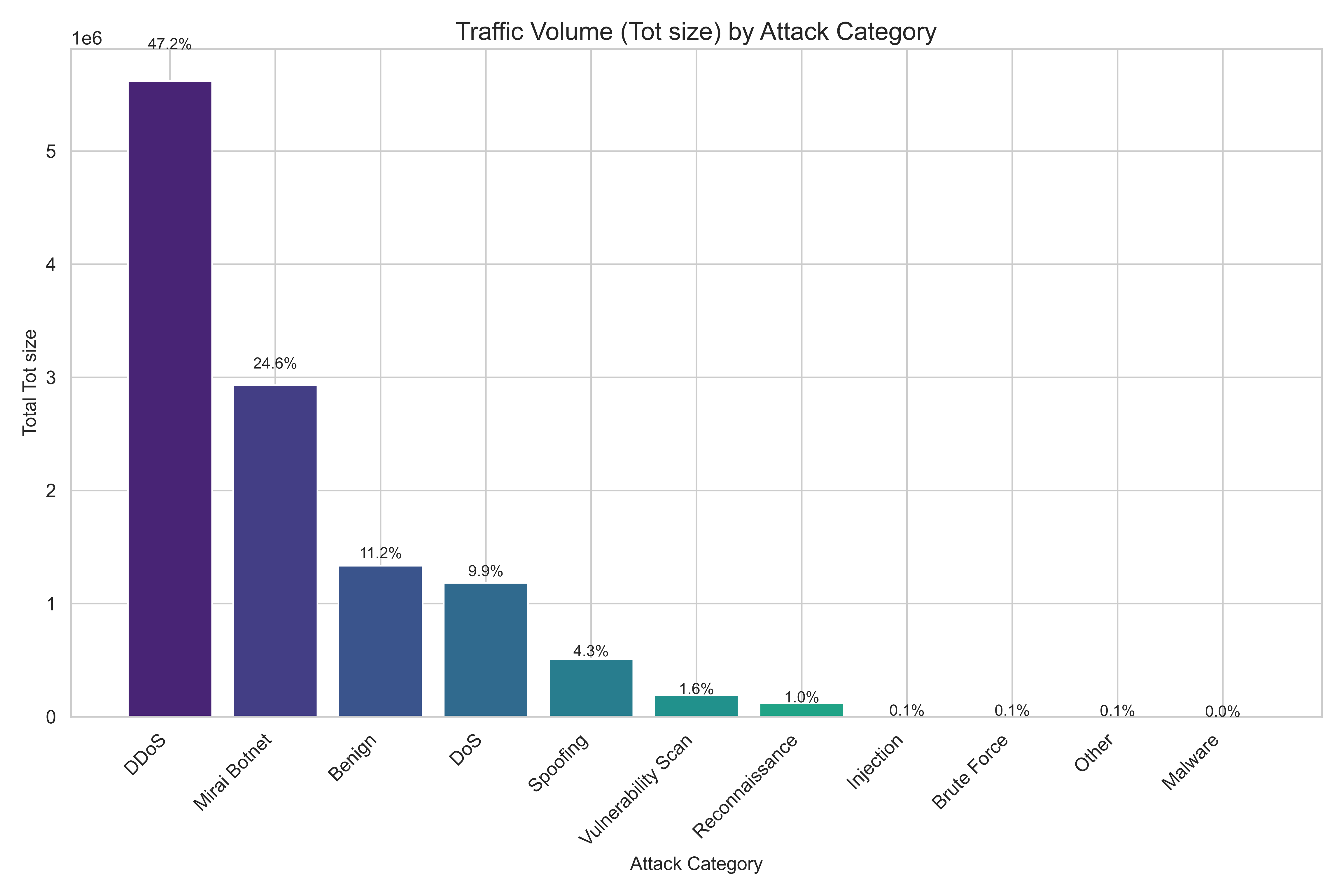
IoT Security Threat Detection for SMEs:

