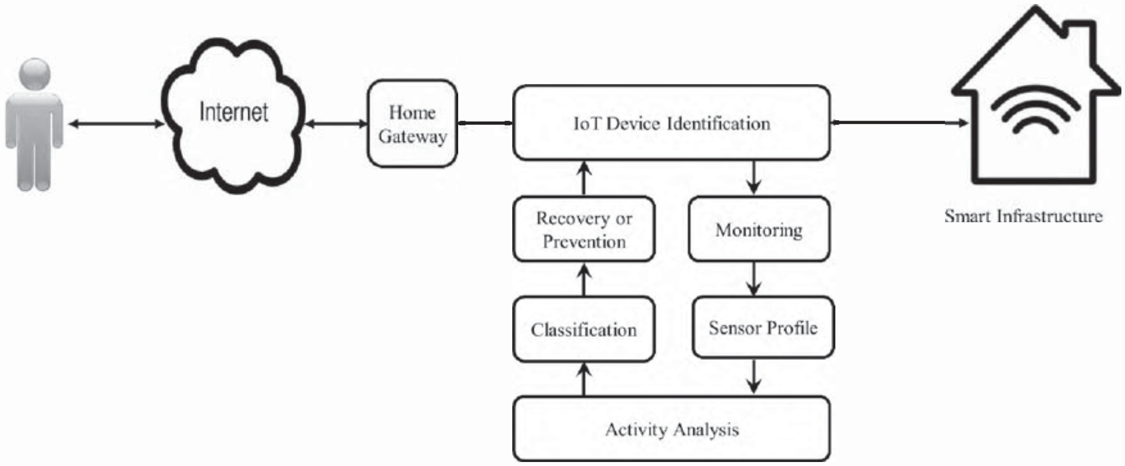

Fig.3. IoT Device Identification Framework for Smart Infrastructure

A. Monitoring

IoT network traffic is collected by a network management tool. The monitoring process will take place between the gateway and the smart infrastructure. The benefit of using this technique is to detect malicious IoT devices' operations before the gateway is reached. Packet capture software such as Wireshark is used to capture network communication. Wireshark traffic includes the source IP, the source port, the target IP, the destination port, and packets' content. IoT device data will be collected from the payload to produce the profile of the device. The data from all devices will be analyzed to identify device behavior.

B. Sensor profile

The module for the sensor profile describes a data structure for the normal operation of IoT sensors. Machine learning methods have been used to describe the regular function of IoT sensors. A detailed model should be used to establish all potential states of normal sensor behavior. This article focuses on the sensors' traffic analysis to classify IoT devices across the network. The network administrator will detect infected sensors in the IoT networks through device identification.

C. Analysis

The previous step's sensor profile will be used as a benchmark to search for possible inconsistencies in the received IoT system's communications. A runtime profile is generated for the sensor communications, and any deviation from the base model should be considered an abnormality. The probability distribution is used to verify the likelihood of natural behavior beyond the lower or upper bound. It will be considered as an irregular system if the collected data rate exceeds the specification parameters.

D. Classification

The classification functionality determines the irregularity after the analysis module finds a malicious in the received communications. Classification of irregularities allows IT, managers or consumers, to recognize the type of abnormality more precisely.

E. Action (Prevention and Recovery)

Several recovery measures can be taken to protect the IoT networks, e.g., reject received information, adjust the network configuration, de-authenticate the sensor, etc. If the classification unit does not identify an IoT device, it is possible to reset the IoT device and ask the unit to re-authenticate itself. This paper concentrates on the sensor profile to identify IoT devices by using network traffic analysis. A hacker can use IoT device identification to discover infected IoT devices by performing proactive traffic analysis of the network. Sensor profile and device recognition can assist the network administrator in identifying compromised devices in the IoT networks. The IT manager can also use the sensor profile to enforce different security policies for different IoT devices.

