International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Unique Identification Using M2M Approach Based on Internet of Things

Md Mahmud and Mr. Sudhir Goswami

Department of Computer Science and Engineering, School of Research and Technology, People's University, Bhopal, Madhya Pradesh, India

ABSTRACT

The Internet of Things (IoT) is progressively becoming ingrained in daily life. Smart homes, farms, and industries can all benefit from IoT devices. Thousands of IoT devices are linked to the internet every day. Nevertheless, IoT devices are gaining popularity despite their modest build-up and limited capability. This creates security issues since an attacker may compromise the entire network by using one IoT device as an access point. Furthermore, the number of businesses manufacturing IoT devices is always expanding, even though a large portion of them have little to no knowledge of IoT security. As a result, an attacker can get access to a network by using a compromised Internet of Things device. Nevertheless, the likelihood of hackers exploiting IoT device security flaws that rely on widely accepted techniques is decreasing. Mobile Network Operators (MNOs) regularly examine the traffic produced by these devices-whether they are rogue, malfunctioning, or authentic-to determine their identities (fingerprints) and characteristics. When an Internet of Things device is hacked and connected to the network, the finger-printing technique works fast to determine its real identity. We describe an ML-based approach to IoT device identification in this work. Recognizing the IoT devices connected to a network is crucial for locating and eliminating rogue devices as well as for comprehending the security flaws in a network as a whole. Several earlier studies have tried to apply ML to this procedure.

Keywords: Iot, Fingerprinting, DSN, F1 score.

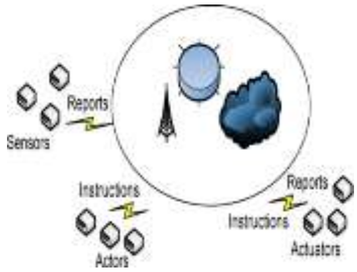
1. Introduction

The "internet of things" refers to the network of numerous devices—cars, buildings, and other machinery—that are outfitted with sensors, actuators, electronics, software, and network connectivity so they can collect and exchange data. The term "Internet of Things" was first used in 1999 by Kevin Ashton. However, the largest IoT research and development initiatives started in 2010. The Internet of Things (IoT) is expected to create new applications and bridge many technologies by connecting physical objects to enable intelligent decision-making.

Thanks to the invention of TCP/IP in the early 1980s, two computers could communicate with one another through a computer network before the Internet of Things was widely researched. A brief introduction to layered architecture can be found in the next section. Since the internet was made available for commercial use, for example, the generation has been able to improve technology at a quick pace. Access to the World Wide Web (www) was made feasible later, in 1991. The following phase of the internet's development saw mobile devices connected to the network, giving rise to the mobile-Internet. The term "internet of things" refers to the future state in which every object in our environment will be able to connect to one another.

1.1 Uniqueness

De-duplication of demographic and biometric data is the uniqueness process. Each person only needs to enroll once for Aadhaar, and after de-duplication, just one Aadhaar card would be issued. After every person receives an Aadhaar number, the de-duplication process checks the resident's biometric and demographic information against the CIDR, compares it with data that has already been recorded, and collects more information during the registration process. One could look through the records in the UIDAI database to find out if a resident is already listed or not. All someone has to do is register for Aadhaar once.



Anything Any Device
Anyone Anybody
Any Service Any Business
Any Path Any Network
Any Place Anywhere
The INTERNET of THINGS

Figure 1 Network model for wireless M2M communication Figure 2 Internet of things

1.2 Machine-To-Machine (M2m) & Internet of Things (IoT) Network

M2M stands for machine-to-machine communication. It is a direct communication system that uses wired or wireless communication routes between the devices without involving any human involvement. It collects data and sends it to other devices that are connected. It's a technique that allows devices to communicate with each other without utilizing the internet. Numerous industries, including defense, industry, tracking and monitoring, and facility management, use M2M communications.

M2M technology is present in many settings, such as residences, workplaces, and shopping centers. A classic example of a machine to machine communication system is the ability to manage electrical equipment like fans and lightbulbs using Bluetooth on a smartphone. Here, the electrical devices and the smartphone are the two devices interacting with each other.

Machine-to-machine (M2M) technology aims to transmit sensor data to a network. M2M systems typically employ public networks and access methods like Ethernet or cellular in order to be less expensive than SCADA or other remote monitoring technologies. The fundamental components of an M2M system are sensors, RFID, a Wi-Fi or cellular communications link, and software for autonomic computing, which helps a network device read data and make choices. These M2M apps translate the information so that preprogrammed, automatic actions can be started.

Though many people use the terms interchangeably, M2M and IoT are not the same. IoT is not necessary for M2M, but M2M is necessary for IoT. Although both connected and standalone devices are described, M2M systems are often distinct, stand-alone networks of devices. Machine-to-machine (M2M) communication is raised to a new level by IoT technologies, which integrate several systems into a single, vast ecosystem. While M2M systems use point-to-point connections between machines, sensors, and hardware across cellular or wired networks, IoT systems rely on IP-based networks to send data received from IoT-connected devices to gateways, the cloud, or middleware platforms.

1.3 Internet of Things

The innovative MIT Auto-ID Center's founders, Kevin Ashton and D. Brock, presented the idea of the Internet of Things (IoT) in 1999 and 2021, respectively, nearly 15 years ago. According to Kevin et al.'s projection, there is connectivity among all electronic devices, and all physical and electronic objects have electronic labels that contain pertinent data. "The Internet of Things allows people and things to be connected Anytime, Anyplace, with Anything and Anyone, ideally using Any path/network and Any service," is how the term "Internet of Things" is defined (see Figure 2).

