Machine Learning based Classification and Identification of IoT devices

A M. Tech Project Report Submitted
in Partial Fulfillment of the Requirements
for the Degree of

Master of Technology

by

Param Bharatbhai Patel & Pratik Mahendra Kamble
(234101062 & 234101039)

under the guidance of

Dr. Tamarapalli Venkatesh

to the

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
GUWAHATI - 781039, ASSAM

CERTIFICATE

*This is to certify that the work contained in this thesis entitled "**Machine Learning based Classification and Identification of IoT devices**" is a bonafide work of **Param Bharatbhai Patel & Pratik Mahendra Kamble** (**Roll No. 234101062 & 234101039**), carried out in the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati under my supervision and that it has not been submitted elsewhere for a degree.*

Supervisor: **Dr. Tamarapalli Venkatesh**

Professor,

Nov, 2024 Department of Computer Science &

Engineering,

Guwahati. Indian Institute of Technology Guwahati, Assam.

Acknowledgements

We would like to express our deepest gratitude to our thesis supervisor, Prof. Tamarapalli Venkatesh, from the Department of Computer Science and Engineering at IIT Guwahati. His constant support, insightful guidance, and encouragement have been instrumental in shaping this project. Working under his mentorship has not only deepened our understanding but also inspired us to push our boundaries. The lessons we've learned from him will undoubtedly guide us in the years to come.

We are also incredibly thankful to our parents and friends for their unwavering support and motivation throughout this journey. Their belief in us and their constant encouragement have been our driving force. Finally, a heartfelt thank you to Prof. Venkatesh once again, whose dedication and guidance have made this project possible.

Contents

Chapter 1

Introduction

The Internet of Things (IoT) represents a transformative network of physical objects or devices embedded with sensors, software, and other technologies. This network facilitates seamless connectivity and communication among devices and systems, enabling automation, monitoring, and intelligent decision-making. Through local networks or the internet, IoT devices exchange data to enhance user experiences and operational efficiency.

IoT devices :

IoT (Internet of Things) devices are physical objects embedded with sensors, software, and other technologies to connect and exchange data with other devices and systems over the internet. These devices range from household appliances like smart bulbs, thermostats, and security cameras to industrial tools, medical monitors, and environmental sensors. They are designed to enhance automation, monitoring, and data-driven decision-making in smart environments like homes, cities, and industrial facilities.

Classification of IoT devices :

The classification of IoT devices involves categorizing them based on their characteristics, behavior, or functionality to enhance network management and security. This is typically achieved using network traffic patterns, such as packet size, activity cycles, port usage, and signaling protocols, or by analyzing device behavior like communication intervals and sleep cycles[SSG+17]. Machine learning techniques, such as Random Forests or Decision

Trees, are commonly employed to identify patterns and classify devices accurately. IoT devices can also be grouped by roles, including smart home appliances, industrial sensors, and medical monitors.

Identification of IoT devices :

The identification of IoT devices refers to the process of uniquely recognizing and distinguishing devices within a network based on their characteristics, behavior, or communication patterns. It ensures that each device can be accurately monitored and managed, addressing critical security and operational challenges. Identification often involves techniques like device fingerprinting, which uses features such as MAC addresses, communication protocols, packet size, and timing patterns to create unique profiles for devices. Advanced methods leverage machine learning to analyze network traffic and identify devices by extracting meaningful features from packet headers or flow behaviors[KJL22].

Importance of IoT device Identification :

The identification of IoT devices is crucial for ensuring network security, efficient management, and seamless operation, as it enables the unique recognition of devices to detect and isolate rogue or unauthorized entities, preventing security breaches. It also supports behavioral monitoring for anomaly detection, identifying deviations like unusual traffic patterns, unauthorized access, or protocol changes, which may indicate security threats or device malfunctions. Anomaly detection further facilitates timely intervention to address issues, leveraging lightweight and efficient methods, such as machine learning, suitable for resource-constrained IoT devices. Accurate identification and anomaly detection are particularly vital in heterogeneous IoT networks, where devices from diverse vendors and protocols coexist, enabling vulnerability management, regulatory compliance, and enhanced reliability while protecting against evolving cyber threats.

