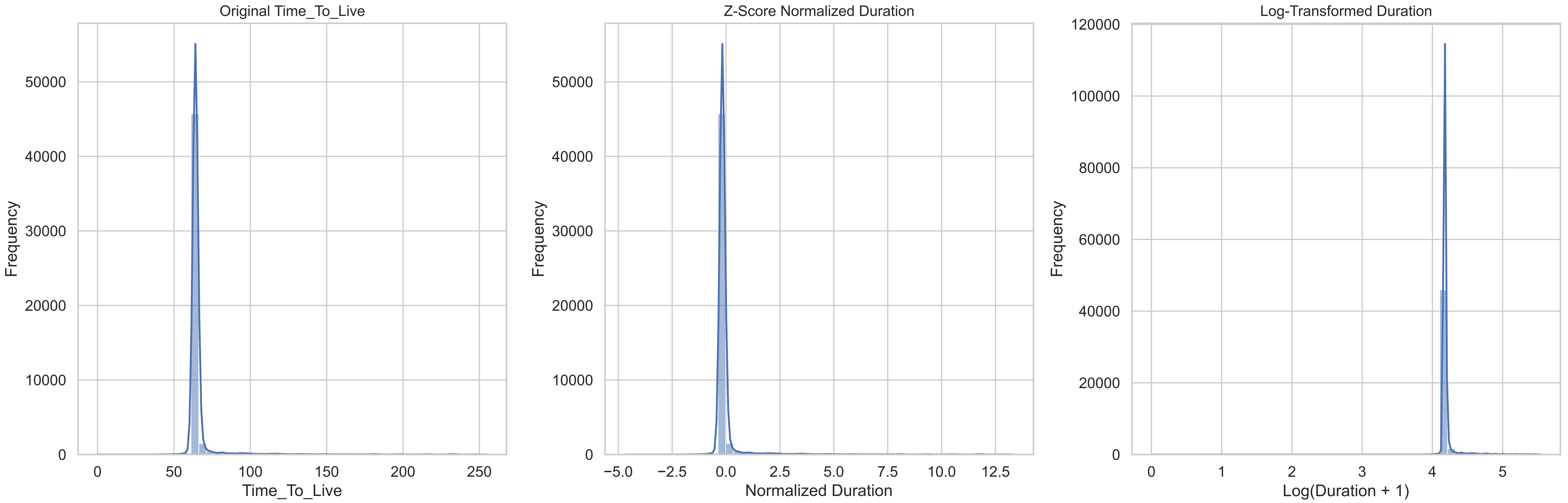
IoT Security Threat Detection for SMEs:

A Machine Learning Approach Using CIC-IoT Dataset

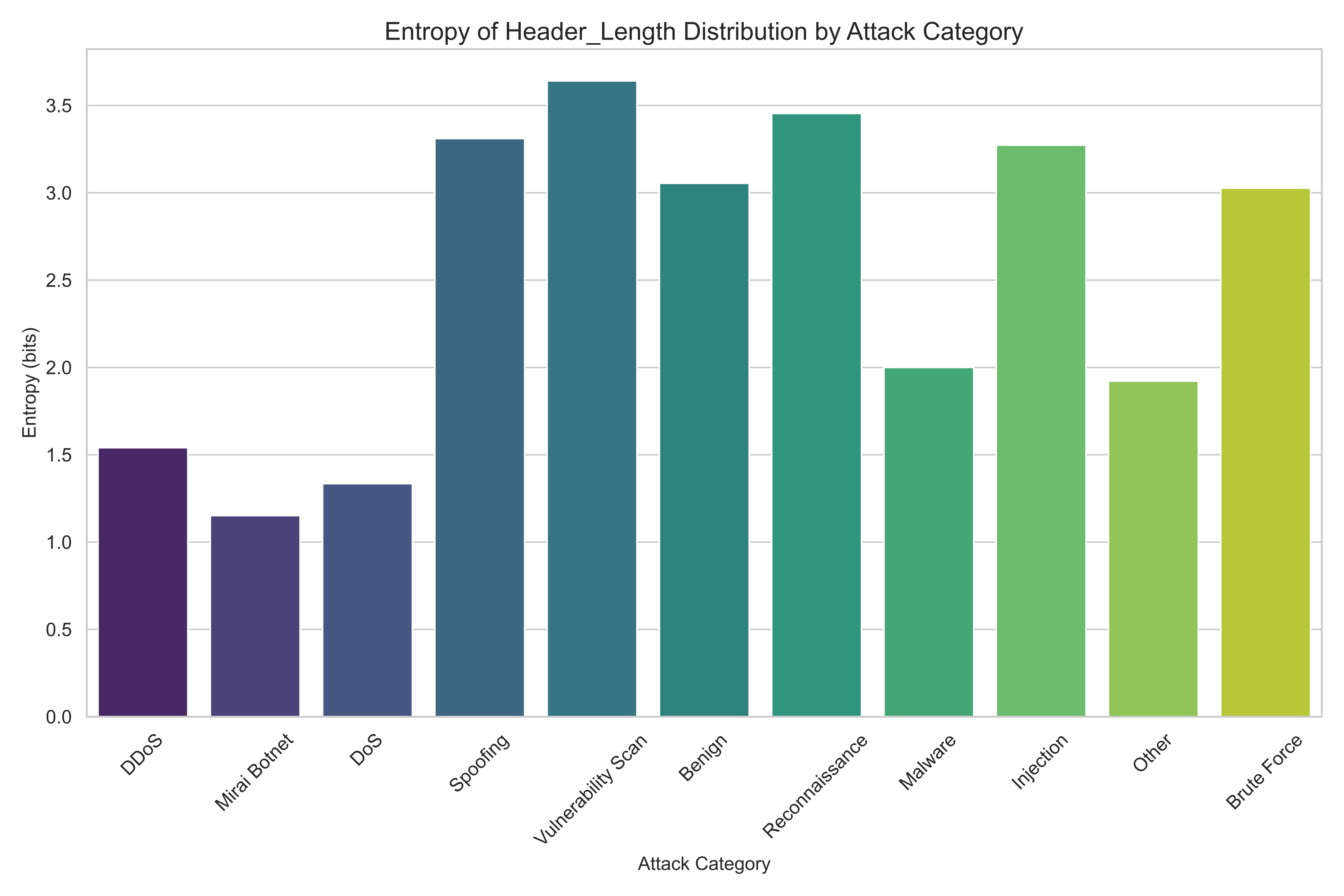
STAGE 2, STEP 2: FEATURE ENGINEERING

This report presents feature engineering techniques for IoT security threat detection, creating derived features that enhance attack detection capabilities while optimizing for the resource constraints of SME environments.



Original Time_To_Live
Frequency
Time_To_Live
Z-Score Normalized Duration
Frequency
Normalized Duration
Log-Transformed Duration
Frequency
Log(Duration + 1)

This figure shows the transformation of flow duration (Time_To_Live) to improve its utility for machine learning models. The left panel shows the original distribution, which is often skewed with extreme outliers that can dominate distance-based models. The center panel displays the Z-score normalized version, which centers the distribution around zero with unit variance, making it more suitable for algorithms sensitive to feature scales. The right panel shows a log-transformed version, which compresses the range of extreme values while preserving relative ordering. These transformations are particularly important for IoT security monitoring in SMEs, where flow duration can vary dramatically between normal traffic and attack patterns like DDoS. The normalized features improve model accuracy while reducing the impact of outliers that might trigger false positives in production environments.