1.4 Objective

- Demand-driven, portable service delivery-UID is beneficial for demand-driven and portable service delivery system since entity can authenticate their Aadhaar anywhere.
- Access to relevant MIS (Management Information System) and empowerment of beneficiary -Aadhaar can be used to and provide self-help facilities for activities such as checking their entitlements, services delivery timeline, log grievances etc through self-service kiosks, mobile phones, call centers etc.
- Address verification -To provide an authentic verification mechanism,
- Our method uses an enhanced combination of features from previous work and includes an approach for dealing with unbalanced device data via data augmentation.
- We further demonstrate how to enhance device identification via a group-wise data aggregation.
- We provide a comparative evaluation of our method against two recent identification methods using five public IoT datasets (Aalto University, UNSW-Sydney IEEE TMC, IoTFinder, UNSW-Sydney ACM SOSR, and IoT Network Intrusion Dataset) which together contain data from over 100 devices, two of which include both benign and malicious data.

- Through our evaluation we demonstrate improved performance over previous results with F1 scores above 99%, with considerable improvement gained from data aggregation

2. Literature Review

A. Aksoy and M. H. Gunes 2019, studied on The Internet of Things (IoT) has grown in acceptance and permeated every aspect of our daily lives. Though they don't have the processing power to run cybersecurity software, these gadgets are nevertheless vulnerable to hacking just like any other computing system. Isolating IoT devices and preventing communications from a firewall or gateway to the device is a crucial step in preventing attacks on these kinds of devices. Therefore, network management and security benefit greatly from the IoT device's identity. In this research, we provide approach IDentifier (SysID), an automated approach for classifying device features based on their network traffic. SysID employs a single packet sent from the device to identify the type of packet it is [1].

A. Sivanathan et al. 2018, explained on the Internet of Things (IoT) is being hailed as the next wave that is reshaping our society, and smart homes, businesses, and cities are getting more and more IoT devices installed. However, owners of these smart environments might not even be completely aware of all of their IoT assets, let alone know if every IoT device is secure from cyberattacks and operating as intended. In this study, we tackle this problem by creating a solid framework that uses traffic characteristics measured at the network level to classify IoT devices. they offer four distinct contributions. Initially, we outfitted a smart setting with twenty-eight distinct Internet of Things (IoT) gadgets, including cameras, lighting, outlets, motion detectors, appliances, and health monitors. [2].

Ankur Utsav et al. 2020, outlined the concept of a system that uses a sensor to scan the body for heat, storing the user's daily record in a database. A one-time generating method called QR code development is used to identify each user. With the creation of unique QR codes, a new, creative, inventive, and intelligent microcontroller-based temperature monitoring system concept is suggested for keeping an eye on the most recent coronavirus discovery (COVID-19A) [3].

Bastida, D. and Lin, F. 2020, The growing number of IoT applications can be attributed to their diverse and intermittent character, which makes scalability a crucial research concern for developing cloud-based IoT/M2M systems. A dynamic SDN-based network slicing mechanism is proposed in this study to address the scalability issues brought on by the heterogeneity and fluctuation of IoT application requirements. For every new kind of Internet of Things application, the suggested approach may automatically generate a network slice on the spot and dynamically modify the slice's QoS attributes to meet the application's evolving needs [5].

D. D. N. Nguyen et al. 2022, Physical unclonable features (PUF) of IoT devices or radio frequency fingerprinting (RF) were also employed in studies on wireless Internet of Things (IoT) node authentication techniques. Models based on machine learning are essential to these strategies. In this letter, we propose a unique and efficient method for IoT node authentication based on Chi-square distribution theory and Mahalanobis Distance correlation. [6].

Haghnegahdar, L. et al. 2022, outlined the growth of the industrial Internet of things (IIoT) to enable the benefits of more intelligent control and sophisticated production equipment. All production divisions, including the newest and quickly developing technology of additive manufacturing (AM), or 3D printing, have the opportunity to change their patterns thanks to cloud-based technologies that enable remote data gathering, intelligent machine interconnection, and sensor monitoring. AM is a sort of direct manufacturing that uses cutting-edge technology to facilitate complex production and a formation that can expedite supply chain and manufacturing procedures. [10].

I. H. Sarker, 2021, explained in Data from the Internet of Things (IoT), cybersecurity, mobile, social media, business, health, and other sources abound in the digital world. The key to developing intelligent analyses of these data and correspondingly clever and automated applications is understanding artificial intelligence, namely machine learning. There are many different kinds of machine learning algorithms in the field, including supervised, unsupervised, semi-supervised, and reinforcement learning. Furthermore, deep learning, a subset of a larger class of machine learning techniques, is capable of large-scale, intelligent data analysis. provide a thorough overview of these machine learning methods that can be used to improve an application's intelligence and capabilities [13].

J. Brownlee, 2020, outlined the process of converting unprocessed data into a format that machine learning algorithms can use to model it. With the help of concise explanations, commonly used Python libraries, and comprehensive tutorial sessions, you will learn how to securely and efficiently get your data ready for machine learning predictive modeling [14].

R. Osei, et al. 2022, suggested a productive way to fingerprint Internet of Things devices that combines dimensionality reduction methods like PCA and Autoencoder with established classifiers. The latter methods decrease the quantity of IoT data to be processed while extracting the most pertinent elements needed for precise fingerprinting. We tested our suggested method in multiple trials using an actual dataset from an Internet of Things network. The outcomes demonstrate that the Autoencoder for dimensionality reduction using a Decision Tree Algorithm keeps the prediction of the IoT device fingerprints very good (97%) while reducing the number of features from 14 to 5 [24].

