RES MILITARIS
Social Science Journal

Enhancing IoT Device Classification: Leveraging Machine Learning for Efficient Network Management and Security

Dr. Sk. Mahaboob Basha[1*], V. Chandana[1], K. Sreeja[1], K. Naveen[1], K. Nivas[1]

[1]Department of CSE, Sree Dattha Institute of Engineering and Science, Sheriguda, Hyderabad, Telangana, India

Corresponding E-mail: mahaboob@sreedattha.ac.in

Abstract

Given the widespread use of IoT devices in areas such as smart homes, healthcare, and industry, it is crucial to have effective techniques for categorizing and overseeing these devices. Understanding the behavior of devices connected to a network is essential for effective network traffic analysis. Through the analysis of network traffic characteristics, one can deduce the type and behavior of IoT devices. In the past, the classification of IoT devices was based on manual inspection or simple heuristics, but these methods were not sufficient to handle the increasing complexity and diversity of IoT devices. Similar to a data scientist, the traditional systems faced limitations in terms of scalability, accuracy, and resource intensiveness, which called for more advanced approaches. This project is highly significant as it has the potential to automate and improve the efficiency of device classification, resulting in better network management, enhanced security, and optimized resource utilization. With the help of machine learning algorithms such as Random Forest, the system ensures flexibility in adapting to changing network patterns. It also offers a user-friendly interface for seamless interaction, ultimately working towards the larger objective of building smarter and safer IoT environments.

Keywords: Internet of Things, Network traffic analysis, Machine learning, Random forest classifier.

1. Introduction

The proliferation of the Internet of Things (IoT) has led to the integration of a diverse array of smart devices into various environments, including homes, businesses, and cities. These devices, ranging from cameras and lights to motion sensors and health monitors, are designed to enhance convenience, efficiency, and security. However, the rapid expansion of IoT technology introduces significant challenges related to device management and cybersecurity. Operators often lack visibility into the full spectrum of IoT devices within their environments, making it difficult to ensure that each device functions correctly and is safeguarded against cyber threats. In this study, we propose a comprehensive framework for classifying IoT devices based on network traffic characteristics. By leveraging traffic data collected from a smart environment equipped with 28 different types of IoT devices over six months, we can gain valuable insights into the unique traffic patterns and behaviors of these devices. This dataset, partially released for public use, serves as the foundation for our analysis and the development of a robust classification system. The rise of IoT technology has transformed smart environments, embedding a multitude of interconnected devices designed to improve daily operations and user experiences. However, this technological advancement comes with significant challenges. One of the primary issues is the lack of comprehensive visibility and control over the diverse range of IoT devices deployed within these environments. Operators often struggle to maintain an up-to-date inventory of their IoT assets, leading to difficulties in monitoring device functionality and security status.

IoT devices are inherently diverse, with varying communication protocols, operational behaviors, and security mechanisms. This diversity complicates efforts to develop standardized methods for device classification and monitoring. Without a reliable classification framework, it becomes challenging to detect anomalous behavior that might indicate device malfunctions or cyber-attacks. Consequently, operators face increased risks of security breaches, privacy violations, and operational disruptions. Existing methods for IoT device management typically rely on manual inventory processes or specialized monitoring equipment, both of which are impractical for large-scale deployments. Manual methods are time-consuming and error-prone, while specialized equipment adds to the cost and complexity of the infrastructure. There is an urgent need for an automated, efficient, and scalable solution that can classify and monitor IoT devices based on readily available network traffic data.

2. Literature Survey

The number of devices connecting to the Internet is ballooning, ushering in the era of the "Internet of Things" (IoT). IoT refers to the tens of billions of low cost devices that communicate with each other and with remote servers on the Internet autonomously. It comprises everyday objects such as lights, cameras, motion sensors, door locks, thermostats, power switches and household appliances, with shipments projected to reach nearly 20 billion by 2020 [1]. Thousands of IoT devices are expected to find their way in homes, enterprises, campuses and cities of the near future, engendering "smart" environments benefiting our society and our lives.