V. Results and Analysis

Our research aims to provide a way for IoT devices to be defined based on their flow-based network behavior. We used the Pcap files of the [17] dataset to extract network features to create a new dataset. We used the CICflowmeter [18] to extract features from Pcap files and construct the CSV format of the IoT-23 dataset. IoT-AD-20 is the name of the new dataset [19]. The IoT-AD-20 dataset comprises 80 network features and a label feature. The label feature identifies the type of IoT device. The testbed for the IoT-23 [17] dataset comprises the IoT devices and the networks that interconnect these devices. A standard smart home structure was developed, consisting of the Amazon Echo device, the Philips Hue device, the Somfy door lock device to generate the dataset. Our proposed new database includes several network and flow features. One challenge is how much training set data should be obtained before it is compared with the IoT system's behavioral profile. The total number of instances collected for different devices is presented in Table 1. We analyzed and tested various instances sizes for each device to verify our proposed model's capacity to classify different IoT devices.

Table 1. IoT-AD-20 Dataset Instances

Device Name	Instances
Amazon Echo	29495
Philips Hue	24633
Somfy Door Lock	8286

A data preparation approach was needed for the IoT-AD-20 dataset because data types are not suited to machine learning techniques for such attributes. To standardize the IoT-AD-20 dataset, we used the column normalization technique. Our proposed model is tested for full dataset features, reduced features, and flow-based features. To evaluate the proposed model for device identification, we used numerous machine learning algorithms. This study used multiple classifiers, but optimum results were achieved by a classifier using a decision tree. Table 2 shows the accuracy, precision, recall, and F scores of the entire data set.

Table 2. Accuracy, Precision, Recall, and F Score

Device	Accuracy	Precision	Recall	F Score
Amazon Echo	100	100	100	100
Philips Hue	100	100	100	100
Somfy Door Lock	100	100	100	100

The learning curve illustrates a connection between the training and testing of an algorithm over several test samples. The learning curve addresses how the algorithm will improve the detection capacity by presenting further data, or sufficient data is provided for the optimal output of the algorithm. Figure 4 shows the F score learning curve for detecting IoT devices using a decision tree algorithm. Since the F score is the harmonic means of precision and recall, so we used the F score for the learning curve. The learning curve concludes that a minimum of three hundred instances is sufficient for IoT device identification to achieve better decision tree efficiency.

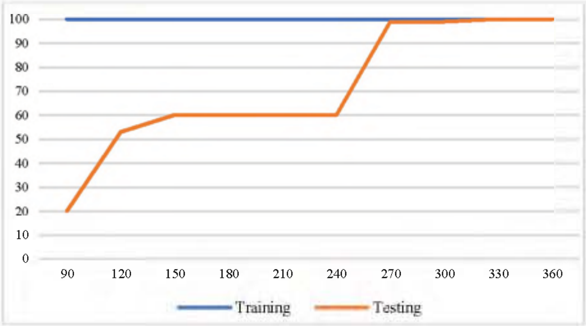

Fig.4. Learning Curve for Device Type Identification

Two important areas of machine learning are training and testing. Initially, we made a random selection to produce training data; thus, the suggested method for training and testing selects a randomized sample each time. A fully grown tree can overfit; thus, we divide the fully grown decision tree using our proposed model's maximum depth range. We used the K-fold cross-validation test to measure the feasibility and over-fitting of the proposed model. An overfitted model yields an optimal fit to the training data but may perform poorly on other data sets. Such a model is not of much value in the real world since it cannot forecast new events. We checked the validity of the model proposed with a 3, 5, and 10-fold cross-validation test. The effect of the 10-fold cross-validation score is shown in Table 3.