S. Zhang, et al. 2021, outlined how a basic problem with the Internet of Things is device identification (IoT). Accurately identifying every distinct device within a network is essential for the provision of numerous essential services, such as intrusion prevention and access control. But in the Internet of Things, device identification has numerous difficulties. As machine learning-assisted methods possess the ability to automate tasks and record dynamic device behaviors, they have great potential for device identification. Techniques with supervised machine learning support show good accuracy in identifying devices. They do, however, come with a challenge: a high quantity of labeled datasets. However, without labeled datasets, unsupervised machine learning can also achieve high accuracy [26].

Salama, R. et al. 2023, when assets or consumer products are equipped with sensors and intelligence and linked to the Internet, the amount of information that can be evaluated online will increase. Machine-to-machine (M2M) communications and the Internet of Things (IoT) are terms used to describe this process. Although the idea and paradigm are not new, there has been a surge in the number and kind of connected devices in addition to advancements in information collection, processing, and sharing technologies. [28].

3. Proposed Methodology

IOT is enabled by the latest developments in (Radio Frequency Identification) RFID, Machine to Machine (M2M) communication, and Near Field Communication (NFC), sensors, actuators, mobile phones, etc. The current revolution in the communication field is the internet of things. In the world of new technologies RFID seemed to be necessary for the IOT.

- Resident provides Aadhaar Number, necessary demographic and biometric details to terminal devices belonging to the AUA/SA (or merchant/operator appointed by AUA/SA) to obtain a service offered by the AUA/SA.
- Aadhaar authentication enabled application software that is installed on the device packages these input parameters, encrypts, and sends it to AUA server over either a mobile/broadband network using AUA specific protocol.
- AUA server, after validation adds necessary headers (AUA specific wrapper XML with license key, transaction id, etc.), and passes the request through ASA server to UIDAI CIDR.
- Aadhaar authentication server returns a “yes/no” based on the match of the input parameters.
- Based on the response from the Aadhaar authentication server, AUA/SA conducts the transaction.

3.1 Problem Statement

One method of securing an Internet of Things (IoT) device network is device identification, which allows devices that are deemed suspect to be removed from the network. We provide IoTDevID, a unique device identification (fingerprinting) technique that models the behavior of IoT devices based on the network packets they exchange via machine learning. Our method incorporates a methodology for handling unbalanced device data via data augmentation, and it makes use of an improved mix of characteristics from earlier work. We also show how a group-wise data aggregation might improve device identification. Using five public IoT datasets (Aalto University, UNSW-Sydney IEEE TMC, IoTFinder, UNSW-Sydney ACM SOSR, and IoT Network Intrusion Dataset*), which collectively contain data from over 100 devices, two of which include both benign and malicious data, provide a comparative evaluation of our method against two recent identification methods. With F1 scores exceeding 99%, our review shows better performance over the past results, with significant improvement coming from data aggregation.

3.2 Requirements and Infrastructure

Python 3.6 was used to create the application files. The technical features of the computer used for experiments are given below.

Table 3.1 Library and their task

Library	Task
Scapy	Packet (Pcap) crafting
Sklearn	Machine Learning & Data Preparation
Imblearn	Data Augmentation
Numpy	Mathematical Operations
Pandas	Data Analysis
Matplotlib	Graphics and Visuality

Table 3.2 Technical feature of the computer used for experiments

Central Processing Unit	Intel(R) Core (TM) i7-7500U CPU @ 2.70GHz 2.90 GHz or above
Random Access Memory	8 GB
Operating System	Windows 10 Pro 64-bit
Graphics Processing Unit	AMD Readon (TM) 530

3.3 Implementation

The implementation phase consists of 5 steps, which are:

- Fingerprinting
- Initial Fingerprint Method Evaluation

- Data Augmentation
- Augmentated and Aggregated Fingerprint Method Evaluation
- Malicious Device Dataset Evaluation

3.3.1 Fingerprinting

This step contains the PCAP2CSV. This converts with pcap extension to single packet-based, csv extension fingerprint (IoT Sentinel, IoTSense, IoTDevID individual packet based feature sets) and makes labeling.

3.3.2 Initial Fingerprint Method Evaluation

This step contains the Classification of Individual packets for Aalto University Dataset. This makes machine learning application for individual packets for Aalto University and allows to compare 3 different feature sets (IoT Sentinel, IoTSense, IoTDevID individual packet based feature sets). It uses these algorithms: RF (Random Forest), NB (Naïve Bayes), kNN (k-Nearest Neighbours), GB (Gradient Boosting), DT (Decision Trees), and SVM (Support Vector Machine).

3.3.3 Data Augmentation

This step divides the datasets into two as train and test. It then applies data augmentation for the required classes using resampling and SMOTE methods.

3.3.4 Augmentated and Aggregated Fingerprint Method Evaluation

This step contains these 4 steps:

1. Using the IoTDevID technique, Aalto University results with augmentation and aggregation enable machine learning (RF) application for an augmented version of the dataset based on each packet level. It then uses the packet aggregation approach to give results for four different group sizes.
2. Using IoTDevID approach, machine learning (RF) application for enhanced version of IoTfinder dataset based on individual packet level is made possible by IoTfinder results with augmentation and aggregation. It then uses the packet aggregation approach to give results for four different group sizes.
3. Using the IoTDevID approach, UNSW_benign_ findings with augmentation and aggregation create an augmented version of the UNSW-Sydney IEEE TMC dataset for machine learning (RF) applications based on individual packet level. It then uses the packet aggregation approach to get results for 4 distinct group sizes.
4. Using the IoTDevID approach, Aalto University results with integrated labels enable machine learning (RF) application for an enriched version of the dataset based on individual packet level. The packet aggregation approach is then used to generate results for four distinct group sizes. Nonetheless, in this file, very similar devices are gathered under the same label and treated as a group in the Aalto University dataset.