Challenges in IoT device identification :

- **Device Heterogeneity:** The wide variety of IoT devices with diverse functionalities complicates the development of universal identification models.

- **Dynamic Network Environments:** IoT devices operate in varied network settings, requiring adaptive identification methods that generalize across different environments.

- **Scalability:** The rapid growth of IoT devices demands solutions that scale effectively while maintaining high accuracy.

1.1 Motivation

The proliferating integration of Internet of Things (IoT) devices into daily life and industry has enhanced efficiency but introduced critical security and operational challenges. Accurate identification of IoT devices is vital to mitigate these risks. It enables the swift containment of security breaches by isolating compromised devices, such as those infected with malware or part of botnets, reducing damage and preventing further attacks. Device identification also prevents spoofing by verifying legitimacy, reducing risks of impersonation. Moreover, understanding device behavior aids in anomaly detection, such as identifying irregular activities that may signal breaches. It further supports network traffic analysis to detect suspicious patterns like DDoS attacks or unauthorized data transfers. These capabilities make IoT device identification essential for securing networks and fostering trust in IoT ecosystems.

1.2 Objective

To develop or refine a Machine Learning-based approach for the classification and identification of IoT devices, emphasizing the creation of a generalizable method that performs reliably across diverse environments and network contexts.

Chapter 2

Literature Survey

2.1 Overview of the IoT device Identification

The identification of IoT devices is a critical task in securing and managing increasingly heterogeneous networks. With the widespread adoption of IoT devices in homes, enterprises, and cities, the ability to classify and monitor devices based on their network behavior has become essential for mitigating security risks, preventing unauthorized access, and ensuring proper functionality. This field has evolved significantly over the years, transitioning from simple rule-based and static methods to advanced machine learning-based systems capable of identifying devices in real time.

IoT device identification presents unique challenges. The diversity of device types, communication protocols, and resource constraints complicates the creation of unified, scalable solutions. Traditional approaches relied on static identifiers like MAC addresses or environmental attributes, such as DNS queries and port numbers, which often failed to generalize across diverse settings. Moreover, the use of non-IP and low-energy protocols like ZigBee and Bluetooth in IoT ecosystems has necessitated more adaptable and protocol-independent methods.

Machine learning has emerged as a powerful tool in addressing these challenges. By leveraging network traffic features, researchers have developed systems that classify IoT devices based on behavioral patterns. Recent methods emphasize the importance of robust feature extraction, selection, and classification techniques to enhance the accuracy and generalizability of these systems. This literature survey delves into the methods and techniques proposed in the field, categorized across feature engineering, model selection, and performance evaluation, while highlighting the advancements and limitations of key approaches.

2.2 IoT Device Data

An IoT device dataset is a structured collection of data that captures the network behaviors, communication patterns, and physical characteristics of IoT devices. These datasets typically include raw network traffic (.pcap files), extracted features such as packet size, inter-arrival times, and protocol usage, along with behavioral metrics like activity cycles[SSG$^+$17], signaling patterns, and cipher suites[SGL$^+$18]. They are often annotated with labels indicating device types, vendors, or instances, making them suitable for supervised learning tasks. Metadata, including the deployment environment and device configurations, may also be included to provide context. Such datasets are collected through passive monitoring of network traffic, controlled experiments in testbeds, simulated scenarios, or crowdsourced data from IoT gateways, ensuring coverage of diverse devices and settings.

2.3 Traditional Identification methods and Adaption of Machine Learning

Traditional identification methods for IoT devices primarily relied on static identifiers such as MAC addresses, DNS queries, and port numbers, as well as rule-based systems leveraging predefined traffic patterns or protocol-specific attributes[SSG$^+$17]. These methods were computationally simple and effective in controlled environments but faced significant limitations in dynamic, large-scale IoT networks. They struggled with issues like spoofed identifiers, evolving device behaviors, and handling non-IP or low-energy protocols such as ZigBee and Bluetooth. To overcome these challenges, the adoption of machine learning (ML) has introduced behavior-based approaches, which analyze network traffic patterns, device activities, and protocol-independent features. ML-driven methods provide adaptability, scalability, and improved accuracy in diverse and dynamic IoT ecosystems, paving the way for generalizable solutions that address the shortcomings of traditional approaches.