Table 3. 10-Fold Cross-Validation for Device Identification

Device	Accuracy	Precision	Recall	F Score
Amazon Echo	100	100	100	100
Philips Hue	100	100	100	100
Somfy Door Lock	100	100	100	100

Next, using a reduced data set, we evaluated our proposed framework. To choose important and high-rank features, we used a recursive feature selection model. Feature selection plays an important function in machine learning. Feature selection is a strategy for choosing relevant and required features to increase the prediction ability. The feature selection method eliminates overfitting, improves the model's predictive ability, decreases the machine learning algorithm's training and testing time, and decreases the model's difficulty level [20]-[26]. We implemented the RFE technique to select significant attributes after extracting all correlated features from the IoT-AD-20 dataset. A random forest algorithm was used as an estimator by RFE for feature ranking. The RFE methodology ranks feature based on their importance. We used a cross-validation test to verify the selection of the extracted features and the over-fitting of the RFE model. We have used accuracy, precision, recall, and F score as performance parameters to check the RFE model. The outcomes were compared with the results of the full dataset of functionality. Our study concluded that an optimal collection of features is known to be a selection of 15 to 40 features. Table 4 displays the outcomes of 15, 30, and 40 features.

Table 4. Accuracy, Precision, Recall, and F Score for Selected Features

Device	Accuracy	Precision	Recall	F Score
15 Features				
Amazon Echo	100	100	100	100
Philips Hue	100	100	100	100
Somfy Door Lock	100	100	100	100
30 Features				
Amazon Echo	100	100	100	100
Philips Hue	100	100	100	100
Somfy Door Lock	100	100	100	100
40 Features				
Amazon Echo	100	100	100	100
Philips Hue	100	100	100	100
Somfy Door Lock	100	100	100	100

We used three cross-validation tests to determine the feasibility and over-fitting of our implemented IoT device recognition methodology via a reduced dataset. There was no change in the results of the cross-validation test. A model based on the flow-based features of the IoT network is proposed in [25]. Flow-based classification techniques only examine header data to classify the activity of the network. Flow-based features have been validated and assessed by the IoT Botnet dataset via numerous machine learning algorithms. This proposed framework is a two-level architecture designed to enable the detection of abnormal activity on IoT networks. The first level of the model aims to classify the device's stream as normal or abnormal, while the category of malicious behavior is described by the level two model. If the device flow is abnormal in the first level model, then the system moves the flow to the level two model to determine the type of abnormality. In this paper, 17 flow features for the proposed flow-based device identification model are empirically selected from our adopted IoT-AD-20 dataset. Table 5 describes the flow features used by our proposed model. A high detection rate can be achieved if a perfect classification model is created. A performance analysis was carried out to choose the most excellent classification method. A classifier that uses a decision tree produces strong prediction outcomes. The proposed model obtained a 100 % accuracy, precision, recall, and F score for all device recognition, as seen in Table 6. We used 3, 5, and 10-fold cross-validation tests to evaluate our implemented flow-based IoT device identification methodology. We present 10-fold cross-validation tests of our proposed flow based model in Table 6. The result of our proposed IoT device identification technique remains unchanged.

Table 5. Selected Flow Features IoT-AD-20 Dataset

Feature Name	Feature Name
Flow ID	Src Port
Dst Port	Protocol
Flow Duration	Flow Byts/s
Flow Pkts/s	Flow IATMean
Flow IAT Std	Flow IAT Max
Flow IAT Min	Fwd IAT Tot
Fwd IAT Mean	Subflow Fwd Pkts
Subflow Fwd Byts	Subflow Bwd Pkts
Subflow Bwd Byts	Device Type

Table 6. Accuracy, Precision, Recall, and F score Based on Flow Features

Device	Accuracy	Precision	Recall	F Score
Amazon Echo	100	100	100	100
Philips Hue	100	100	100	100
Somfy Door Lock	100	100	100	100
10-Fold Cross-Validation				
Amazon Echo	100	100	100	100
Philips Hue	100	100	100	100
Somfy Door Lock	100	100	100	100

Meidan et al. [27] have created an IoT botnet dataset. They created the dataset using nine IoT devices. The dataset includes 115 attributes for normal and malicious networks stream. The data set consists of different normal network activities for each industrial device to maintain regular network activity in the training set. The accuracy, precision, recall, and F score was 96 % using the full dataset shown in Table 7. We applied the learning curve to check the minimum number of instances needed for training to reach the highest detection rate. The learning curve concludes that a minimum of seven thousand instances is necessary to achieve better detection results for IoT device identification. Figure 5 presents the learning curve for the IoT botnet dataset. We used three K-fold cross-validation tests to determine the possible validity and performance of the IoT system recognition methodology that we implemented for the IoT botnet data set. In Table 8, we present a 10-fold cross-validation test of our suggested model IoT botnet dataset.