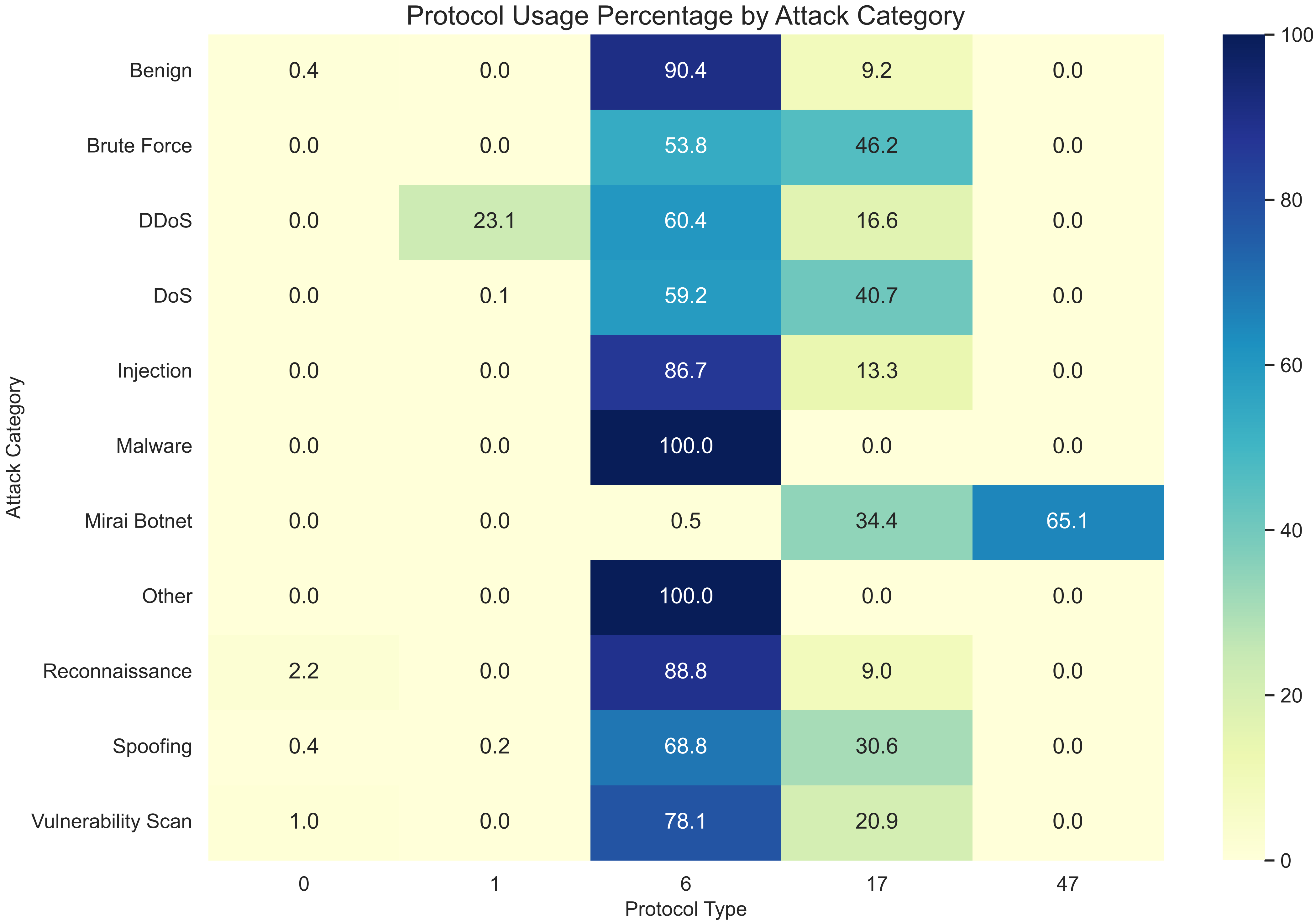


Entropy of Header_Length Distribution by Attack Category
Entropy (bits)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
DDoS
Mirai Botnet
DoS
Spoofing
Vulnerability Scan
Benign
Reconnaissance
Malware
Injection
Other
Brute Force
Attack Category