2.4 Feature Extraction Methodologies

Feature extraction is a critical process in IoT device identification, where raw network traffic data is analyzed and transformed into structured attributes that characterize the behavior and communication patterns of devices. These features serve as the foundation for machine learning models and other analytical methods, enabling systems to classify, identify, and monitor IoT devices effectively. Key attributes typically include packet size, inter-arrival times, protocol usage, flow volume, and signaling events like DNS queries or cipher suites. Such features encapsulate the dynamic and often unique communication behaviors of IoT devices, allowing systems to distinguish between device types and even individual devices.

The choice of features plays a pivotal role in determining the accuracy and generalizability of identification systems. Well-selected features ensure that the system can accurately recognize devices under varying network conditions and configurations, making it robust against environmental changes. Conversely, poorly chosen or overly specific features, such as those tied to particular protocols or network infrastructures, can hinder the system's ability to generalize across diverse IoT environments. Therefore, feature extraction must strike a balance between capturing enough device-specific details and maintaining adaptability, ensuring the system remains effective in heterogeneous and dynamic IoT ecosystems.

2.4.1 Statistical Traffic Analysis

Sivanathan et al. conducted a detailed study by collecting long-term network traffic from 28 IoT devices over six months, extracting features like flow volume, duration, average rate, and signaling events (e.g., DNS queries and cipher suites). These features effectively captured device activity patterns and were useful for classification and identification. However, their applicability is often tied to the specific network environment, as factors like protocol prevalence and infrastructure significantly influence the extracted features. This dependency limits their generalizability to other settings, highlighting the need for more protocol-agnostic approaches to ensure broader applicability across diverse IoT ecosystems.[SGL+18]

2.4.2 Protocol-Based Metrics

IoT Sentinel employs a fingerprinting methodology that leverages 23 protocol-based features extracted from the first 12 packets transmitted during a device's setup phase. These features include attributes such as source and destination IP addresses, packet sizes, and specific protocol usage, including DNS, HTTP, TCP, and others. The extracted data is organized into a 23×12 matrix, where each row represents a feature, and each column corresponds to one of the 12 initial packets. This matrix serves as a unique representation of the device's behavior during its setup, allowing the system to classify and identify devices with high accuracy.While this approach is effective in capturing distinctive characteristics during the setup phase, it faces limitations when applied to ongoing traffic analysis. Since the methodology relies on the initial burst of packets, it does not account for dynamic behavior or long-term traffic patterns that may arise as devices operate over time. Additionally, IoT Sentinel struggles with non-IP-based devices, such as those using low-energy protocols like ZigBee or Bluetooth, as the fingerprinting process depends heavily on IP-based traffic features. These constraints make IoT Sentinel highly suitable for initial device classification but less adaptable for continuous monitoring or environments with heterogeneous protocol requirements.[MMH+17]

2.4.3 Single-Packet Features

SysID simplifies feature extraction by focusing on single-packet headers, making the process computationally efficient and straightforward. By utilizing features such as TCP sequence numbers, port numbers, and IP IDs, it reduces the need for processing large volumes of traffic data or performing complex multi-packet analyses. This streamlined approach is particularly beneficial for resource-constrained environments, where low-latency and lightweight solutions are critical. However, this simplicity comes with trade-offs. By relying solely on single-packet data, it may lack the contextual insights that can be obtained through the analysis of packet sequences or traffic flows. For instance, multi-packet analysis can capture temporal patterns, inter-arrival times, and other behavioral dynamics that are often critical for identifying subtle variations between similar devices or detecting anomalies. While this approach ensures scalability and speed, its focus on isolated packet-level attributes might limit its effectiveness in scenarios requiring deeper behavioral understanding or complex anomaly detection.([AG19])