Table 7. Accuracy, Precision, Recall, and F Score Botnet Dataset

Device	Accuracy	Precision	Recall	F Score
Bmonitor	96	99	99	99
DDoorbell	96	99	99	99
EDoorbell	96	97	98	97
SCamera1002	96	99	99	99
SCamera1003	96	97	97	97
SCamera737	96	90	90	90
SCamera838	96	88	88	88
Thermostat	96	99	98	98
Webcam	96	98	98	98

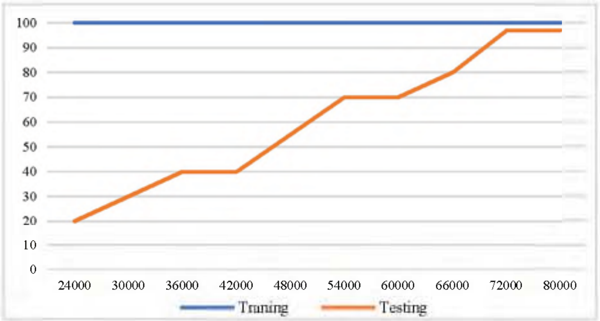

Fig.5. Learning Curve for Device Type Identification

Table 8. Accuracy, Precision, Recall, and F Score 10-Fold Cross-Validation Botnet Dataset

IoT Botnet	Accuracy	Precision	Recall	F Score
20 Features	92	92	92	92
40 Features	94	94	94	94
60 Features	94	94	94	94

VI. Discussion

Our proposed model obtained better performance using the IoT-AD-20 dataset for IoT device identification. The IoT-AD-20 dataset generated from the IoT-23 dataset's pcap file provides better performance. The dataset is generated based on flow-based feature-capability, so each instance of the dataset has the same source IP, source port number, destination IP, and destination port number. The full data set comprises 80 network features and a label feature. A full features dataset produced 100 % device identification for all devices. A 15 features model also generates 100 % identification of all devices. The amount of training data needed to recognize various IoT devices is critical in classifying IoT devices. We test and analyze our adapted dataset via a different set of instances. We have used the learning curve to search an optimal range of instances that create a higher detection rate. The learning curve shows that three hundred instances are ideal for identifying IoT devices to achieve a higher detection performance. We also used various sets of features selected through the technique of recursive function elimination. It is found that, in order to achieve the highest rate of identification, 15 to 30 features are considered to be an optimal collection of features. Table 4 shows a 100% accuracy, precision, recall, and F score was achieved through 15 features of the IoT-AD-20 dataset.

Our adapted IoT-AD-20 dataset includes 17 flow-based features, so to evaluate our proposed model for flow-based capabilities, we chose all flow-based features empirically. Our proposed model has achieved 100 % recognition for all IoT devices using flow-based features, as shown in Table 6. We have compared the proposed adapted dataset to the IoT botnet dataset. The detection rate achieved via the IoT botnet dataset was not satisfactory. In addition, the RFE feature selection model was applied to the IoT botnet dataset, and 20, 40, 60 IoT botnet dataset features were chosen. Subsets have been shown to decrease further the detection capabilities of the proposed IoT botnet dataset model. As we reduce the number of instances to 1000 for each device, IoT devices' detection rate is further reduced. The accuracy, precision, recall, and F score was measured as 82 percent, respectively. Our proposed model offers the following benefits, high-rank features; few instances are required to train the model, short time for training and testing, more simplified selection of features due to the high-rank features, high detection capability, many flow-based features, identification of devices based on flow features.