3.3.5 Malicious Device Dataset Evaluation

The UNSW_Malicious_ findings with augmentation and aggregation are contained in this phase. This enables the use of the IoTDevID approach to base machine learning (RF) application for UNSW-Sydney ACM SOSR and IoT Network Intrusion datasets at the individual packet level. The packet aggregation approach is then used to generate results for four distinct group sizes. But in contrast to previous steps, this one includes both harmful and benign data generated by the same devices. The idea is not to stop these attacks, but rather to demonstrate that the device may be identified if it exhibits unusual behavior. As such, not all information from malicious datasets is utilized. Only instances where IoT devices are attacked are included in the data. We separated the malicious and benign parts of the pcap files before extracting the fingerprints for this operation. The datasets website makes apparent all the information needed for the filtering process [11]. Wireshark can be used for these activities.

4. Result and Discussion

In this section, we use dimensionality reduction and machine learning approaches to suggest an effective solution for fingerprint IoT devices. The following succinctly describes the contributions of our suggested solution.

- We describe an accurate fingerprinting technique for Internet of Things devices.
- We present various feature extraction methods and demonstrate how they affect the IoT device's fingerprinting.
- We use a real IoT dataset that includes the pcap traffic of an IoT network to verify the efficacy of our solutions.

Assume for the moment that a number of IoT devices are connected to an IoT network. Our goal is to identify (fingerprint) the identities of these IoT devices by analyzing the traffic that they generate, such as the device's name. The massive amount of network traffic generated by IoT devices is well acknowledged. As a result, our main goal is to determine the bare minimum of characteristics required to accurately fingerprint IoT devices. In order to do this, we test a variety of dimensionality reduction strategies, including PCA and autoencoders, in an effort to identify the ideal collection of characteristics for IoT device fingerprinting.

4.1 Data Collection

Here, the data was gathered using two datasets that are accessible to the public [11]. Real device data is present in both. First, there are 31 devices in the Aalto dataset. Secondly, network traffic statistics from the UNSW dataset [2] can be utilized to model IoT communication patterns. The sole common use for IP addresses is in network intrusion detection. The modelling approach was developed using the smaller Aalto data set, then we used the bigger UNSW data set to assess its broader generality. The data also showed a small imbalance caused by different kinds of IoT devices generating traffic based on their functions. Out of the thirty-one IoT devices in the latter, twenty-seven fall into different categories (two devices each kind). The different types of devices are listed in Table 4.1, and the dataset schema and number of instances are shown in Table 4.2.

Table 4.1 List of IoT Devices of the dataset used

No.	Device	No.	Device	No.	Device
1.	Aria	10.	EdimaxPlug2101W	19.	TP-LinkPlugHS110
2.	D-LinkCam	11.	EdnetCam	20.	WeMoInsightSwitch
3.	D-LinkDayCam	12.	EdnetGateway	21.	WeMoLink
4.	D-LinkDoorSensor	13.	HomeMaticPlug	22.	WeMoSwitch
5.	D-LinkHomeHub	14.	HueBridge	23.	Withings
6.	D-LinkSensor	15.	HueSwitch	24	EdimaxPlug1101W
7.	D-LinkSiren	16.	iKettle	25	TP-LinkPlugHS100
8.	D-LinkSwitch	17.	Lightify	26	EdimaxCam
9.	D-LinkWaterSensor	18.	MAXGateway	27.	SmarterCoffee

Table 4.2 Dataset Schema

Dataset	Number of Devices	Type
Aalto University	31	Benign
UNSW-Sydney IEEE TMC	31	Benign
IoTFinder	51	Benign
UNSW-Sydney ACM SOSR	28	Benign & Malicious
IoT Network Intrusion Dataset	2	Benign & Malicious

4.2 Data Preprocessing

Subsequently, the system examines the information it has collected and extracts relevant characteristics. These collected attributes are the basis for creating device fingerprints from network packets. Three categories of approaches exist: individual, aggregated, and mixed. Individual packets are based on distinct identifiers, like MAC or IP addresses, to differentiate between devices. However, the ability of this technology to discriminate between various devices and behaviors might have certain limitations. To overcome this, an aggregated technique is applied, which first separates packet ML labels before merging and then groups packets with the same identification property. By rejecting packets with inaccurate labeling, this technique aims to increase accuracy. The hybrid approach handles issues with the aggregating process and employs unique labels for exceptions. This method is recommended for handling scenarios in which transfer issues affect MAC addresses and cause merging issues. Therefore, although these fingerprinting techniques provide a novel way to classify network packets, careful consideration of their implications in different network scenarios is needed to achieve reliable device detection. The label encoding, data shuffle, data normalization, and data cleaning processes are collectively referred to here as “data preprocessing.”

4.3 Data Cleaning

Real-world scenarios cannot yield a flawless dataset due to human error, technological constraints, and errors occurred during data gathering or collecting. Because of this, if any data does not fit the pattern or is considered unnecessary, we remove it from the dataset during the data cleaning process. To clean up the dataset that is being analyzed, the following steps are taken:

- Eliminating rows when there are duplicate or missing values.
- Elimination of inconsistent values—that is, instances that don't match features—based on the type of feature data.

4.4 Feature Selection

During the feature reduction phase, features that are less important or needless are removed. Reducing the number of features ensures a lower processing cost for the learning model and sometimes improves the model's accuracy. This is the aim of feature selection. A statistically based feature selection method that assesses the correlation between attributes and labels is called chi-squared feature selection. Then, it chooses the attributes having the highest association with the label features using a genetic algorithm.