2.4.4 Behavior-Based Features

It introduced a behavior-centric approach to feature extraction that significantly improves the robustness and adaptability of IoT device identification. Traditional methods often rely on static identifiers such as MAC or IP addresses to recognize devices. However, these identifiers have notable limitations: they can be easily spoofed by malicious actors, are sometimes unavailable in low-energy protocols like ZigBee, Bluetooth, or Z-Wave, and may not generalize well across different environments. Recognizing these challenges, it shifts the focus from static identifiers to dynamic, behavior-based features.It extracts protocol-independent attributes at the packet level, which capture the intrinsic communication patterns and behaviors of devices. These features include packet sizes, inter-arrival times, protocol usage, and other dynamic characteristics that remain consistent even when network conditions or configurations change. This approach not only enhances the system's resilience to environmental changes but also allows it to identify devices operating on non-IP-based protocols, making it applicable to a wider range of IoT ecosystems. By relying on behavior-centric features, IoTDevID ensures a higher degree of accuracy, security, and generalizability compared to traditional identifier-based methods.([KJL22])

There were certain limitaions observed with every feature extraction methodologies.Limitations such as feature extraction methods that rely on environment-specific attributes, such as DNS queries or cipher suites, face challenges in generalization. Additionally, single-packet approaches may miss nuanced behaviors observable only through aggregated traffic patterns.

2.5 Feature Selection Techniques

Feature selection is a crucial step in IoT device identification systems that narrows down the extracted attributes to retain only the most informative ones, ensuring the data used for classification is both relevant and efficient. By eliminating redundant or irrelevant features, feature selection reduces noise in the dataset, leading to improved model accuracy and generalization. Additionally, it enhances computational efficiency by minimizing the complexity of the training process and reducing the resources required for real-time analysis, making it particularly valuable for resource-constrained IoT environments.

Various techniques have been employed in previous studies for the feature selection process. Methods like genetic algorithms (GAs) are used to identify optimal feature subsets by simulating an evolutionary process, ensuring the chosen features maximize classification performance. Recursive Feature Elimination (RFE) and statistical methods rank features based on their contribution to the model's predictive power. Machine learning-based approaches, such as Random Forests or Support Vector Machines (SVMs), incorporate embedded feature selection, prioritizing attributes that improve decision-making during training. These techniques not only improve the performance of IoT identification systems but also ensure robustness by focusing on features that generalize well across different devices and network settings. Feature selection, therefore, plays a key role in building scalable and accurate IoT classification systems.

2.5.1 Heuristic-Based Selection

IoT Sentinel used predefined feature sets, focusing on protocol presence and payload attributes.The selected features include protocol-specific attributes like source/destination IPs, packet sizes, and DNS queries, focusing on the first 12 packets exchanged by devices.It is effective in controlled environment but its dependence on predefined, protocol-based features fails to adapt to diverse networks.[MMH+17]

2.5.2 Rigorous Manual Selection

IoTDevID adopted a systematic approach to analyze the relevance of features, focusing on removing environment-dependent attributes that could hinder generalizability across diverse datasets. Features such as port numbers, session IDs, and IP counts, while potentially informative in specific environments, were excluded as they are heavily influenced by the network setup and device interactions unique to a particular dataset. By eliminating these environment-specific attributes, IoTDevID ensured that the selected features represented intrinsic device behaviors, allowing the model to perform robustly even in unseen environments.

In addition to manual feature selection, IoTDevID incorporated advanced optimization techniques, such as Genetic Algorithms (GAs), to further refine the feature set. GAs simulate an evolutionary process, iteratively evaluating combinations of features to identify subsets that maximize classification performance. This automated refinement process complements manual selection by systematically reducing redundancy and retaining only the most predictive features. By combining manual expertise with algorithmic optimization, IoTDevID achieved a lean yet highly effective feature set, enhancing the model's accuracy, scalability, and adaptability in heterogeneous IoT ecosystems. ([KJL22])

2.5.3 Genetic Algorithm (GA)

SysID automated feature selection using GA, which evaluated subsets of features based on their contribution to classification accuracy and computational efficiency.

$$\text{Fitness} = 0.9 \times \text{Accuracy} + 0.1 \times \left(1 - \frac{|\text{SelectedFeatures}| - 1}{|\text{AllFeatures}| - 1}\right)$$

The fitness function in GA balanced accuracy (90%) and feature reduction (10%), resulting in a streamlined feature set that optimized system performance.([AG19])

There were several limitations and trade-offs observed in the previous studies such as while GA automates feature selection, its computational overhead can be significant for large datasets. Similarly, manual selection requires domain expertise, which may not scale effectively for highly diverse datasets.Therefore some studies[KJL22] first does manual selection then uses GA on the smaller subset to reduce it further.