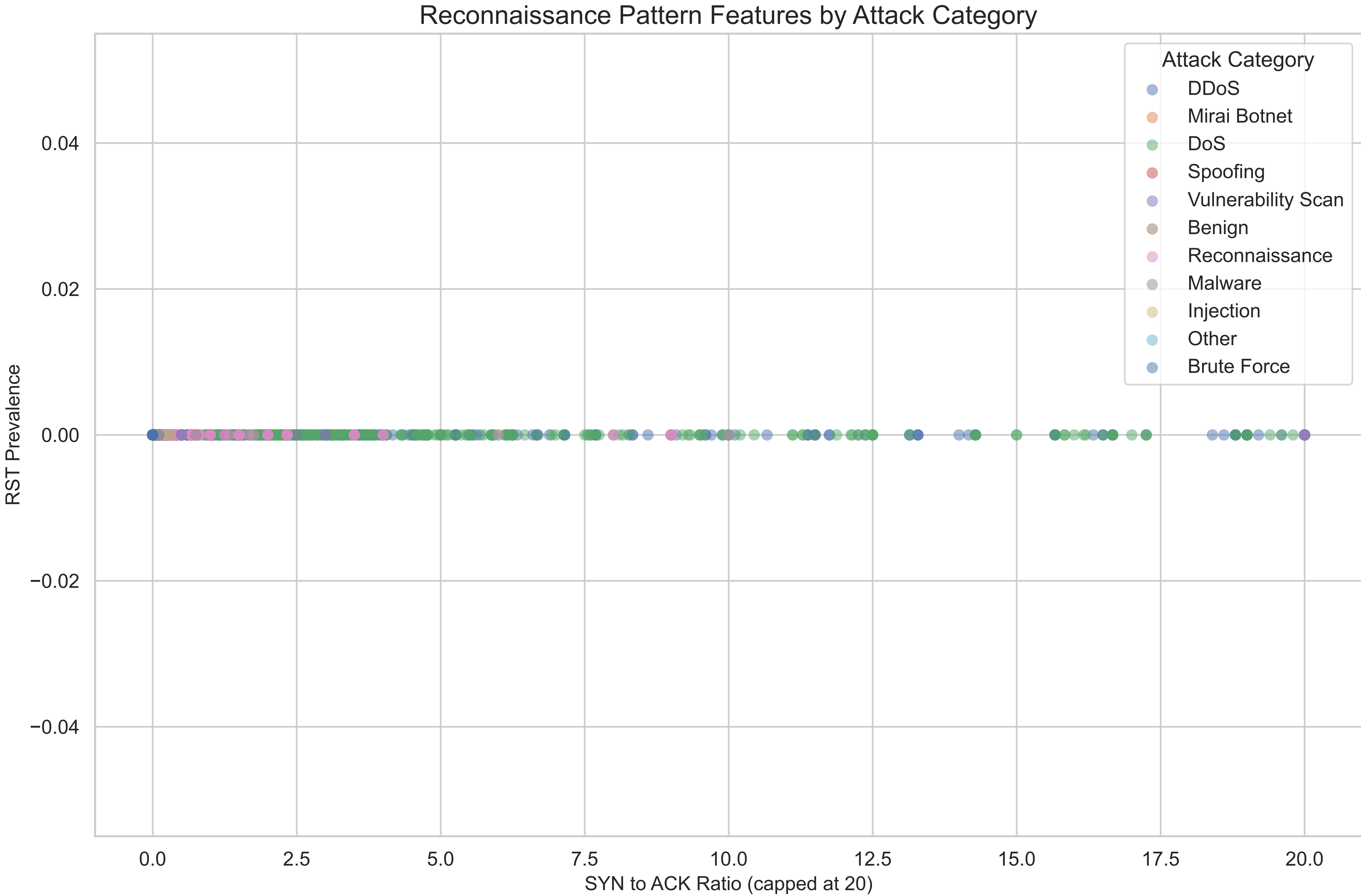
This figure visualizes the entropy of packet size distributions (Header_Length) across different attack categories.

Entropy measures the uncertainty or randomness in a distribution, with higher values indicating more variability. This feature is particularly valuable for distinguishing between attack types: DDoS and flooding attacks often show low entropy due to their repetitive, uniform packet sizes, while benign traffic typically has higher entropy reflecting diverse normal activities. Reconnaissance activities may show moderate entropy due to their structured probing patterns. By calculating entropy as a derived feature,

