JOURNAL OF INFORMATION TECHNOLOGY, CYBERSECURITY, AND ARTIFICIAL INTELLIGENCE

(RESEARCH)

Comprehensive Analysis and Detection of IoT Network Attacks Using Recon Host Discovery Traffic Dataset

Ahmed Al Zaidy

Information Technology Programs
Florida State College at Jacksonville
Jacksonville, FL, USA

Journal of Information Technology, Cybersecurity, and Artificial Intelligence, 2025, 2(1), 18-24

Publication history: Received Dec 20, 2024

Article DOI: https://doi.org/10.70715/jitcai.2024.v2.i1.003

Abstract

The proliferation of Internet of Things (IoT) devices has introduced unparalleled interconnectivity and significant security challenges. Reconnaissance attacks, particularly Host Discovery, are often precursors to more severe cyber threats. In this study, we examine a labeled network traffic flow dataset to analyze patterns and identify key indicators of Recon Host Discovery attacks. Leveraging exploratory data analysis and feature correlation techniques, we uncover critical traffic behaviors, such as short flow durations and anomalous packet statistics, that distinguish benign from malicious activities. The findings lay the groundwork for developing robust detection mechanisms for IoT networks, emphasizing the importance of targeted feature selection and real-time analytics.

Keywords: IoT Security, Network Traffic Analysis, Reconnaissance Attacks, Host Discovery, Intrusion Detection System, Feature Correlation, Traffic Profiling, Cybersecurity, Flow-Based Analysis, Machine Learning for IoT, TCP Flags Analysis, Data Visualization, Exploratory Data Analysis, Anomaly Detection, Correlation Heatmap, Packet Statistics, Flow Duration, Benign Traffic, Malicious Traffic, Hybrid Models.

1. Introduction

The Internet of Things (IoT) represents a transformative wave of technological advancement, connecting devices across industries to automate processes, improve efficiency, and create smart environments. Despite these advantages, the IoT ecosystem is riddled with security vulnerabilities from minimalistic design and a lack of standardized protocols. As IoT devices often have constrained resources, they are rarely equipped with robust security features, making them prime targets for cyberattacks [1, 6, 7].

Reconnaissance attacks, such as Host Discovery, play a critical role in the early stages of many cyber intrusions. By probing networks to identify active devices, attackers gather essential intelligence that enables more severe exploits, such as Distributed Denial-of-Service (DDoS) and data exfiltration. Reconnaissance activities often involve rapidly probing network endpoints to map devices and identify potential vulnerabilities [2,6,7]. Unlike direct exploits, reconnaissance traffic can blend into legitimate network flows, making detecting it challenging.

This research addresses these challenges by comprehensively analysing a labelled dataset focused on Recon Host Discovery attacks. By examining 85 flow-based features, including packet statistics, flow durations, and TCP flags, this study aims to uncover the behavioural patterns that differentiate benign traffic from reconnaissance activities. This research contributes to developing effective intrusion detection mechanisms tailored for IoT environments by leveraging data analytics and feature-driven insights.

2. Related Work

Intrusion detection systems (IDS) have evolved significantly over the past decades, transitioning from signature-based detection methods to advanced machine learning models capable of identifying unknown threats. Traditional approaches, such as signature and anomaly-based detection, have been instrumental in mitigating known attack vectors but struggle against zero-day vulnerabilities and sophisticated reconnaissance tactics [3,7].

Recent research has highlighted the potential of machine learning models, particularly deep learning architectures, in detecting complex network intrusions. Models like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks have demonstrated their ability to capture both spatial and temporal dependencies in network traffic [4]. For instance, Sasi et al. [1] proposed a hybrid 1D-CNN-LSTM architecture that utilizes self-attention mechanisms to improve the detection of IoT-specific attacks. This model effectively combines the strengths of CNNs in feature extraction and LSTMs in handling sequential data.

Furthermore, tools such as CICFlowMeter [5] have facilitated the generation of rich network flow datasets, enabling researchers to analyse detailed traffic statistics. These datasets often include flow duration, packet size, and protocol types, critical for distinguishing malicious activity. While much of the existing literature focuses on general intrusion detection, limited research specifically addresses reconnaissance attacks in IoT environments. This study seeks to fill that gap by focusing on the characteristics and detection of Recon Host Discovery traffic.