2.6 Model Selection and Classification Methods

Classification is a critical step in IoT device identification, where selected features are mapped to device labels using machine learning models. This process assigns each device to a specific category or type based on its extracted behavioral attributes. The choice of classification model is pivotal, as it directly impacts the system's accuracy, scalability, and interpretability. An effective model ensures reliable device identification, handles large-scale deployments efficiently, and provides insights into its decision-making process, which can be valuable for security and network management.

Model selection and the choice of classification methodology are therefore central to the success of an IoT identification system. Different studies have employed a variety of approaches, each tailored to specific requirements or datasets. For instance, supervised learning models like Random Forests and Support Vector Machines (SVMs) have been widely used for their ability to handle complex feature sets and high-dimensional data. Decision Tree-based methods offer transparency and interpretability, making them useful in scenarios where understanding the classification logic is important. Other approaches, such as Neural Networks, provide the capacity to model non-linear relationships and handle large-scale datasets but may trade off interpretability for accuracy. Some studies also employ ensemble methods, combining multiple models to improve overall performance and robustness.[NLY$^+$24]

In addition, researchers have explored hybrid techniques, where machine learning is integrated with domain knowledge to refine classification processes. The methodologies adapted by previous studies demonstrate the trade-offs between computational complexity, accuracy, and adaptability, highlighting the importance of aligning model selection with the specific challenges and goals of IoT device identification. This diversity underscores the need for careful evaluation of classification models to achieve a balance between performance and practical deployment requirements in IoT ecosystems.

2.6.1 Rule-Based Models

Use of interpretable models like Decision Trees (J48), PART, and OneR, which handle high-dimensional data well and provide explainable outputs.As rules are derived from the actual attribute values present in the packets, the algorithms can identify distinctive header characteristics to effectively classify devices. However, their effectiveness can diminish with highly dynamic traffic behaviors.

2.6.2 Wrapper based Model Selection

This approach involves evaluating multiple machine learning algorithms on the selected features to identify the model that performs best in terms of accuracy, generalization, and scalability. The process is systematic and data-driven, ensuring that the chosen algorithm aligns well with the nature of the IoT traffic data and the desired classification outcomes.Also the factor of training time and testing time is used in decision making process.[KJL22]

2.6.3 Multi-Stage Frameworks

([SGL$^+$18]) used a hierarchical approach to first distinguish IoT from non-IoT traffic, followed by specific device classification. In the first stage, classifiers are applied to each nominal attribute separately, using a Bag-of-Words (BoW) representation to encode values.Naive Bayes Multinomial classifier is employed to predict initial labels.The second stage integrates the outputs from the first stage with quantitative features, such as traffic rate and DNS intervals, using a Random Forest classifier.

2.7 Performance Evaluation

Performance evaluation is a critical aspect of assessing the reliability and effectiveness of IoT device identification systems. Metrics such as accuracy, precision, recall, and F1-score are commonly used to determine how well these models perform in identifying and classifying devices. Systems designed for IoT device identification typically achieve high levels of accuracy when tested on controlled datasets,often exceeding 0.95. However, challenges arise when models are deployed in more diverse and dynamic environments, particularly for devices utilizing non-IP protocols or operating in different network conditions.

Generalizability is a key focus in modern IoT identification frameworks. Effective systems validate their models across multiple datasets to ensure consistent performance in varied scenarios. This approach involves selecting features that are independent of specific environments, such as protocol-agnostic attributes, to enhance adaptability. By avoiding features tied to specific session IDs, port numbers, or DNS patterns, models achieve greater robustness and applicability across different network infrastructures[KJL22].

The evaluation process also relies on robust validation techniques. Many frameworks implement cross-validation to test model stability across data splits, ensuring that the results are not overly dependent on a specific dataset partition. Others employ separate training, validation, and testing datasets to prevent data leakage and guarantee reliable performance metrics[KJL22]. These practices help establish confidence in the scalability and adaptability of the models.