The proliferation of IoT, however, creates an important problem. Operators of smart environments can find it difficult to determine what IoT devices are connected to their network and further to ascertain whether each device is functioning normally. This is mainly attributed to the task of managing assets in an organization, which is typically distributed across different departments. For example, in a local council, lighting sensors may be installed by the facilities team, sewage and garbage sensors by the sanitation department and surveillance cameras by the local police division. Coordinating across various departments to obtain an inventory of IoT assets is time consuming, onerous, and error-prone, making it nearly impossible to know precisely what IoT devices are operating on the network at any point in time. Obtaining "visibility" into IoT devices in a timely manner is of paramount importance to the operator, who is tasked with ensuring that devices are in appropriate network security segments, are provisioned for requisite quality of service, and can be quarantined rapidly when breached. The importance of visibility is emphasized in Cisco's most recent IoT security report [2], and further highlighted by two recent events: sensors of a fishtank that compromised a casino in Jul 2017 [3], and attacks on a University campus network from its own vending machines in Feb 2017 [4]. In both cases, network segmentation could have potentially prevented the attack and better visibility would have allowed rapid quarantining to limit the damage of the cyber-attack on the enterprise network.

One would expect that devices can be identified by their MAC address and DHCP negotiation. However, this faces several challenges: (a) IoT device manufacturers typically use NICs supplied by third-party vendors, and hence the Organizationally Unique Identifier (OUI) prefix of the MAC address may not convey any information about the IoT device; (b) MAC addresses can be spoofed by malicious devices; (c) many IoT devices do not set the Host Name option in their DHCP requests [5]; (d) even when the IoT device exposes its host name it may not always be meaningful; and lastly (e) these host names can be changed by the user (e.g. the HP printer can be given an arbitrary host name). For these reasons, relying on DHCP infrastructure is not a viable solution to correctly identify devices at scale.

In this project, we address the above problem by developing a robust framework that classifies each IoT device separately in addition to one class of non-IoT devices with high accuracy using statistical attributes derived from network traffic characteristics. Qualitatively, most IoT devices are expected to

send short bursts of data sporadically. Quantitatively, our preliminary work in [6] was one of the first attempts to study how much traffic IoT devices send in a burst and how long they idle between activities. We also evaluated how much signalling they perform (e.g., domain lookups using DNS or time synchronization using NTP) in comparison to the data traffic they generate. This paper significantly expands on our prior work by employing a more comprehensive set of attributes on trace data captured over a much longer duration (of 6 months) from a testbed comprising different IoT devices.

There is no doubt that it is becoming increasingly important to understand the nature of IoT traffic. Doing so helps contain unnecessary multicast/broadcast traffic, reducing the impact they have on other applications. It also enables operators of smart cities and enterprises to dimension their networks for appropriate performance levels in terms of reliability, loss, and latency needed by environmental, health, or safety applications. However, the most compelling reason for characterizing IoT traffic is to detect and mitigate cybersecurity attacks. It is widely known that IoT devices are by their nature and design easy to infiltrate [7], [8], [9], [10], [11], [12]. New stories are emerging of how IoT devices have been compromised and used to launch large-scale attacks [13]. The large heterogeneity in IoT devices has led researchers to propose network-level security mechanisms that analyse traffic patterns to identify attacks (see [14], [15]); success of these approaches relies on a good understanding of what "normal" IoT traffic profile looks like.

3. Proposed Methodology

By meticulously following these step-by-step procedures, our research endeavors to contribute significantly to the field of IoT device classification within smart environments, paving the way for enhanced monitoring, functionality assessment, and cybersecurity measures in IoT ecosystems. Figure 1 shows the proposed system mode. The detailed operation is as follows:

Step 1: The first step in this research journey involves gathering data. Specifically, we focus on the network traffic generated by IoT devices in smart environments. To accomplish this, we establish an experimental setup within a smart environment, equipping it with a diverse range of 28 IoT devices spanning various categories such as cameras, lights, motion sensors, and health monitors. Over a period of 6 months, we diligently collected and synthesize traffic traces emanating from these devices. Moreover, to foster collaboration and advancement within the research community, we selectively release a portion of this collected data as open data for wider usage and exploration.