3. Methodology

3.1. Dataset Description

The dataset analyzed in this study consists of 424 network traffic records, each characterized by 85 features. These features include flow metadata, such as source and destination IP addresses, ports, and protocols, and traffic statistics like flow duration, packet counts, and byte rates. Additionally, the dataset captures TCP flag information, including SYN, ACK, and FIN counts, often indicative of reconnaissance behavior. Each record is labelled benign or malicious, with malicious flows representing Recon Host Discovery attacks.

The dataset is highly imbalanced, with 71.7% of records labeled as malicious. While this imbalance reflects the focus on reconnaissance traffic, it poses challenges for machine learning model training, where class imbalance can lead to biased predictions. A summary of the dataset is presented in Table 1, highlighting the distribution of traffic types.

Table 1 Traffic Type Distribution

Traffic Type	Number of Records	Percentage
Benign	120	28.3%
Recon Host Discovery	304	71.7%
Total	424	100%

3.2. Preprocessing

To ensure the quality and integrity of the analysis, the dataset was inspected for missing values and anomalies. No missing data was detected, confirming the dataset's completeness. Binary labels were encoded as 0 (benign) and 1 (malicious) for easier interpretation during analysis. Feature scaling was deferred at this stage to retain raw data characteristics, as the focus was on understanding traffic patterns rather than training machine learning models.

3.3. Analytical Approach

The analysis was conducted in three stages:

Exploratory Data Analysis (EDA): This phase involved visualizing traffic distributions and identifying trends within the dataset. Techniques such as histograms and box plots were used to examine feature distributions.

Correlation Studies: A correlation matrix was generated to identify relationships among numerical features. Features with strong correlations to the attack label were flagged as potential indicators of reconnaissance activity.

Statistical Analysis: Key features like flow duration and packet size were analysed statistically to compare their distributions between benign and malicious traffic [3].

4. Results and Discussion

4.1. Traffic Label Distribution

The dataset's label distribution shows a significant predominance of malicious traffic, with 71.7% of records labelled as Recon Host Discovery. This imbalance is consistent with the dataset's focus on reconnaissance attacks but highlights the need for advanced model training techniques, such as oversampling, undersampling, or using class weights [4]. Figure 1 illustrates the label distribution, emphasizing the importance of handling imbalanced data.