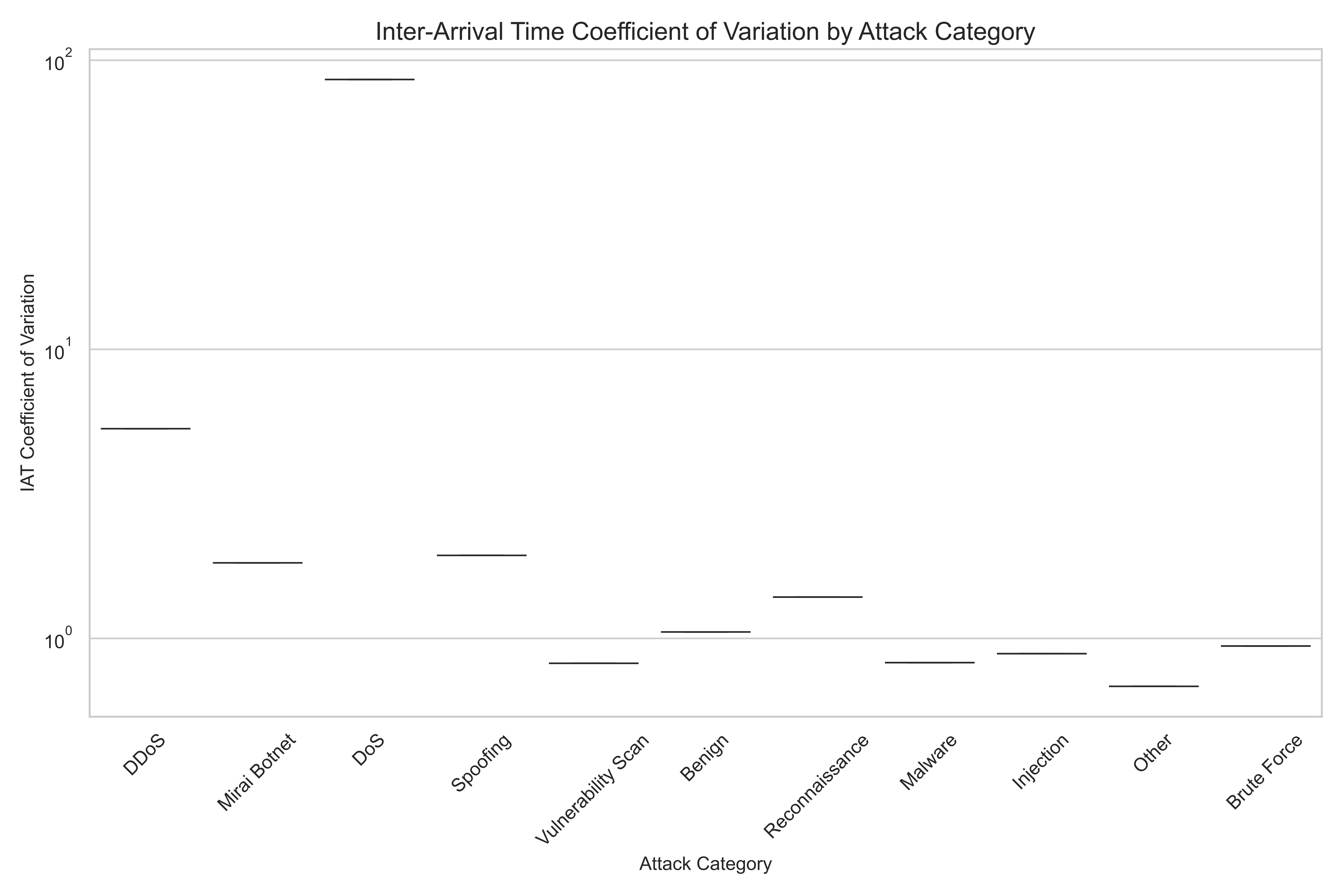
we capture complex distribution characteristics in a single value, which can significantly enhance detection accuracy. For SMEs, entropy-based features offer effective, lightweight indicators of suspicious network behavior patterns that might be missed by simpler threshold-based approaches.


Protocol Usage Percentage by Attack Category
Attack Category
Protocol Type
0
1
6
17
47
Benign 0.4 0.0 90.4 9.2 0.0
Brute Force 0.0 0.0 53.8 46.2 0.0
DDoS 0.0 23.1 60.4 16.6 0.0
DoS 0.0 0.1 59.2 40.7 0.0
Injection 0.0 0.0 86.7 13.3 0.0
Malware 0.0 0.0 100.0 0.0 0.0
Mirai Botnet 0.0 0.0 0.5 34.4 65.1
Other 0.0 0.0 100.0 0.0 0.0
Reconnaissance 2.2 0.0 88.8 9.0 0.0
Spoofing 0.4 0.2 68.8 30.6 0.0
Vulnerability Scan 1.0 0.0 78.1 20.9 0.0
100
80
60
40
20
0

This heatmap displays the protocol usage percentage across different attack categories, revealing distinctive protocol preferences for various attack types. The color intensity and numeric values represent the percentage of traffic using each protocol within an attack category. This derived feature transforms raw protocol counts into a normalized distribution that highlights behavioral patterns regardless of sample size differences. For example, certain DDoS attacks heavily favor specific protocols like UDP or ICMP, while reconnaissance activities predominantly use TCP. This protocol distribution profile serves as a powerful feature for attack classification, as it captures the fundamental behavior of different attack techniques. For SMEs, monitoring these protocol distributions provides an efficient way to detect deviations from established baselines without requiring deep packet inspection, making it suitable for deployment on resource-constrained monitoring systems.