A limitation of our research is that we did not have enough devices to test. This is a common limitation among most IoT-related works because of the lack of publicly accessible datasets for many devices. Only three devices make up the experimental connected smart home. Another weakness is that IoT devices are solely part of our testing network, which would not be the case in the actual world. Most traffic is created by smartphones or laptops inside a general-purpose network. The model should be trained and evaluated in wider IoT networks comprising many IoT devices and other network devices.

VII. Conclusion and Future Work

This paper aimed to establish a framework for analyzing network traffic flow data to classify IoT devices. To describe network behavior, we built a dataset with 80 network attributes. Three subsets of the dataset evaluate our proposed model by analytical evaluation (full features, reduced features, and flow-based features). Our proposed model and dataset's validity were tested using accuracy, precision, recall, and F score. Our suggested model achieves 100% accuracy, precision, recall, and F score. We have demonstrated that IoT devices can be identified accurately by our proposed model. This paper focuses on the identification of IoT devices by analyzing flow-based network traffic. An intruder can use IoT device classification to expose vulnerable IoT devices by conducting a constructive network traffic flow analysis. Device profile and device identification can allow the network administrator to identify infected sensors in IoT networks. The IT manager can also use the sensor profile to enforce various security policies for different IoT devices.

For future work, we plan to validate our proposed model on IoT networks containing a large number of IoT devices.

REFERENCES

[1] "Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025," [Online]. Available: https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide.

[2] N. Apthorpe, D. Reisman, and N. Feamster, "A Smart Home is No Castle: Privacy Vulnerabilities of Encrypted IoT Traffic," 2017, [Online]. Available: http://arxiv.org/abs/1705.06805.

[3] 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," Proc. - Int. Conf. Distrib. Comput. Syst., pp. 2177–2184, 2017, doi: 10.1109/ICDCS.2017.283.

[4] S. Doyne, "Hackers Used New Weapons to Disrupt Major websites Across U.S." [Online]. Available: https://www.nytimes.com/2016/10/24/learning/questions-for-hackers-used-new-weapons-to-disrupt-major-websites-across-us.html.

[5] "FCC Adopts Broadband Consumer Privacy Rules." [Online]. Available: https://www.fcc.gov/document/fcc-adopts-broadband-consumer-privacy-rules.

[6] V. Brik, S. Banerjee, M. Gruteser, and S. Oh, "Wireless device identification with radiometric signatures," Proc. Annu. Int. Conf. Mob. Comput. Networking, MOBICOM, pp. 116–127, 2008, doi: 10.1145/1409944.1409959.

[7] T. D. Nguyen, S. Marchal, M. Miettinen, H. Fereidooni, N. Asokan, and A. R. Sadeghi, "DÏoT: A federated self-learning anomaly detection system for IoT," Proc. - Int. Conf. Distrib. Comput. Syst., vol. 2019-July, pp. 756–767, 2019, doi: 10.1109/ICDCS.2019.00080.

[8] R. Falk and S. Fries, "Managed certificate whitelisting - A basis for internet of things security in industrial automation applications," Secur. 2014 - 8th Int. Conf. Emerg. Secur. Information, Syst. Technol., no. November 2014, pp. 167–172, 2014.

[9] J. Franklin, D. McCoy, P. Tabriz, V. Neagoe, J. van Randwyk, and D. Sicker, "Passive data link layer 802.11 wireless device driver fingerprinting," 15th USENIX Secur. Symp., no. December, pp. 167–178, 2006.

[10] S. V. Radhakrishnan, A. S. Uluagac, and R. Beyah, "GTID: A Technique for Physical Device and Device Type Fingerprinting," IEEE Trans. Dependable Secur. Comput., vol. 12, no. 5, pp. 519–532, 2015, doi: 10.1109/TDSC.2014.2369033.

[11] M. Miettinen et al., "IoT Sentinel Demo: Automated Device-Type Identification for Security Enforcement in IoT," Proc. - Int. Conf. Distrib. Comput. Syst., pp. 2511–2514, 2017, doi: 10.1109/ICDCS.2017.284.