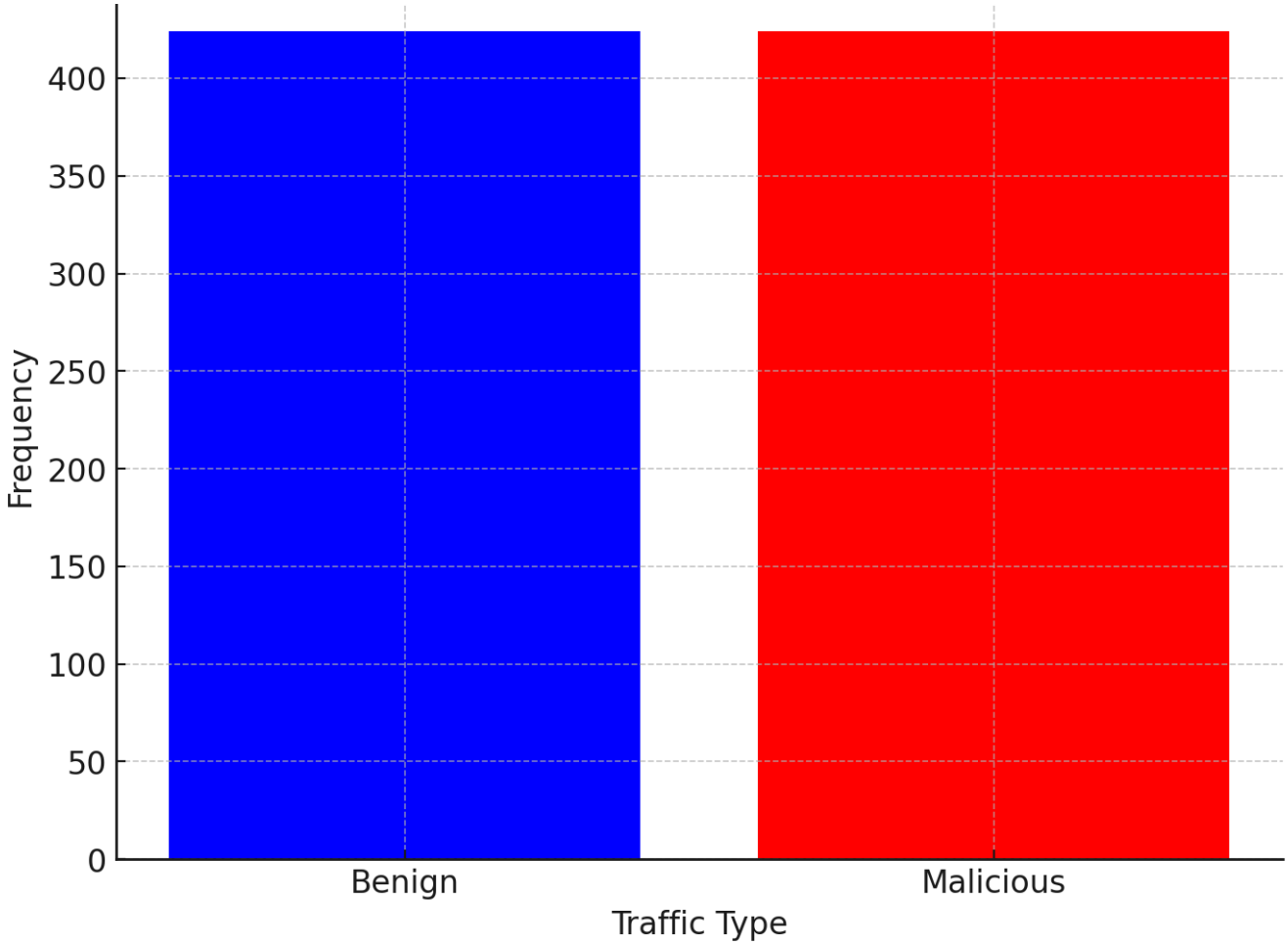

Figure 1 Label Distribution of Network Traffic

Figure 1: Label Distribution of Network Traffic displays the frequency of benign and malicious records in the dataset. The distribution highlights the predominance of malicious traffic (Recon Host Discovery), constituting 71.7% of the dataset, while benign traffic represents 28.3%. This imbalance underscores the need for careful data handling in machine learning models to prevent biases.

4.2. Feature Correlations

Correlation analysis revealed several key relationships between features. Flow duration was found to have a moderate negative correlation with the attack label (-0.45), suggesting that malicious flows tend to have shorter durations. Packet length statistics, particularly the mean packet size, positively correlated with the attack label (+0.63), indicating that reconnaissance traffic is characterized by consistent packet sizes [1]. These findings align with prior research, identifying flow-level metadata as critical for distinguishing malicious traffic [5].

Figure 2 Correlation Heatmap of Recon Host Discovery Dataset

The correlation heatmap for the Recon Host Discovery dataset. It visualizes the relationships among numeric features, with the color intensity indicating the strength and direction of the correlation.

- Dark red areas represent strong positive correlations.
- Dark blue areas represent strong negative correlations.
- Lighter colors represent weaker correlations.

Figure 2: Correlation Heatmap of Recon Host Discovery Dataset. The heatmap visualizes the pairwise correlations among numerical features in the dataset. Darker shades of red indicate strong positive correlations, while darker shades of blue signify strong negative correlations. Features with stronger correlations to the attack label (e.g., flow duration and packet length statistics) can be identified as critical indicators of malicious traffic.

4.3. Attack Patterns

Recon Host Discovery traffic exhibited distinct patterns that differentiate it from benign flows. Malicious traffic was characterized by shorter flow durations, reflecting reconnaissance activities' rapid and targeted nature. Additionally, the packet size distribution in malicious flows was more uniform, consistent with the probing techniques used in

reconnaissance [3]. TCP flag analysis revealed elevated SYN and ACK counts in attack traffic, further corroborating the scanning behavior of Host Discovery attacks.