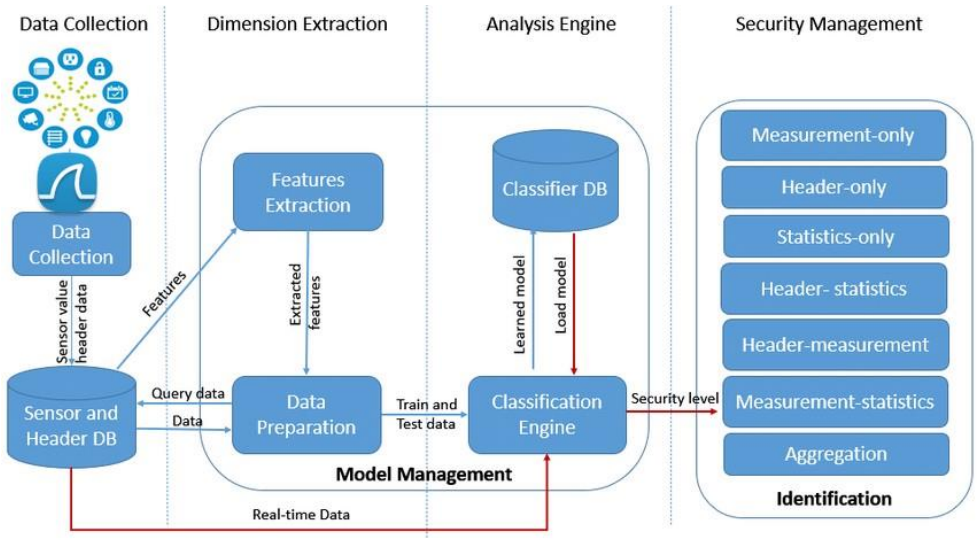

Figure 1: System architecture of proposed IoT device identification using machine learning approach.

Step 2: Existing Naive Bayes Theorem

Building upon the foundation of collected data, the next phase delves into understanding the underlying network traffic characteristics. Leveraging statistical attributes such as activity cycles, port numbers, signaling patterns, and cipher suites, we gain valuable insights into the behaviors exhibited by different IoT devices within the smart environment. This comprehension serves as the cornerstone for our subsequent classification endeavors.

Step 3: Proposed Random Forest Classifier

With a comprehensive understanding of the network traffic characteristics, we proceed to develop a robust classification framework. Central to this framework is a multi-stage machine learning-based classification algorithm. Our approach utilizes the Random Forest Classifier, a powerful ensemble learning technique known for its versatility and effectiveness in handling complex classification tasks. Through meticulous training and validation, we demonstrate the efficacy of our proposed classifier in accurately identifying specific IoT devices based solely on their network activity, achieving an impressive accuracy rate exceeding 99%.

Step 4: Performance Comparison

To provide a comprehensive evaluation of our proposed framework, we undertake a thorough analysis of its performance characteristics. This entails a meticulous examination of the trade-offs between various factors including cost, speed, and classification accuracy. By meticulously dissecting these trade-offs, we offer valuable insights into the practical feasibility and scalability of deploying our classification framework in real-world smart environments.

Step 5: Prediction of Output from Test Data with Trained Model

In the final step of our research procedure, we put our classification framework to the test by predicting outputs from unseen test data. Leveraging the trained Random Forest Classifier, we input test data samples representing network traffic from previously unseen IoT devices. Subsequently, we analyze the predictions generated by our model, thereby evaluating its ability to accurately classify and identify IoT devices based solely on their network activity patterns.

3.1 Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. Figure 2 explains the working of the Random Forest algorithm. Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier: There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result. The predictions from each tree must have very low correlations.