[12] S. Siby, R. R. Maiti, and N. O. Tippenhauer, "IoTScanner: Detecting privacy threats in IoT neighborhoods," IoTPTS 2017 - Proc. 3rd ACM Int. Work. IoT Privacy, Trust. Secur. co-located with ASIA CCS 2017, pp. 23–30, 2017, doi: 10.1145/3055245.3055253.

[13] H. Kawai, S. Ata, N. Nakamura, and I. Oka, "Identification of communication devices from analysis of traffic patterns," 2017 13th Int. Conf. Netw. Serv. Manag. CNSM 2017, vol. 2018-Janua, pp. 1–5, 2017, doi: 10.23919/CNSM.2017.8256018.

[14] A. Rizzi, S. Colabrese, and A. Baiocchi, "Low complexity, high performance neuro-fuzzy system for Internet traffic flows early classification," 2013 9th Int. Wirel. Commun. Mob. Comput. Conf. IWCMC 2013, pp. 77–82, 2013, doi: 10.1109/IWCMC.2013.6583538.

[15] A. Dainotti, A. Pescapé, and C. Sansone, "Early classification of network traffic through multi-classification," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6613 LNCS, pp. 122–135, 2011, doi: 10.1007/978-3-642-20305-3_11.

[16] The Royal Academy of Engineering, Smart infrastructure: the future, vol. 72, no. Fall 2011. 2012.

[17] "Aposemat IoT-23: A labeled dataset with malicious and benign IoT network traffic," [Online]. Available: https://www.stratosphereips.org/datasets-iot23.

[18] A. H. Lashkari, G. D. Gil, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of tor traffic using time based features," ICISSP 2017 - Proc. 3rd Int. Conf. Inf. Syst. Secur. Priv., vol. 2017-Janua, no. September, pp. 253–262, 2017, doi: 10.5220/00061056 02530262.

[19] "IoT-AD-20," [Online]. Available: https://sites.google.com/view/iotdataset1.

[20] I. Ullah and Q. H. Mahmoud, "A Technique for Generating a Botnet Dataset for Anomalous Activity Detection in IoT Networks," IEEE Trans. Syst. Man, Cybern. Syst., vol. 2020-Octob, pp. 134–140, 2020, doi: 10.1109/SMC42975.2020. 9283220.

[21] I. Ullah and Q. H. Mahmoud, "A filter-based feature selection model for anomaly-based intrusion detection systems," Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017, vol. 2018-Janua, pp. 2151–2159, 2017, doi: 10.1109/BigData.2017.8258163.

[22] I. Ullah and Q. H. Mahmoud, A Scheme for Generating a Dataset for Anomalous Activity Detection in IoT Networks, vol. 12109 LNAI. Springer International Publishing, 2020.

[23] I. Ullah and Q. H. Mahmoud, "An intrusion detection framework for the smart grid," Can. Conf. Electr. Comput. Eng., pp. 10–14, 2017, doi: 10.1109/CCECE.2017.7946654.

[24] I. Ullah and Q. H. Mahmoud, "A Hybrid Model for Anomaly-based Intrusion Detection in SCADA Networks," Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017, pp. 2160–2167, 2020, doi: 10.1109/ACIT50332.2020.9299965.

[25] I. Ullah and Q. H. Mahmoud, "A two-level flow-based anomalous activity detection system for IoT networks," Electron., vol. 9, no. 3, 2020, doi: 10.3390/electronics9030530.

[26] I. Ullah and Q. H. Mahmoud, "A Two-Level Hybrid Model for Anomalous Activity Detection in IoT Networks," 2019 16th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC), Las Vegas, NV, USA, vol. 9, no. 3, 2020, doi: 10.3390/electronics9030530.

[27] Y. Meidan et al., "N-BaIoT-Network-based detection of IoT botnet attacks using deep autoencoders," IEEE Pervasive Comput., vol. 17, no. 3, pp. 12–22, 2018, doi: 10.1109/MPRV.2018.03367731.