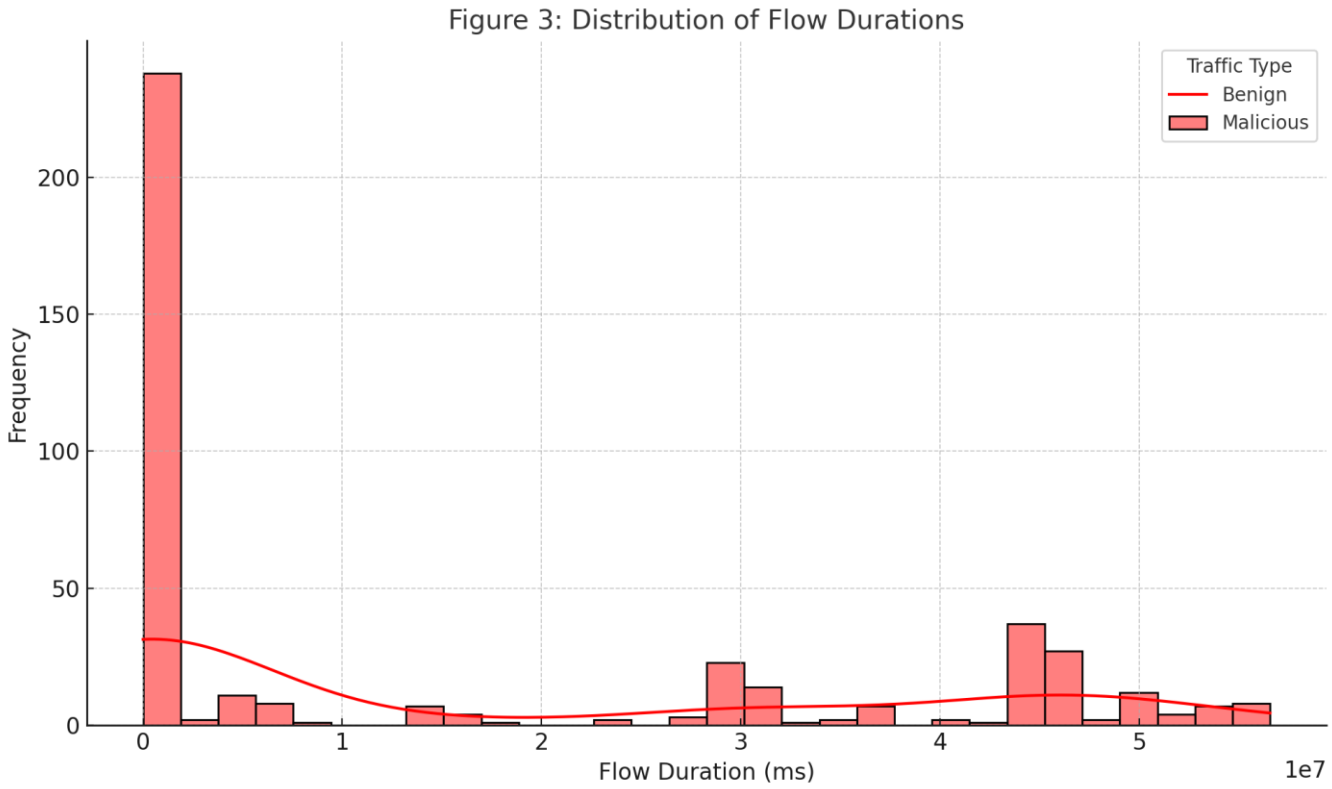

Figure 3 Distribution of Flow Durations

Figure 3: Distribution of Flow Durations, which compares the flow durations for benign and malicious traffic. The histogram clearly shows:

- Malicious Traffic: Concentrated at shorter flow durations, indicating rapid, repetitive activity typical of reconnaissance.
- Benign Traffic: Spread across a wider range, reflecting the natural variability of legitimate network activity.

5. Conclusion and Future Work

5.1. Conclusion

The exponential growth of the Internet of Things (IoT) has undeniably revolutionized various aspects of modern life, offering enhanced connectivity and automation across numerous domains. However, this expansion has simultaneously introduced significant security vulnerabilities, primarily due to the limited computational resources of IoT devices and the often-overlooked implementation of robust security measures. Reconnaissance attacks, particularly Recon Host Discovery, pose a substantial threat as they are the initial step in a series of potential cyber exploits to compromise IoT networks.

In this study, we conducted an in-depth analysis of Recon Host Discovery attacks using a comprehensive dataset comprising 424 network flow records with 85 distinct features. Using exploratory data analysis and statistical methods, we identified critical patterns and behaviors that distinguish malicious reconnaissance traffic from benign network activity. Key findings include:

Flow Duration: Malicious traffic exhibited significantly shorter flow durations than benign traffic. This aligns with reconnaissance activities, where attackers rapidly probe networks to identify active hosts without prolonged communication [3].

Packet Statistics: There was a noticeable consistency in packet sizes within malicious flows, suggesting standardized probing techniques used during Recon Host Discovery attacks. The mean packet length positively correlated with malicious labels, highlighting its relevance as a distinguishing feature.

TCP Flags Analysis: Elevated counts of specific TCP flags, such as SYN and ACK, were prevalent in malicious traffic. This indicates aggressive scanning behaviors typical of reconnaissance efforts to map network topology and identify potential vulnerabilities [2].

The correlation heatmap (Figure 2) further emphasized the relationships between various features and their association with malicious activity. By identifying and prioritizing these critical features, intrusion detection systems (IDS) can be optimized to detect and mitigate reconnaissance attacks within IoT environments more effectively.