- It takes less training time as compared to other algorithms.

- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase. The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

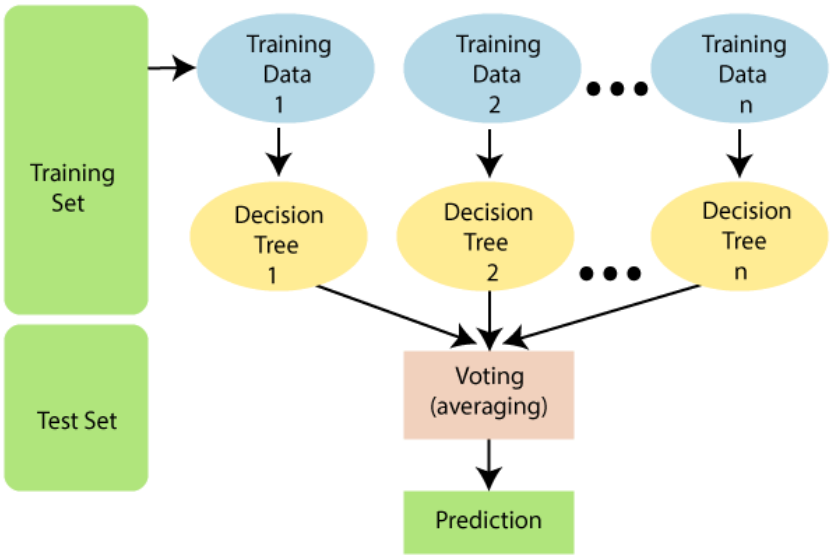

Figure 2. RFC block diagram.

4. Results and Discussion

Figure 3 illustrates the output after running the Bag of Goods (BOG) and Naive Bayes (NB) processes on the uploaded dataset. The BOG process transforms the raw data into a format suitable for machine learning by extracting relevant features like Rate, Port, Domain, Cipher, and Device. The displayed sample data includes entries such as "54,6,443,47559,0" and "97,6,443,53911,0", where each entry represents a distinct network traffic characteristic mapped to a specific device class. This preprocessing step is essential for accurate classification.

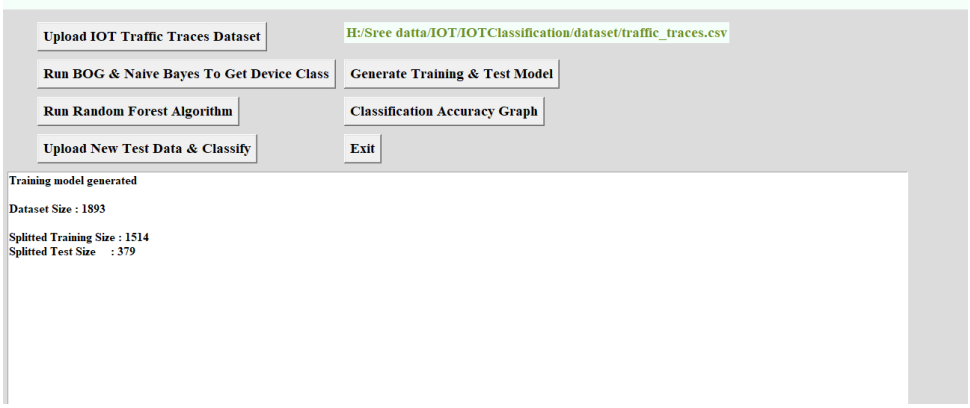

Figure 4: After generate training & test model.

The Figure 4 shows that after generating the training and test models, the GUI displays the details of the dataset split. The total dataset size is 1893 entries, with 1514 entries allocated for training and 379 entries for testing. This split is typically done to ensure that the model is trained on a substantial amount of data while reserving a portion for evaluating the model's performance. The model is now ready to be trained using the Random Forest Classification (RFC) algorithm.

