Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 15, No. 3, 2024

IoT Device Identification Based on IP Traffic

Yuanchen Jiang*

Department of Computer Science and Technology, Qingdao University, Qingdao, CO 266071, China
* **Corresponding author Email:** jiangyuanchen@qdu.edu.cn

Abstract: This study aims to address the challenge of IoT device classification by proposing a method based on the random forest classifier. By analyzing the network traffic characteristics of four common household IoT devices, we constructed a feature set and utilized the RF classifier for device identification. The experimental results demonstrate that the random forest classifier performs excellently in terms of precision, recall, and F1 score. By analyzing the network traffic characteristics of four common household IoT devices, we constructed a feature set that includes 20 important device feature information which can effectively represent device identity characteristics, and used the RF classifier for device identification. We conducted numerous experiments on a publicly available dataset and achieved an accuracy rate of 97.22%. The findings offer valuable references for the development of the IoT device identification field and point out potential directions for future research to further enhance the performance and adaptability of the classifier.

Keywords: IoT Security, IoT Device Identification, Random Forest, Network Traffic.

1. Introduction

With the rapid development of Internet of Things (IoT) technology, IoT devices have been widely adopted across various industries, including smart homes, industrial automation, intelligent healthcare, and intelligent transportation systems [1-2]. These interconnected devices enable seamless communication and data exchange through networks, significantly enhancing the convenience and efficiency of daily life and industrial operations [3]. However, the rapid proliferation of IoT devices has also introduced unprecedented challenges in the realm of cybersecurity. Due to the vast diversity and complexity of IoT devices, each device often exhibits unique communication patterns and behavioral characteristics. These characteristics, while beneficial for functionality, can also be exploited by cyber attackers to infiltrate networks by forging critical identity information such as MAC and IP addresses [3-4]. This raises significant concerns regarding network security, making the accurate identification of IoT devices an urgent and critical problem that must be addressed. In recent years, the advancement of machine learning technologies has offered promising solutions to these challenges, providing innovative approaches for network traffic analysis and device identification [5]. Unlike traditional methods such as rule matching and statistical analysis, which often struggle with scalability and adaptability, machine learning methods excel in their ability to automatically learn complex patterns from data. By constructing intelligent models tailored to the characteristics of IoT devices, researchers can extract highly discriminative features from large-scale network traffic data, enabling both efficient and precise device identification [6-7]. Machine learning models also demonstrate superior generalization capabilities, allowing them to perform effectively in dynamic and heterogeneous network environments where traditional methods fall short [8]. This makes them particularly well-suited for addressing the challenges posed by the rapid evolution and increasing complexity of IoT ecosystems. This paper aims to advance the field of IoT device identification by exploring methods based on network traffic analysis, with a particular focus on understanding how different traffic features impact identification performance. We propose a machine learning-based classification framework that extracts 21 carefully selected and highly effective features from raw network traffic to serve as input vectors for the model [9]. By conducting a comprehensive comparative analysis of various machine learning algorithms, we achieved a device identification accuracy rate of 97%, highlighting the effectiveness of our approach. The primary goal of this work is to strike a balance between efficiency and accuracy, enabling the proposed method to adapt to the increasingly complex and dynamic IoT network environments. This research provides a robust and practical solution for improving device identification performance, contributing to enhanced network security in IoT systems [10].

2. Proposed Method

2.1. Attack Model

As shown in Figure 1, the attacker infiltrates the IoT system by disguising their MAC address, exploiting vulnerabilities in network security. The left side of the diagram depicts a hacker figure, symbolizing the attacker, who utilizes a laptop as the primary attack tool [11-12]. This representation highlights the simplicity with which an attacker can carry out such actions, emphasizing the importance of robust security measures. By employing sophisticated technical means, the attacker manipulates their MAC address to make it appear identical to that of a legitimate device within the IoT system. This method allows them to bypass authentication protocols and gain unauthorized access to the network. In the center of the diagram, various IoT devices are illustrated, showcasing the diversity and ubiquity of devices typically present in an IoT ecosystem. Examples include surveillance cameras, smart bulbs, smart locks, smart refrigerators, and more [13]. These devices are interconnected via wireless networks and rely on cloud platforms to facilitate remote control, data exchange, and seamless integration into the broader IoT ecosystem. While this connectivity enhances convenience and functionality, it also increases the system's vulnerability to external attacks, such as MAC address spoofing. By