Despite these advancements, there are limitations. While achieving high accuracy is a common goal, it often comes at the expense of computational efficiency. Multistage[SGL+18] and ensemble-based models, while effective in improving accuracy, can be resource-intensive, making them less suitable for real-time scenarios or large-scale IoT deployments. Addressing this trade-off between performance and efficiency remains a key area for future research in IoT device identification.

2.8 Conclusion of the Literature Survey

The field of IoT device identification has witnessed significant advancements, evolving from rule-based and static approaches to sophisticated machine learning-driven methods. This literature survey highlights the diverse methodologies employed at each stage of the identification process, including feature extraction, feature selection, model classification, and performance evaluation.

Feature extraction remains a critical step, with approaches ranging from statistical traffic analysis and protocol-specific metrics to behavior-based and single-packet features. While these methods have demonstrated effectiveness, challenges persist in achieving generalization across heterogeneous environments. Feature selection techniques, such as heuristic-based, manual, and genetic algorithm-driven methods, play a crucial role in refining feature sets, but trade-offs between computational efficiency and scalability remain.

In classification, the adoption of rule-based models, wrapper-based model selection, and multi-stage frameworks reflects the field's drive towards improving accuracy and interpretability. Each methodology offers unique advantages, but dynamic traffic behaviors and diverse IoT protocols continue to pose challenges.

Performance evaluation has underscored the importance of accuracy, precision, recall, and generalizability. Cross validation and multi dataset testing have become standard practices to ensure models can adapt to varied environments.

Moving forward, development of more generalizable and robust method with support for non-IP and low energy protocols with high accuracy will be critical.By addressing these limitations, future research can pave the way for more robust, scalable, and efficient IoT device identification frameworks, capable of meeting the demands of increasingly complex and dynamic IoT ecosystems.

Chapter 3

Implementation and Validation of Research work

3.1 Methodology Implementation

3.1.1 Introduction and Objective

The implementation focuses on the implementation and evaluation of [KJL22], a behavior-based IoT device identification framework designed to classify IoT devices using network traffic analysis. It focuses on the use of protocol-independent features, avoiding static identifiers like MAC and IP addresses to ensure generalizability across diverse environments. By analyzing packet-level behaviors, the framework classifies devices accurately, including those using non-IP protocols like ZigBee and Bluetooth.

3.1.2 Dataset explanation

The evaluation relies on two publicly available datasets: the Aalto University IoT Dataset and the UNSW IoT Dataset. These datasets were chosen to ensure a diverse and realistic representation of IoT devices and their network behaviors, allowing for robust testing of the framework's generalizability and accuracy.

Aalto University Dataset :

It includes data from 27 different device types across 31 smart home IoT devices. The dataset contains network traces of devices in different operational states, such as active and idle. Captures data over multiple protocols, including HTTP, DNS, and TCP/IP.

UNSW Dataset :

This dataset contains data from various IoT devices, including both IP and non-IP devices.Covers devices operating under diverse scenarios, capturing both normal and anomalous traffic patterns.Features communication patterns of devices using protocols like ZigBee, Z-Wave (non-IP), as well as IP-based protocols.

3.1.3 Methodology

The implementation involves a systematic approach, divided into several stages such as Feature Extraction,Feature Selection,Algorithm Selection and Performance Evaluation.

Feature Extraction :

This process is desgined to transform raw network data(.pcap) files into structured attributes. The datasets were divided into training and testing sets which are in isolation with each other. Aalto Dataset was split like 16 sessions for training and 4 sessions for testing. UNSW dataset was split according to the data collection on day-wise.This isolation ensures there is no leakage between training and test sets.Features like payload entropy,port class and protocol info were included with those extracted from headers.

Feature Selection :

This process is designed to systematically narrow down the initial feature pool to the subset of that which contains most informative and generalizable attributes.26 features were removed from the initial pool by six feature voting techniques using xverse pyhton package.Session-specific attributes, such as IP IDs and TCP sequence numbers, are excluded for their lack of generalizability, with raw port numbers replaced by behavioral port classes.Genetic Algorithm was then used to further reduce the feature set.The resulting feature set emphasizes protocol-independent characteristics like payload entropy, size, and protocol classes.