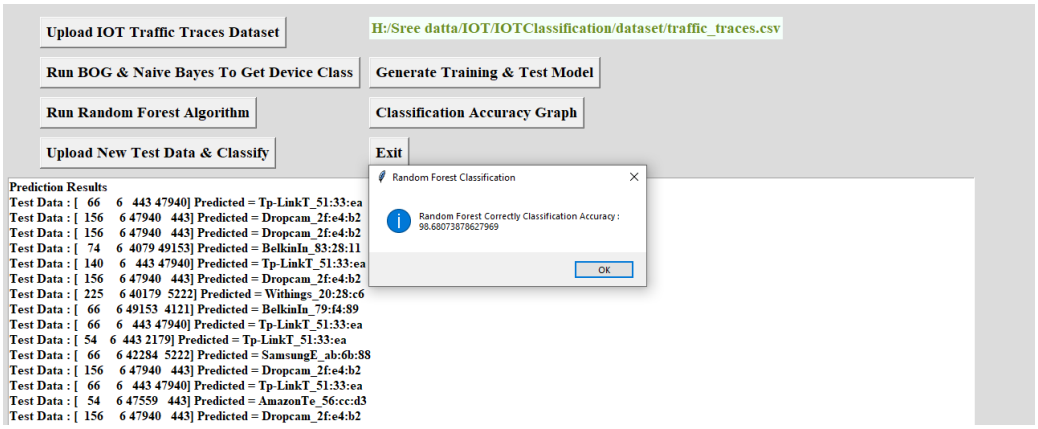

Figure 5: Applied RFC Algorithm

This figure 5 shows the results of applying the Random Forest Classification (RFC) algorithm to the test data. The algorithm predicts the device labels based on their network activity. Sample predictions include:

- Test Data: [66, 6, 443, 47940] Predicted = Tp-LinkT_51:33:ea
- Test Data: [156, 6, 47940, 443] Predicted = Dropcam_2f:e4:b2

These predictions demonstrate the model's capability to identify specific IoT devices. The accuracy of the RFC algorithm is noted as 98%, indicating a high level of precision in classification. This high accuracy is crucial for ensuring reliable monitoring and management of IoT devices in smart environments.

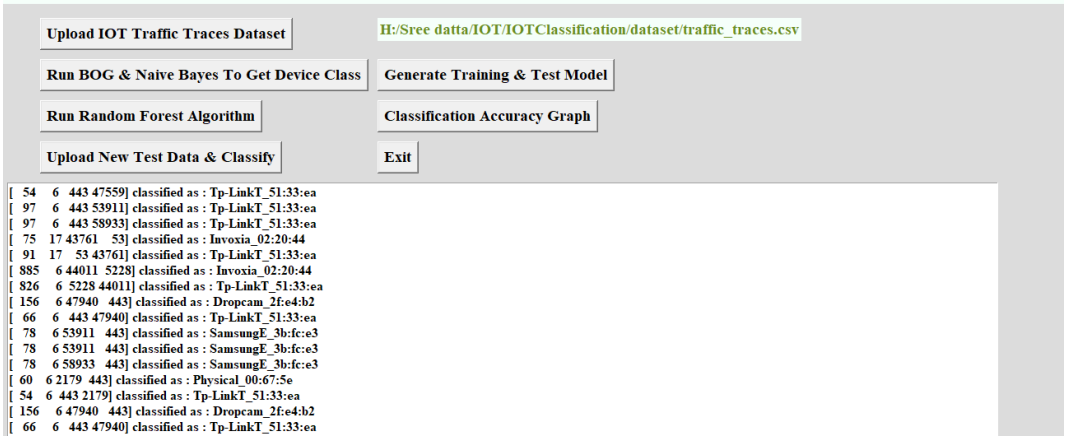

Figure 6: Output of New test data

In this figure 6, the output of classifying new test data is displayed. The new data entries are classified into specific IoT devices, showcasing the model's real-time application:

- [74, 6, 443, 46960] classified as : Tp-LinkT_51:33:ea
- [66, 6, 46960, 443] classified as : SamsungE_3b:fc:e3
- [75, 17, 43761, 53] classified as : Invoxia_02:20:44

These classifications demonstrate the model's practical utility in identifying devices based on their network traffic characteristics. The consistency and accuracy of these classifications affirm the model's effectiveness in a live environment, making it a valuable tool for IoT device management.