disguising their MAC address, the attacker effectively impersonates a legitimate IoT device. This deception allows them to infiltrate the IoT system and circumvent the network's security measures designed to detect and block unauthorized devices [14]. Once inside the network, the attacker gains access to critical resources and can execute a range of malicious activities. These actions may include stealing sensitive data, such as user credentials, private communications, or operational logs, which can compromise the privacy and security of both users and the system. Additionally, the attacker can disrupt the functionality of connected devices, rendering them inoperable or causing them to behave erratically. For instance, they might disable surveillance cameras, unlock smart locks without authorization, or manipulate the settings of smart appliances, thereby causing significant inconvenience or harm. Furthermore, the attacker may use their foothold within the network to launch further attacks, spreading malware, initiating denial-of-service (DoS) attacks, or creating a pathway to compromise other connected networks [15]. This scenario underscores the critical need for advanced security mechanisms in IoT systems, such as more robust device identification methods, encryption protocols, and anomaly detection techniques. By addressing these vulnerabilities, IoT ecosystems can better protect against sophisticated attacks and ensure the security and reliability of connected devices and their associated networks.

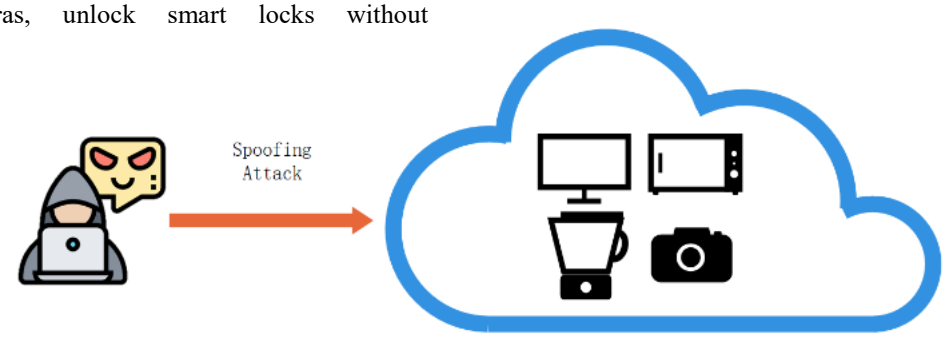

Figure 1. Attacker infiltrating the Internet of Things by disguising MAC addresses

This type of attack poses a significant and serious threat to the security and integrity of IoT systems [16]. Due to the inherent vulnerabilities and lack of robust security measures in many IoT devices, attackers can often exploit these weaknesses with relative ease. Most IoT devices are designed with a primary focus on functionality and convenience, often at the expense of comprehensive security features. This oversight leaves them particularly susceptible to sophisticated attacks, such as identity spoofing or unauthorized access. Moreover, the sheer number and widespread distribution of IoT devices further exacerbate the issue [17]. IoT devices are deployed in various environments, ranging from homes and offices to industrial settings and public infrastructure. This diversity makes it challenging to implement and maintain consistent security protocols across all devices. Once an attacker successfully compromises a single device within the network, it can serve as a gateway to infiltrate and compromise other connected devices. The interconnected nature of IoT systems enables attackers to quickly propagate their influence throughout the network, amplifying the damage caused [18]. The potential consequences of such an attack are far-reaching. Compromised IoT devices can be used to steal sensitive information, disrupt critical services, and even launch large-scale cyberattacks, such as distributed denial-of-service (DDoS) attacks [19]. In industrial settings, such breaches could lead to operational failures, financial losses, or even physical harm. In a smart home or healthcare environment, the repercussions could include compromised privacy, data breaches, or threats to personal safety. The lack of adequate security measures combined with the distributed nature of IoT systems underscores the urgent need for enhanced cybersecurity strategies. Solutions such as end-to-end encryption, secure device authentication, and real-time anomaly detection are critical to mitigating these risks. By addressing these vulnerabilities proactively, organizations can strengthen the resilience of their IoT systems, protect sensitive data, and maintain the trust and safety of users [20].

2.2. Design and Implementation of IoT Device Identification Scheme

2.2.1. Solution Overview

The system first captures network data packets in real-time through sensors deployed at key network nodes. These packets contain raw information about device communication, providing a foundation for subsequent feature extraction. These features can reflect the behavioral patterns and communication characteristics of devices. Based on the extracted features, the system constructs a feature set to describe the unique behaviors of different IoT devices. The construction of the feature set takes into account the distinctiveness, stability, and availability of features to ensure the accuracy and robustness of the model. Using the constructed feature set, the system employs machine learning algorithms for device classification and identification. Through training datasets, the model learns the characteristic patterns of different devices and optimizes classification boundaries to improve the accuracy of identification.