Algorithm Selection :

This process is designed to select the best choice of Machine Learning algorithm for device identification. Six classic ML algorithms were evaluated after hyparameter tuning and nested cross-validation. Algorithms evaluated include Random Forest,Decision Tree,kNN,Gradient Boosting,Support Vector Machine,Naive Bayes.These models were trained on features extracted from network traffic and tested for classification accuracy, computational efficiency, and stability across 27 IoT device classes.Decision Tree was chosen as the optimal algorithm due to its balance of high accuracy and low inference time.

Table 3.1 Performance Comparison of Machine Learning Models

ML	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Train Time**	**Test Time**
DT	0.70439	0.77426	0.70549	0.72727	0.12788	0.00411
GB	0.66079	0.65619	0.62928	0.62473	2641.88	7.32185
kNN	0.70517	0.75175	0.70540	0.71801	0.00488	20.2171
NB	0.61736	0.58443	0.62852	0.55874	0.44444	0.43289
RF	0.70856	0.76842	0.70777	0.72727	3.74166	0.33265

Performance Evaluation :

Performance evaluation assesses the model's effectiveness using metrics like accuracy, F1-score, and inference times across multiple evaluation methods.Evaluation was done on three approaches individual packet-based classification, aggregation of packet-level predictions, and a mixed method combining the two.

Individual Method packets : Classifies each packet independently, avoiding merging errors caused by shared MAC/IP addresses.

Aggregated Method packets : Groups packets by MAC address and assigns the most frequent label from ML predictions.

Mixed Method packets : Combines individual and aggregated methods by using individual labels for ambiguous MAC addresses and aggregated labels for the rest, ensuring better accuracy in complex scenarios.

As the dataset is imbalanced a relevant metric for performance evaluation was considered, i.e. F1 Score.A group size of 13 was selected for aggregation algorithm as it provides the perfect balance between performance and practicality.

Evaluation Results :

Individual method packets achieved a F1 score of 0.7282 for Aalto dataset and F1 score of 0.8281 for UNSW dataset.

Aggregated method packets achieved a F1 score of 0.8099 for Aalto dataset and F1 score of 0.9205 for UNSW dataset.

Mixed method packets achieved a F1 score of 0.8551 for Aalto dataset and F1 score of 0.9316 for UNSW dataset.

Challenges Identified :

Data Imbalance : Imbalanced datasets were a significant challenge, as certain device classes had disproportionately fewer samples. This imbalance affected the model's ability to generalize and accurately classify underrepresented devices, especially for non-IP protocols like ZigBee and Bluetooth.

Non-IP Device Traffic : Non-IP devices using protocols like ZigBee and Z-Wave posed a challenge due to limited data availability and differences in protocol structures. This made it difficult to extract and standardize behavioral features for these devices.

3.2 Validation on different Dataset

A further validation of [KJL22]'s methodology was carried out on new dataset.

3.2.1 Dataset explanation

The evaluation relies on the publicly available dataset : CIC IoT 2022 dataset. It is a rich and diverse dataset designed to simulate real-world IoT environments.The dataset consists of traffic data for 40 IoT devices, with IP and non-IP protocols(ZigBee,Z-Wave).It includes normal operations and also attack conditions with both active and idle states of the devices.

Feature Extraction and Labelling :

By help of the tools like WireShark,Scapy and Python, an initial set of about 100 features was generated. From which only the subset of the features that were relevant to original study were retained.Labelling is done by mapping source MAC addresses to device names, organizing packets into sessions (active or idle), and grouping devices of the same brand or model under a single label. Sessions are further categorized using binary session IDs to reflect device presence, addressing challenges like data imbalance and session-specific variations.

Dataset Construction :

The dataset consists of data of two sessions i.e. idle and active. For training and testing subsets, it was divided to create four categories : idle-active,idle-idle,active-idle,active-active. Former being the training and later being the testing. Specific adjustments were also made according to the need of handling the data absence or imbalance.Adjustments include duplicating idle state data into active state data,removing the device due to very low device presence.