5. Conclusion

In the pursuit of enhancing smart environments, the classification of IoT devices based on network traffic characteristics emerges as a crucial endeavor. This paper presents a comprehensive framework addressing the challenges associated with identifying and monitoring IoT assets within smart environments. Through extensive experimentation, the study showcases the effectiveness of machine learning techniques in accurately classifying various IoT devices, achieving a remarkable accuracy rate of over 99%. One of the significant contributions of this research lies in the creation of an extensive dataset comprising traffic traces from a diverse array of IoT devices. By synthesizing and analyzing this data, the study provides valuable insights into the underlying network traffic patterns exhibited by different IoT devices. Statistical attributes such as activity cycles, port numbers, signaling patterns, and cipher suites are employed to characterize and differentiate between devices. This not only aids in device identification but also offers a deeper understanding of their behavior within smart environments. The proposed framework for classifying IoT devices based on network traffic characteristics lays a solid foundation for future advancements and extensions. While the current study achieves remarkable accuracy in device identification, there are several avenues for expanding and refining the framework's features and capabilities.

References

[1] I. Spectrum. (Last accessed July 2017.) Popular Internet of Things forecast of 50 billion devices by 2020 Is outdated. https://goo.gl/ 6wSUkk.

[2] Cisco, "Cisco 2017 Midyear Cybersecurity Report," Tech. Rep., 2017.

[3] A. Schiffer. (2017) How a fish tank helped hack a casino. https: //goo.gl/SAHxCX.

[4] Ms. Smith. (2017) University attacked by its own vending machines, smart light bulbs & 5,000 IoT devices. https://goo.gl/ cdNJnE.

[5] S. Alexander and R. Droms, "DHCP Options and BOOTP Vendor Extensions," Internet Requests for Comments, RFC Editor, RFC 2132, March 1997. [Online]. Available: https: //tools.ietf.org/rfc/rfc2132.txt

[6] A. Sivanathan et al., "Characterizing and Classifying IoT Traffic in Smart Cities and Campuses," in Proc. IEEE Infocom Workshop on Smart Cities and Urban Computing, Atlanta, USA, May 2017.

[7] S. Notra et al., "An Experimental Study of Security and Privacy Risks with Emerging Household Appliances," in Proc. M2MSec, Oct 2014.

[8] F. Loi et al., "Systematically Evaluating Security and Privacy for Consumer IoT Devices," in Proc. ACM CCS workshop on IoT Security and Privacy (IoT S&P), Texas, USA, Nov 2017.

[9] I. Andrea et al., "Internet of Things: Security vulnerabilities and challenges," in 2015 IEEE Symposium on Computers and Communication (ISCC), July 2015.

[10] K. Moskvitch, "Securing IoT: In your Smart Home and your Connected Enterprise," Engineering Technology, vol. 12, April 2017.

[11] N. Dhanjani, Abusing the Internet of Things: Blackouts, Freakouts, and Stakeouts. O'Reilly Media, 2015.

[12] E. Fernandes et al., "Security Analysis of Emerging Smart Home Applications," in 2016 IEEE Symposium on Security and Privacy (SP). IEEE, may 2016.

[13] T. guardian. (2016) Why the internet of things is the new magic ingredient for cyber criminals. https://goo.gl/MuH8XS.

[14] T. Yu et al., "Handling a Trillion (Unfixable) Flaws on a Billion Devices: Rethinking Network Security for the Internet-of-Things," in Proc. ACM HotNets, Nov 2015.

[15] A. Sivanathan et al., "Low-Cost Flow-Based Security Solutions for Smart-Home IoT Devices," in Proc. IEEE ANTS, Nov 2016.