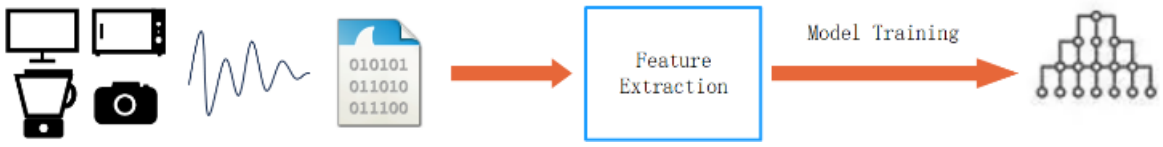

Figure 2. System Model

2.2.2. Solution Overview

As shown in Table 1, this is a carefully designed feature set that provides a comprehensive and in-depth basis for network behavior analysis for machine learning models. These

features not only include basic network layer information, such as packet size (Pck_size), IP header length (IP_ihl), IP time to live (IP_ttl), but also delve into the transport and application layers, including specific flags for TCP and UDP (such as TCP_ACK, TCP_FIN, UDP_len), as well as BOOTP and DNS parameters related to device configuration and network services (such as BOOTP_sname, BOOTP_file, DNS_qr, DNS_r).

The diversity and depth of these features enable the model to capture device behavior patterns from multiple perspectives, thereby improving the accuracy of identification. For instance, the length features of TCP and UDP can reveal the amount of data communicated by devices, while the TCP flags can reflect changes in connection status. IP options and flags provide additional information about the packet transmission path and network configuration, which is crucial for identifying specific types of network traffic. BOOTP and DNS features involve how devices obtain network configuration and domain name resolution services, and this information is significant for understanding device network interactions and locating potential security risks. By analyzing these features, the activity patterns of devices in the network can be identified, including both normal and abnormal behaviors, thus providing security for network operations. Moreover, the extensibility of these features means that as IoT technology evolves and new protocols emerge, the feature set can be continuously updated and expanded to adapt to new network environments and challenges.

Table 1. Feature Set

Pck_size	TCP_ACK
TCP_FIN	TCP_dataofs
TCP_window	UDP_len
IP_DF	IP_flags
IP_ihl	IP_len
IP_options	IP_tos
IP_ttl	BOOTP_hlen
BOOTP_flags	BOOTP_sname
BOOTP_file	BOOTP_options
DNS_qr	DNS_rd

3. Experiments and Analysis

In this study, we utilized a publicly available dataset to conduct a detailed classification study of four common household IoT devices. These devices include the D-Link camera (DlinkCam), D-Link home gateway (DlinkHomeHub), Philips Hue smart switch (HueSwitch), and Withings smart devices (Withing). These devices are widely used in daily life, covering various aspects such as home security monitoring, network management, smart lighting, and health monitoring.

To validate the feasibility and effectiveness of our IoT device classification scheme, we conducted an in-depth comparative analysis using five different machine learning techniques.

3.1. K-Nearest Neighbors (KNN)

KNN algorithm is a straightforward instance-based learning method that classifies data points based on their proximity to other labeled examples in the feature space. It operates on the principle that similar instances are likely to belong to the same category. KNN uses a distance metric, such as Euclidean or Manhattan distance, to measure the closeness between data points, which makes it particularly useful for applications with clearly separable clusters. Despite its simplicity and intuitiveness, KNN has certain limitations, such as sensitivity to noise in the data and higher computational costs for large datasets, as it requires computing the distance to every other point during classification. Nevertheless, its ease of implementation and effectiveness on smaller datasets make it a commonly used benchmark in classification tasks.

3.2. Decision Tree (DT)

DT algorithms predict the target variable by iteratively splitting the dataset based on feature values. Each split corresponds to a decision rule, forming a tree-like structure that leads to a predicted class label or continuous value at the leaves. Decision trees are highly interpretable and allow users to visualize the decision-making process, which is particularly valuable in applications where explainability is crucial. They handle both categorical and numerical data and are robust to missing values. However, decision trees tend to overfit the data if not properly pruned, especially when dealing with noisy datasets. Their versatility and interpretability make them a popular choice for a variety of machine learning problems.