4.5 Feature Extraction

We retrieved features from the packet headers of pcap (packet capture) files. We first separated the pcap data into training and test sets in order to completely isolate these sets. In the Aalto dataset, every device has 20 sessions (pcap files). They were divided into two groups (80%:20%): four sections of test data and sixteen sections of training data. Each day's items in the UNSW dataset are generated into a different pcap file. We used data collected on different days to construct training and test data (pcap files) in order to isolate. Following the extraction of our first features, we used the xverse (pypi.org/project/xverse/) package and a feature-importance-based voting method to remove features that were deemed unnecessary. This method uses six different techniques to establish each device's feature relevance ratings. It then uses voting to determine which features (features and vote rates) should be included. The chi-square best variables, variable importance using extra trees classifier, variable importance using RF, information value using the weight of evidence, and L1-based feature selection are the six scoring methodologies used by this method. We removed the 26 features from our feature pool that received no votes on any device using any of these six methods.

4.6 ML Algorithms

The created prediction model and, in turn, the classification efficiency are significantly impacted by the quantity of the dataset and the ratio of training to testing and validation in the machine learning process. The training/testing ratio of 80/20, which uses 80% of the dataset for model training, 20% for testing, and 20% for model validation, is taken into consideration in this work. A variety of machine learning approaches use the dimensionality-reduced dataset as their training set. The selection of algorithms was based solely on supervised learning machine learning. Using the Python SKLearn module, the recovered components of the dimensionality methods were split 60/20/20 percent for training, testing, and validation. The ML algorithms are as follows: Random Forest (RF), Decision Tree (DT), K-Nearest Neighbors (KNN), Super Vector Machine (SVM), and Naive Bayes (NBC).

Table 4.3 makes it clear which algorithms perform best in terms of DI: RF, DT, and GB. The fastest inference times are achieved by DT and NB, however NB's precision is not very good. Although kNN and GB attain relatively high accuracy, their slowness renders them unsuitable for practical application. In terms of speed and accuracy, the SVM algorithm is also impractical. As seen in Figures 4.1 to 4.4, kNN, GB, and SVM are especially harmful due to their extraordinarily lengthy inference durations.

Table 4.3 Comparison of ML algorithms with average and standard deviation (SD) of 100 repeats on the Aalto dataset. t is time, in seconds

ML	Accuracy	Precision	Recall	F1score	Train-t	Test-t
DT	0.705	0.774	0.706	0.727	0.128	0.004
GB	0.699	0.789	0.693	0.725	918.3	8.312
kNN	0.705	0.752	0.705	0.718	0.005	20.20
NB	0.617	0.584	0.629	0.559	0.433	0.032
RF	0.708	0.768	0.708	0.727	3.742	0.333
SVM	0.680	0.697	0.634	0.649	101.3	64.80

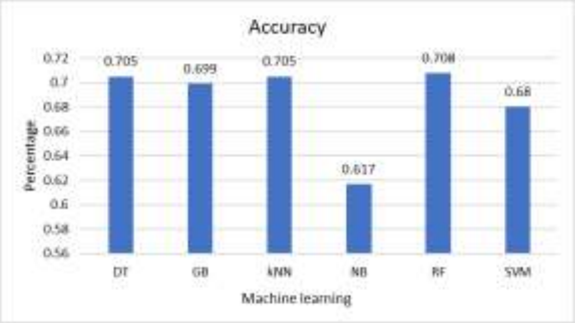


Figure 4.1 Machine learning vs percentage for accuracy

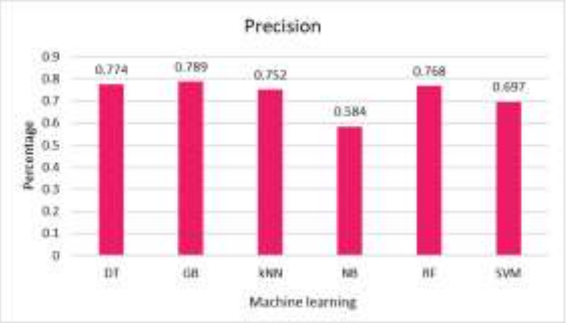


Figure 4.2 Machine learning vs percentage for precision

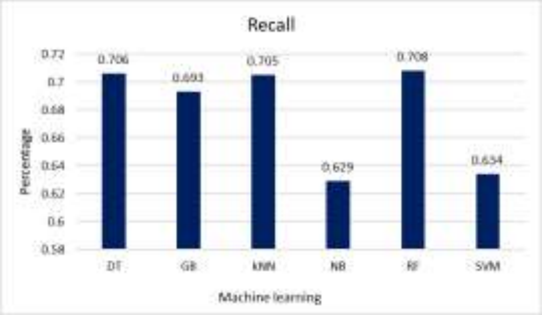


Figure 4.3 Machine learning vs percentage for recall

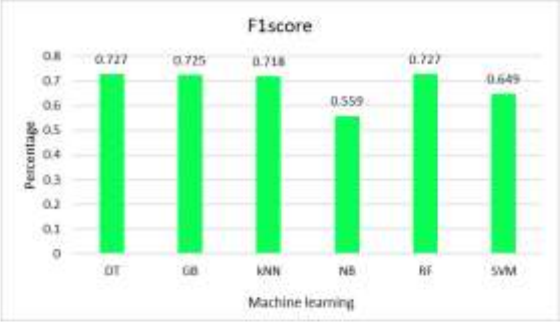


Figure 4.4 Machine learning vs percentage for F1 score

For a real-time device detection system operating in a millisecond-level processing environment, like network traffic, inference times expressed in seconds are not acceptable. Based on this facts, we use DT in the remaining work because it offers the best trade-off between speed and accuracy.

4.7 Performance Analysis

A comparison of the models obtained for every ML algorithm and feature extraction approach previously discussed yields the top performing trained models. The machine learning algorithm and feature extraction technique that simultaneously yield the fewest features and the highest accuracy are the best. The performance metrics that are used are F1-Score, recall, and precision. We verify our approach on a widely used dataset of IoT network traffic statistics. The pcap files from this dataset are extracted using Zeek and concatenated into a single CSV file. After applying dimensionality reduction, a new collection of features is created, and these are taught using machine learning methods.