Performance Evaluation

The evaluation includes both individual level analysis and aggregated algorithm.Individual packet models achieved F1 scores above 0.81, with the active-idle scenario performing best at 0.911.The aggregation algorithm further improved performance. The evaluation demonstrates the effectiveness of the original paper's methodology. However it highlights the need for addressing the data imbalance.

Chapter 4

Conclusion and Future Work

This study demonstrates the evolution of IoT device identification methods from traditional static approaches to advanced behavior-based and machine learning-driven techniques.The implementation and validation of [KJL22]'s methodology highlight the robustness of protocol-independent features and innovative classification techniques, achieving high accuracy across diverse datasets like Aalto, UNSW, and CIC-IoT-2022. By combining individual packet classification, aggregation, and mixed methods, the framework effectively addresses challenges in identifying devices across varied network environments, including non-IP protocols like ZigBee and Z-Wave.

However certain challenges were identified during the study which include ,

Data Imbalance : Imbalanced datasets poses a major challenge in IoT device identification. As observed in the implementation of [KJL22]'s methodology it affected the model's ability to generalize and accurately classify underrepresented devices, especially for non-IP proto-cols like ZigBee and Bluetooth.

4.1 Future Work Directions

The proposed solution for handling data imbalance is discussed in this section.

4.1.1 Meta Learning with stacking

Base Models(0-level models) :

Random Forest : Provides robust generalization but may slightly favor majority classes.

Gradient Boosting (e.g., XGBoost or LightGBM): Focuses on harder-to-classify instances, naturally benefiting minority classes.

Logistic Regression: A simple model for capturing linear relationships, ensuring lightweight predictions.

Meta Model (1-level models) :

Weighted Logistic Regression or LightGBM : Incorporate class weights to emphasize minority classes. Learn from base model outputs, focusing on combining their strengths for minority class prediction.

4.1.2 Exploration of Additional Datasets

Future work will involve experimenting with different datasets to assess the model's generalization across various IoT environments and network traffic patterns.

4.1.3 Feature Selection Techniques

Different feature selection methods, such Recursive Feature Elimination (RFE), clustering-based feature selection,Principal Component Analysis (PCA), will be explored to identify the most relevant features and reduce model complexity.

4.1.4 Deep Learning Approaches

Further exploration of deep learning models such as neural networks will be done to enhance model performance and capture complex patterns in the data.

References

[AG19] Ahmet Aksoy and Mehmet Hadi Gunes. Automated iot device identification using network traffic. In *ICC 2019-2019 IEEE international conference on communications (ICC)*, pages 1–7. IEEE, 2019.

[KJL22] Kahraman Kostas, Mike Just, and Michael A Lones. Iotdevid: A behavior-based device identification method for the iot. *IEEE Internet of Things Journal*, 9(23):23741–23749, 2022.

[MMH⁺17] Markus Miettinen, Samuel Marchal, Ibbad Hafeez, Nadarajah Asokan, Ahmad-Reza Sadeghi, and Sasu Tarkoma. Iot sentinel: Automated device-type identification for security enforcement in iot. In *2017 IEEE 37th international conference on distributed computing systems (ICDCS)*, pages 2177–2184. IEEE, 2017.

[NLY⁺24] Kangli Niu, Shenghao Liu, Lingzhi Yi, Xianjun Deng, Suning Chen, Laurence T Yang, and Minmin Cheng. Ensiot: A stacking ensemble learning approach for iot device identification. In *2024 IEEE/ACM 32nd International Symposium on Quality of Service (IWQoS)*, pages 1–6. IEEE, 2024.

[SGL⁺18] Arunan Sivanathan, Hassan Habibi Gharakheili, Franco Loi, Adam Radford, Chamith Wijenayake, Arun Vishwanath, and Vijay Sivaraman. Classifying iot devices in smart environments using network traffic characteristics. *IEEE Transactions on Mobile Computing*, 18(8):1745–1759, 2018.

[SSG⁺17] Arunan Sivanathan, Daniel Sherratt, Hassan Habibi Gharakheili, Adam Radford, Chamith Wijenayake, Arun Vishwanath, and Vijay Sivaraman. Characterizing and classifying iot traffic in smart cities and campuses. In *2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 559–564. IEEE, 2017.