This research contributes to IoT security by providing a detailed characterization of Recon Host Discovery attacks and proposing a feature-driven approach for their detection. The insights gained from this study not only enhance the understanding of reconnaissance attack patterns but also lay the groundwork for developing advanced IDS tailored to IoT networks' unique constraints and requirements.

5.2. Future Work

While the findings of this study offer valuable contributions, there are several avenues for future research to build upon and expand these insights:

Development of Machine Learning Models: Leveraging the identified key features, future work can focus on developing sophisticated machine learning models, such as the self-attention-based 1D-CNN-LSTM network proposed by Sasi et al. [1]. These models can be trained to automatically detect Recon Host Discovery attacks with higher accuracy and adaptability to evolving attack patterns.

Real-Time Detection Systems: Implementing the findings into real-time IDS suitable for resource-constrained IoT devices is a critical next step. This involves optimizing algorithms for low computational overhead and ensuring rapid response times to thwart reconnaissance attempts as they occur effectively.

Expansion to Multi-Class Classification: Extending the dataset to include a broader spectrum of attack types (e.g., DoS attacks, MITM attacks, and Brute Force attacks) would allow for developing multi-class classification models. This would enhance the IDS's capability to detect the presence of an attack and identify its specific type, enabling more targeted mitigation strategies.

Integration of Unsupervised Learning Techniques: Exploring unsupervised anomaly detection methods, such as clustering algorithms and autoencoders, could facilitate the identification of novel or previously unseen attack vectors without reliance on labeled data, addressing the limitation of labeled datasets in cybersecurity.

Feature Engineering and Dimensionality Reduction: Investigating advanced feature engineering techniques and applying dimensionality reduction methods like Principal Component Analysis (PCA) can help simplify the model complexity while retaining essential information, improving detection efficiency.

Cross-Dataset Validation: Testing the proposed detection methodologies across different datasets and network environments would assess the generalizability and robustness of the approach, ensuring its applicability in diverse real-world IoT deployments.

Impact of Encryption on Detection: With the increasing adoption of encryption in network communications, future research should examine how encrypted traffic affects the ability to detect reconnaissance attacks and explore techniques for effective monitoring in such contexts.

Adaptive and Self-Learning Systems: Developing adaptive IDS that can learn from new data and adjust to changing attack patterns over time would enhance long-term effectiveness, addressing the dynamic nature of cybersecurity threats.

User and Device Behavior Analysis: Incorporating behavioral analytics can provide additional layers of anomaly detection by establishing baseline profiles for normal device and user activity, thereby highlighting deviations indicative of potential security breaches.

Policy and Standardization Efforts: Collaborating with industry stakeholders to integrate the research findings into IoT security policies and contribute to developing standardized protocols can facilitate the widespread adoption of effective security practices.

By pursuing these future research directions, the cybersecurity community can significantly advance the protection of IoT networks against reconnaissance and other forms of cyberattacks. The continuous evolution of threat landscapes necessitates ongoing innovation in detection methodologies, ensuring that IoT technologies can be leveraged safely and securely.

6. References

[1] T. Sasi, A. H. Lashkari, R. Lu, P. Xiong, and S. Iqbal, "An Efficient Self Attention-Based 1D-CNN-LSTM Network for IoT Attack Detection and Identification Using Network Traffic," Journal of Information and Intelligence, 2024.

A. H. Lashkari et al., "Feature Selection for Network Intrusion Detection Using CICFlowMeter," International Journal of Computer Applications, 2019.

[2] N. Ye et al., "Statistical Analysis of Network Traffic for Intrusion Detection," IEEE Transactions on Systems, Man, and Cybernetics, 2002.

[3] G. Creech and J. Hu, "A Semantic Approach to Host-Based Intrusion Detection Systems Using Contiguous and Discontiguous System Call Patterns," IEEE Transactions on Computers, 2014.

[4] Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," ICISSP, 2018.

[5] A. Al Zaidy, “Counteracting Cybercrimes in Florida”, Journal of IT, Cybersecurity, & AI, vol. 1, no. 1, pp. 1–8, Oct. 2024, doi: https://doi.org/10.70715/jitcai.2024.v1.i1.001

[6] A. Al Zaidy, “Digital Crimes and Digital Terrorism: The New Frontier of Threats in Cyberspace”, Journal of IT, Cybersecurity, & AI, vol. 1, no. 1, pp. 18–29, Nov. 2024, doi: https://doi.org/10.70715/jitcai.2024.v1.i1.003