Reconnaissance Pattern Features by Attack Category
RST Prevalence
SYN to ACK Ratio (capped at 20)
0.04
0.02
0.00
−0.02
−0.04
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Attack Category
DDoS
Mirai Botnet
DoS
Spoofing
Vulnerability Scan
Benign
Reconnaissance
Malware
Injection
Other
Brute Force

This scatter plot reveals the relationship between two key reconnaissance pattern indicators: the SYN-to-ACK ratio and RST flag prevalence across different attack categories. These derived features are particularly effective at identifying scanning and reconnaissance activities, which typically show distinctive patterns in TCP flag usage. Reconnaissance attacks exhibit higher SYN-to-ACK ratios, indicating many connection attempts with few completed handshakes, and often show elevated RST prevalence from responses to probes of closed ports. The clustering of points by attack category demonstrates the discriminative power of these features. For SMEs, these indicators provide early warning signs of potential attacks during the reconnaissance phase, allowing for preemptive defensive measures before more damaging attack phases begin. The features are computationally lightweight, making them suitable for continuous monitoring in resource-constrained environments.


Inter-Arrival Time Coefficient of Variation by Attack Category
IAT Coefficient of Variation
10^2
10^1
10^0
DDoS
Mirai Botnet
DoS
Spoofing
Vulnerability Scan
Benign
Reconnaissance
Malware
Injection
Other
Brute Force
Attack Category

This visualization displays the Coefficient of Variation (CV) of Inter-Arrival Times (IAT) across different attack categories. The CV measures the ratio of the standard deviation to the mean, providing a normalized measure of dispersion that captures the consistency or inconsistency of packet timing. This derived feature is particularly effective at distinguishing automated attack traffic (which often shows low variation due to programmatic generation) from human-generated or normal traffic (which typically shows higher, more natural variation). For example, DDoS attacks frequently exhibit very low CV values due to their regular, machine-generated packet patterns, while benign traffic shows higher variability. For SMEs, monitoring this temporal consistency metric provides an efficient way to detect automated attacks with minimal computational overhead, offering a robust indicator that complements traditional volume-based detection methods.

Summary of Engineered Features for IoT Security

Feature	Base Features	Effectiveness	Compute Cost	Description
Normalized Flow Duration	Time_To_Live	High for DoS, Medium for Reconnaissance	Low	Z-score normalized flow duration, improving scale for ML models
Packet Rate	N/A, Time_To_Live	Very High for DDoS/DoS, Low for Spoofing	Very Low	Number of packets per second, highlighting volumetric attacks
Packet Size Entropy	Header_Length	High for DDoS, High for Exfiltration	Medium	Information entropy of packet size distribution, detecting uniformity
Protocol Distribution	Protocol Type	High for Reconnaissance, Medium for DDoS	Low	Percentage distribution of protocols in traffic flow
DDoS Intensity Score	Multiple rate and size metrics	Very High for DDoS, Low for other attacks	Medium	Composite score optimized for DDoS detection incorporating multiple indicators
Reconnaissance Indicators	SYN, ACK, RST flags	Very High for Scanning, Low for DDoS	Low	Flag usage patterns indicative of scanning and reconnaissance
IAT Coefficient of Variation	IAT	High for Automated Attacks, Medium for Manual Attacks	Medium	Measures consistency of packet timing, distinguishing automated attacks

This summary table presents the engineered features developed for IoT security threat detection, outlining their base components, effectiveness for different attack types, computational cost, and descriptions. These derived features transform raw network traffic data into more discriminative indicators optimized for specific attack detection. The computational cost assessment is particularly relevant for SME environments with limited computing resources, helping organizations prioritize which features to implement. Features like packet rate and protocol distribution offer excellent detection capabilities with minimal overhead, making them suitable for all SME deployments. More complex features like entropy calculations provide enhanced detection at moderate computational cost, appropriate for medium-sized deployments. This framework allows SMEs to select feature engineering approaches scaled to their specific resource constraints and security needs, enabling effective threat detection even in environments with limited monitoring infrastructure.