3.3. Gradient Boosting (GB)

GB is a powerful ensemble learning technique that builds a strong predictive model by combining multiple weak learners, typically decision trees. By iteratively refining the predictions of previous models, GB minimizes the error through a gradient descent optimization process. Each subsequent tree focuses on correcting the errors of the previous ones, resulting in a model that is both accurate and robust. GB algorithms, such as XGBoost, LightGBM, and CatBoost, have gained significant popularity in recent years due to their exceptional performance in various machine learning competitions and real-world applications. While GB is highly effective, it requires careful tuning of hyperparameters to avoid overfitting and ensure computational efficiency, particularly when applied to large-scale datasets.

3.4. Naive Bayes (NB)

NB is a probabilistic classifier grounded in Bayes' theorem, which assumes that all features are independent given the class label. Despite the simplicity of this assumption, which rarely holds true in practice, NB often performs remarkably well, particularly in text classification and spam detection tasks. Its computational efficiency and scalability make it suitable for handling high-dimensional data, where other algorithms might struggle. Naive Bayes is easy to implement and requires minimal training time, as it estimates probabilities directly from the training data. However, its assumption of feature independence can limit its applicability to certain problems, especially when features are highly correlated. Nonetheless, its efficiency and solid performance in specific domains make it a valuable tool in the machine learning toolkit.

As shown in Figure 3, in the comparison of machine learning models for IoT device classification, RF demonstrated the best performance with an accuracy rate of 97.22%, showing its outstanding capability in handling complex datasets. Following closely behind are KNN and GB, with accuracy rates of 96.49% and 96.41%, respectively, exhibiting efficient classification capabilities. However, due to the assumption of feature independence in NB, which may not be easily

met in practical applications, its accuracy rate is significantly lower than that of the other models, at only 71.96%. Therefore, we selected RF, which has the best classification accuracy among the models, as our classification model.

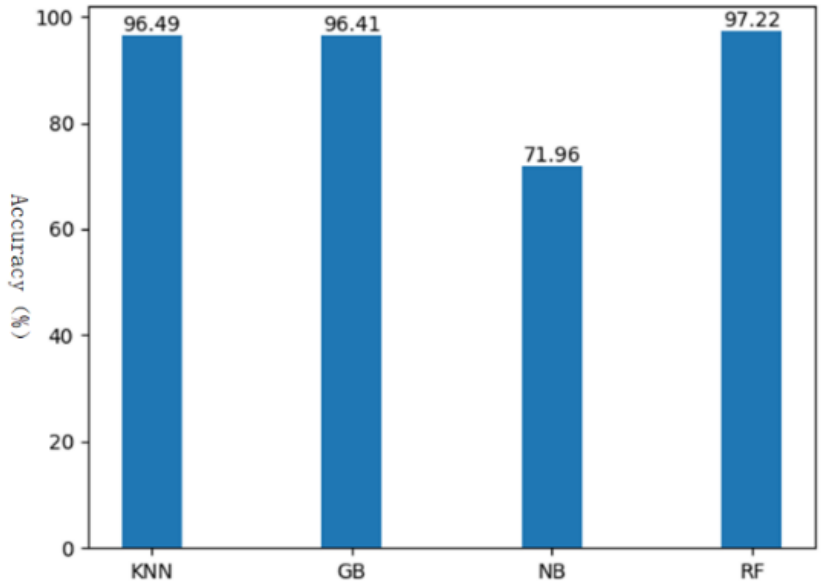

Figure 3. Model Classification Results

Table 2. Evaluation Metrics for Each Device

Device Name	Precision	Recall	F1 Score
D-LinkCam	89.71%	96.06%	92.77%
D-LinkHomeHub	97.03%	91.80%	94.34%
HueSwitch	99.95%	100.00%	99.97%
Withings	100.00%	100.00%	100.00%

The RF classifier performed exceptionally well in the task of identifying four types of IoT devices, especially when identifying HueSwitch and Withings devices, achieving nearly perfect precision, recall, and F1 scores, all at 100%. For D-LinkCam, although the recall rate is high, the precision rate is relatively low, indicating that there is room for improvement in reducing false positives. The identification of D-LinkHomeHub shows a high precision rate and a relatively low recall rate, meaning that the classifier performs well in reducing false positives but may need to further improve recall to ensure all devices are correctly identified. Overall, the RF classifier demonstrated efficiency and accuracy in IoT device classification, providing valuable insights into the performance for different devices, and also pointing out the potential need for performance optimization for specific devices.