We evaluate our proposed approach and plan by testing the trained models and computing several prediction quality metrics, such as Precision, Recall, and F1-Score.

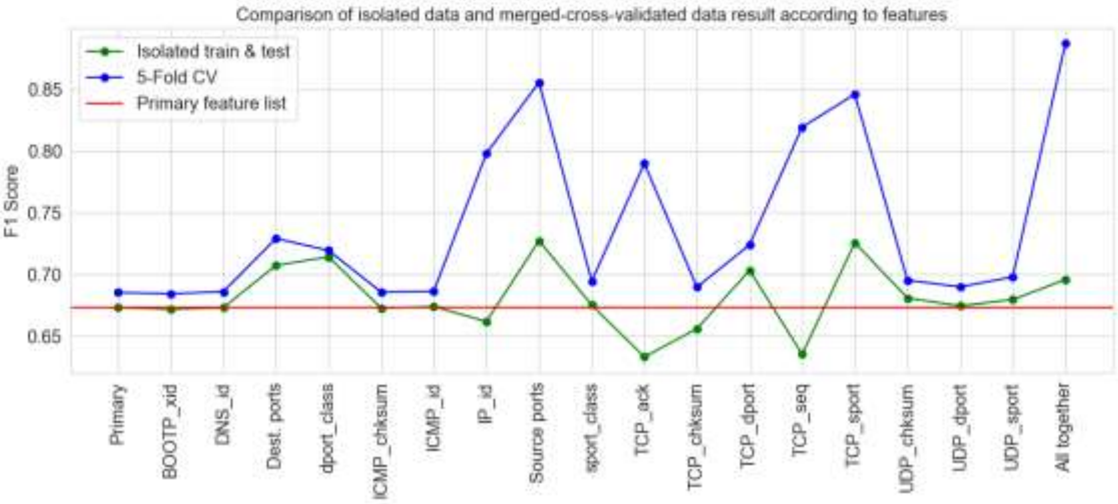


Figure 4.5 Comparison of isolated data and merged-cross-validated data result according to features

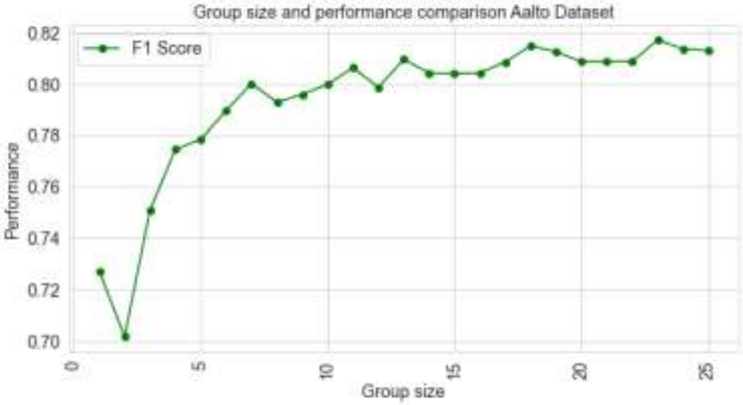


Figure 4.6 Group size vs performance relationship

The connection between performance and total group size, It seems that the plateau is at g = 13.

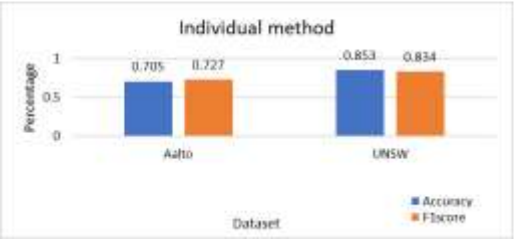
4.8 Performance evaluation

The findings are shown for three different iterations of the method: individual, aggregated, and mixed, depending on the model that was chosen. First, we looked into how the context of the aggregation technique affected the group size vs. performance connection. The positive correlation, as seen in Figure 4.6, implies that larger groups are more productive. Even in cases where a large number of IoT devices connect on a regular basis, larger group sizes may not always be achievable. In light of this, a group size of 13 was selected, which corresponds to the approximate moment at which performance begins to plateau.

Table 4.4 Results and their SD obtained using Individual, Aggregated and Mixed approaches on Aalto and UNSW datasets. t is time in seconds.

Method	Dataset	Accuracy	F1score	Test-t
Individual	Aalto	0.705	0.727	0.004
	UNSW	0.853	0.834	0.008
Aggregated	Aalto	0.745	0.809	0.007
	UNSW	0.943	0.937	0.017
Mixed	Aalto	0.833	0.861	0.008
	UNSW	0.941	0.935	0.022

Table 4.4 shows that the individual packet technique produced an F1 score of 73% and 83%, respectively, in the UNSW and Aalto datasets. However, it was discovered that aggregation greatly increased the ability of both datasets to correctly identify devices. On the Aalto dataset, the overall F1 score rose to 81%, while on the UNSW dataset, it nearly reached 94%. Considering that the UNSW dataset's feature extraction was restricted to the Aalto dataset, it is really fascinating to observe the great level of discriminating displayed in Figures 4.7 to 4.9.