4. Conclusion

In this study, we delved into the challenges of IoT device classification and proposed a solution based on the RF classifier. By comparing and analyzing four common household IoT devices—HueSwitch, Withings, D-LinkCam, and D-LinkHomeHub—we demonstrated the superior performance of the RF classifier in terms of precision, recall, and F1 score. The results not only validate the effectiveness of our approach but also provide new perspectives and methods for the field of IoT device identification.

References

[1] Meidan Y, Bohadana M, Shabtai A, et al. ProfilIoT: A machine learning approach for IoT device identification based on network traffic analysis[C]//Proceedings of the symposium on applied computing. 2017: 506-509.

[2] Salman O, Elhajj I H, Chehab A, et al. A machine learning based framework for IoT device identification and abnormal traffic detection[J]. Transactions on Emerging Telecommunications Technologies, 2022, 33(3): e3743.

[3] Chowdhury R R, Aneja S, Aneja N, et al. Network traffic analysis based iot device identification[C]//Proceedings of the 2020 4th International Conference on Big Data and Internet of Things. 2020: 79-89.

[4] Aksoy A, Gunes M H. Automated iot device identification using network traffic[C]//ICC 2019-2019 IEEE international conference on communications (ICC). IEEE, 2019: 1-7.

[5] Ullah I, Mahmoud Q H. Network traffic flow based machine learning technique for IoT device identification[C]//2021 IEEE international systems conference (SysCon). IEEE, 2021: 1-8.

[6] Shahid M R, Blanc G, Zhang Z, et al. IoT devices recognition through network traffic analysis[C]//2018 IEEE international conference on big data (big data). IEEE, 2018: 5187-5192.

[7] Yousefnezhad N, Malhi A, Främling K. Automated iot device identification based on full packet information using real-time network traffic[J]. Sensors, 2021, 21(8): 2660.

[8] Kotak J, Elovici Y. Iot device identification using deep learning[C]//13th International Conference on Computational Intelligence in Security for Information Systems (CISIS 2020) 12. Springer International Publishing, 2021: 76-86.

[9] Santos M R P, Andrade R M C, Gomes D G, et al. An efficient approach for device identification and traffic classification in IoT ecosystems[C]//2018 IEEE Symposium on Computers and Communications (ISCC). IEEE, 2018: 00304-00309.

[10] Kolcun R, Popescu D A, Safronov V, et al. Revisiting iot device identification[J]. arXiv preprint arXiv:2107.07818, 2021.

[11] Pinheiro A J, Bezerra J M, Burgardt C A P, et al. Identifying IoT devices and events based on packet length from encrypted traffic[J]. Computer Communications, 2019, 144: 8-17.

[12] Luo Y, Chen X, Ge N, et al. Transformer-based device-type identification in heterogeneous IoT traffic[J]. IEEE Internet of Things Journal, 2022, 10(6): 5050-5062.

[13] Sivanathan A, Gharakheili H H, Loi F, et al. Classifying IoT devices in smart environments using network traffic characteristics[J]. IEEE Transactions on Mobile Computing, 2018, 18(8): 1745-1759.

[14] Guo H, Heidemann J. Ip-based iot device detection[C]//Proceedings of the 2018 workshop on IoT security and privacy. 2018: 36-42.

[15] Hussain S, Aslam W, Mehmood A, et al. A machine learning based framework for IoT devices identification using web traffic[J]. PeerJ Computer Science, 2024, 10: e1834.

[16] Hamad S A, Zhang W E, Sheng Q Z, et al. Iot device identification via network-flow based fingerprinting and learning[C]//2019 18th IEEE international conference on trust, security and privacy in computing and communications/13th IEEE international conference on big data science and engineering (TrustCom/BigDataSE). IEEE, 2019: 103-111.

[17] Yin F, Yang L, Ma J, et al. Identifying iot devices based on spatial and temporal features from network traffic[J]. Security and Communication Networks, 2021, 2021(1): 2713211.

[18] Charyyev B, Gunes M H. Locality-sensitive iot network traffic fingerprinting for device identification[J]. IEEE Internet of Things Journal, 2020, 8(3): 1272-1281.

[19] Gu D, Zhang J, Tang Z, et al. IoT device identification based on network traffic[J]. Wireless Networks, 2024: 1-17..

[20] Hao Q, Rong Z. IoTTFID: an incremental IoT device identification model based on traffic fingerprint[J]. IEEE Access, 2023, 11: 58679-58691.