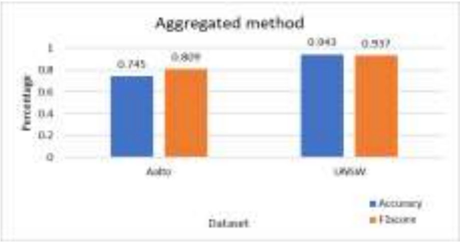


Figure 4.7 Dataset vs percentage for Individual method

Figure 4.8 Dataset vs percentage for Aggregated method

The average separates the device-level performance of the DT models for the Aalto Dataset. Because the data set is considerably imbalanced, we needed to select a statistic that is somewhat insensitive to class size imbalance in order to present a relevant picture of the model's performance at the device level. For this reason, we picked the F1score. Other research have used accuracy for this, but this is an inadequate criterion for imbalanced datasets. Upon analyzing the device-based results from the Aalto dataset, we see that almost all devices experience positive effects from the aggregation process,

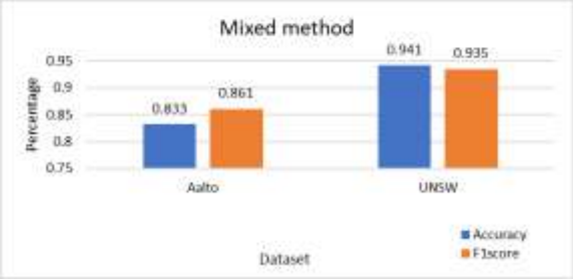


Figure 4.9 Dataset vs percentage for Mixed method

while just four devices experience negative effects. This is a result of the four devices' pairs being affected by a transfer issue. To solve the problem, we used a hybrid strategy that included an exception to our aggregate method. After applying the mixed technique, the overall F1score—which was 81% in the Aalto dataset—rose to 86%. Because the UNSW dataset does not have a transfer problem, the results of this method are not significantly affected by the dataset.

4.9 COMPARISON WITH PREVIOUS WORK

In Table 4.5, we present a comparison of the study's general findings with previously released data. Even while our study doesn't initially appear to yield superior numerical results than earlier research, there are a number of reasons why this comparison is unfair. To begin with, many techniques rely on very specific qualities that are unlikely to be relevant for fresh data sets. result of data inadvertently spilling into the training set from the test set, which occurs frequently

Furthermore, because IoT device datasets inherently have uneven distributions, it can be highly misleading to utilize accuracy numbers in conjunction with imbalanced datasets. Thirdly, different datasets have been employed; some studies only present findings for subsets of the datasets. On the other hand, we made great care to make sure that information leakage was prevented, dataset-specific features were eliminated, appropriate metrics were employed, and all results were presented. Using an additional dataset that we intentionally employed, we assessed the generality of the technique. Because of this, we believe that our set of objective results is more likely to imply probable success on unknown data than results from other studies.

Table 4.5 Comparing the result of IoTDevID with former studies

Study	Dataset	Result	Metric	Feature Types
[1]	Aalto	82%	Accuracy	Packet header
[2]	UNSW	99.98%	Accuracy	Packet header & Flow
[4]	Private	99%	Accuracy	Packet header & Payload
[18]	Aalto	81.50%	Accuracy	Packet header
[25]	Aalto	90.30%	F1 score	Flow statistics
Our Study	Aalto	83.30%	Accuracy	Packet header & Payload
	UNSW	94.30%		
	Aalto	86.10%	F1 score	Packet header & Payload
	UNSW	93.70%		

As from Table 4.5 the accuracy of the previous work after selected Aalto data set is 81.50 % by using Packet header type Feature selected where in our analysis accuracy is 83.30% for the same dataset and Feature selection. But our focus on F1 score by using Packet header & Payload Feature selection both the dataset gets high score as Aalto 86.10% and UNSW 93.70%.

5. Conclusions

In this study, we describe an ML-based method for identifying IoT devices. Finding and removing rogue devices and understanding the security vulnerabilities in a network as a whole depend on being able to identify the IoT devices that are linked to it. ML has been attempted to be used to this process in a few previous research. Nevertheless, we have identified several factors that may limit the relevance of their results, including the use of inappropriate measures and session-based distinguishing characteristics. To develop a robust DI method that reliably extends beyond the training set of data, we have attempted to tackle this process in this study in an open and methodical way. From an ML perspective, we found that decision trees offer the best trade-off between prediction performance and inference time—the latter being essential for putting into practice a model to continually monitor network traffic. In the future, we plan to develop an SDN-based network management system that evaluates the outputs of intrusion detection and device identification, along with an IDS that will work in conjunction with IoTDevID to detect network risks. Thus, the objective is to develop an IoT security system that is workable, devoid of potential weaknesses, and able to repel attacks.

According to prior research, the accuracy of the Aalto data set that was chosen after applying the packet header type feature selection was 81.50%, but in our analysis, the accuracy was 83.30% for the same dataset and feature selection. However, we concentrate on the F1 score by selecting the packet header and payload features, which results in high scores for both datasets (Aalto 86.10% and UNSW 93.70%).

References

[1] A. Aksoy and M. H. Gunes, "Automated IoT device identification using network traffic," in ICC 2019 IEEE, pp. 1-7, 2019.

[2] A. Sivanathan, H. H. Gharakheili, F. Loi, A. Radford, C. Wijenayake, A. Vishwanath, and V. Sivaraman, "Classifying IoT devices in smart environments using network traffic characteristics," IEEE-TMC, 2018.

[3] Ankur Utsav, Amit Abhishek, Kamal Kant, Ritesh Kr.Badhai" Unique Identification for Monitoring of COVID-19 Using the Internet of Things (IoT)" IEEE Xplore. on December 18, 2020.

[4] B. Bezawada, M. Bachani, J. Peterson, H. Shirazi, I. Ray, and I. Ray, "Behavioral fingerprinting of iot devices," in Proceedings of the 2018 Workshop on Attacks and Solutions in Hardware Sec., pp. 41-50, 2018.

[5] Bastida, D. and Lin, F., Enhancing Cloud-Based IoT/M2M System Scalability by Dynamic Network Slicing. Communications and Network, Vol.12, pp.122-154, 2020, doi: 10.4236/cn.2020.123007.

[6] D. D. N. Nguyen, K. Sood, M. R. Nosouhi, Y. Xiang, L. Gao, and L. Chi, "RF fingerprinting based IOT node authentication using mahalanobis distance correlation theory," IEEE Networking Letters, 2022.

[7] E. A. Shammar and A. T. Zahary, "The internet of things (iot): a survey of techniques, operating systems, and trends," *Library Hi Tech*, 2019.

[8] Enjoy Algorithms, "Supervised, Unsupervised, And Semi-Supervised Learning with Real-Life Usecase," https://www.enjoyalgorithms.com/blogs/ supervised-unsupervised-and-semisupervised-learning, accessed on July 07, 2023.

[9] H. Sundmaeker, P. Guillemin, P. Friess, and S. Woelffl´e, "Vision and challenges for realising the internet of things," *Cluster of European research projects on the internet of things, European Commision*, vol. 3, no. 3, pp. 34-36, 2010.

[10] Haghnegahdar, L., Joshi, S. S., & Dahotre, N. B., From IoT-based cloud manufacturing approach to intelligent additive manufacturing: Industrial Internet of Things-An overview. The International Journal of Advanced Manufacturing Technology, pp.1-18, 2022.

[11] Datset available on https://research.aalto.fi/en/datasets/iot-devices-captures

[12] I. Andrea, C. Chrysostomou, and G. Hadjichristofi. Internet of things: Security vulnerabilities and challenges. In Proc. of 2015 IEEE Symposium on Computers and Communication (ISCC), July 2015.

[13] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," SN Computer Science, vol. 2, no. 3, pp. 1-21, 2021.

[14] J. Brownlee, Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python. Machine Learning Mastery, 2020.

[15] J. Tan and S. G. M. Koo. A survey of technologies in internet of things. In Proc. of 2014 IEEE International Conference on Distributed Computing in Sensor Systems, May 2014.

[16] Kahraman Kostas, Mike Just, and Michael A. Lones. IoTDevID: A behaviour-based finger printing method for device identification in the IoT, arXiv preprint, arxiv:2102.08866, 2021.

[17] L. Roselli, C. Mariotti, P. Mezzanotte, F. Alimenti, G. Orecchini, M. Virili, and N. B. Carvalho. Review of the present technologies concurrently contributing to the implementation of the internet of things (IoT) paradigm: RFID, green electronics, wpt and energy harvesting. In Proc. of 2015 IEEE Topical Conference on Wireless Sensors and Sensor Networks (WiSNet), January 2015.

[18] 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 37th Int. Conf. DCS. IEEE, 2017.

[19] M. Wang et al., "Cellular machine-type communications: Physical challenges and solutions,"' IEEE Wireless Commun., vol. 23, no. 2, pp. 126-135, Apr. 2016.

[20] Mainetti, L.; et al., A Software Architecture Enabling the Web of Things, IEEE Journal of Internet of Things. Vol. 2, no. 6, pp. 445-454, 2015.

[21] Marchal Samuel, "IoT devices caputers," https://research.aalto.fi/en/datasets/ iot-devices-captures, accessed on September 06, 2021.

[22] Pradeep Bhanot, "Top 5 IoT applications for IT," https://blazent.com/ top-5-iot-applications/, accessed on May, 2022.

[23] Q. Yang, Y. Liu, W. Yu, D. An, X. Yang, and J. Lin. On data integrity attacks against optimal power flow in power grid systems. In Proc. Of Annual IEEE Consumer Communications and Networking Conference (CCNC), 2017.

[24] R. Osei, L. Habib, M. Malek, and Z. Zhongwen, "Efficient iot device fingerprint- ing approach using machine learning," In Proceedings of the 19th International Conference on Security and Cryptography, pp. 525-533, 2022.

[25] S. A. Hamad, W. E. Zhang, Q. Z. Sheng, and S. Nepal, "IoT device identification via network-flow based fingerprinting and learning," in 18th IEEE TrustCom. IEEE, pp. 103-111, 2019.

[26] S. Zhang, Z. Wang, J. Yang, D. Bai, F. Li, Z. Li, J. Wu, and X. Liu, "Unsu- pervised iot fingerprinting method via variational auto-encoder and k-means," in ICC 2021-IEEE International Conference on Communications. IEEE, 2021, pp. 1-6, 2021.

[27] Sachin Kumar, Prayag Tiwari and Mikhail Zymbler, "Internet of Things is a revolutionary approach for future technology enhancement: a review" J Big Data journal Vol. 6, pp.111, 2019, https://doi.org/10.1186/s40537-019-0268.

[28] Salama, R., Al-Turjman, F., Chaudhary, P., & Yadav, S. P., Benefits of Internet of Things (IoT) Applications in Health care-An Overview, In 2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN) IEEE, pp. 778-784, 2023.

[29] T. Z. Project, "Zeek an open source network security monitoring tool," 2020.

[30] Ta-Wei Yang, Yu-Han Ho, Cheng-Fu Chou, Achieving M2M-device authentication through heterogeneous information bound with USIM card, Future Generation Computer Systems, Vol.110, pp. 629-637, 2020, ISSN 0167-739X, https://doi.org/10.1016/j.future.2019.10.042.

[31] V. Gustavsson, "Machine learning for a network-based intrusion detection system: An application using zeek and the cicids2017 dataset," 2019.

[32] Wang, J.; et al., A distributed algorithm for inter-layer network coding-based multimedia multicast in Internet of Things. Elsevier Computers & Electrical Engineering. Vol. 52, pp. 125-137, 2016.

[33] Yunchuan, S.; Song, H., and Jara, A, Internet of Things and Big Data Analytics for Smart and Connected Communities, IEEE Access. Vol. 4, pp. 766 -773, 2016.