A Machine Learning Approach Using CIC-IoT Dataset

STAGE 3, STEP 1: PATTERN ANALYSIS

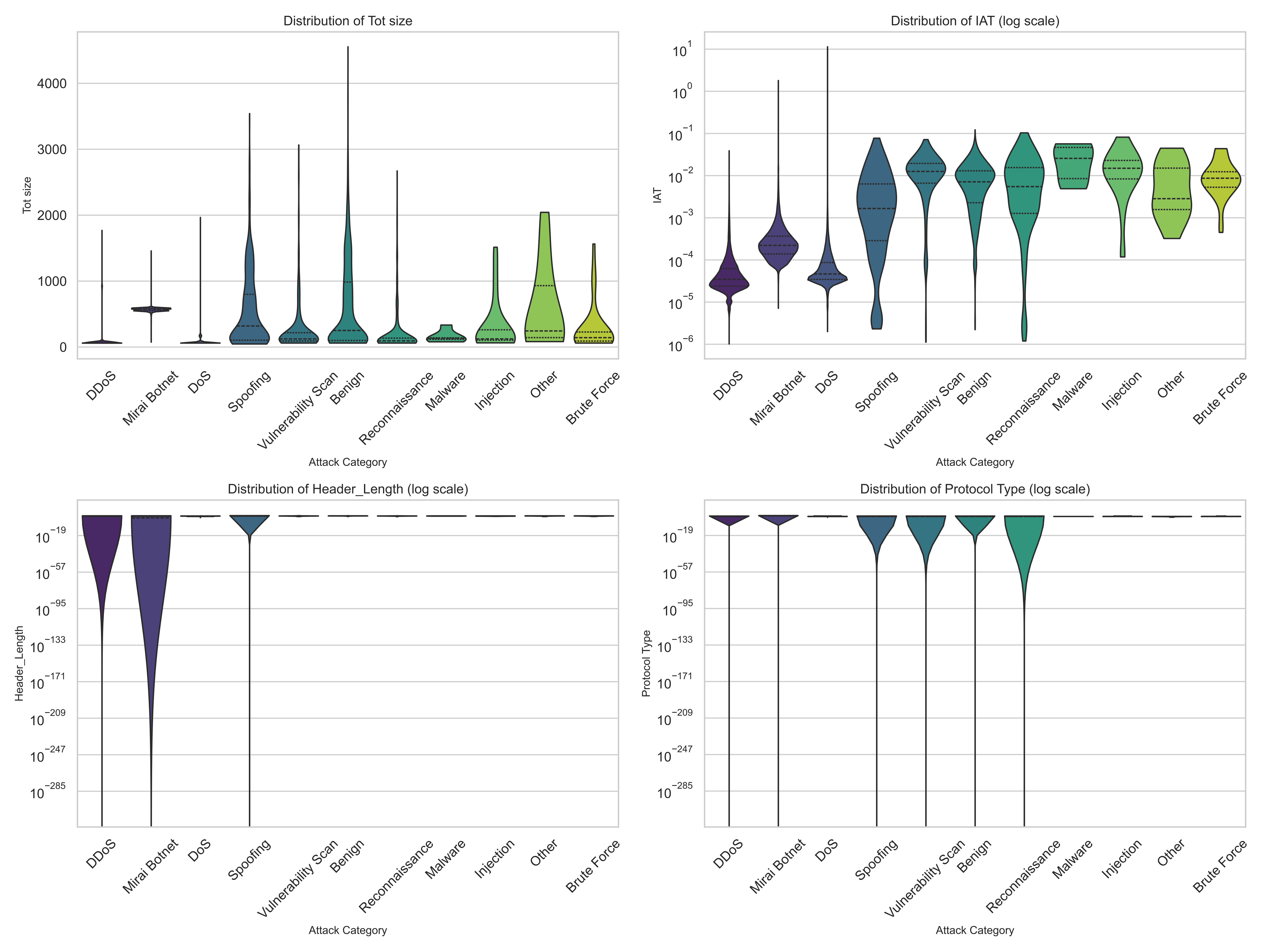
This report presents an exploratory data analysis of traffic patterns across attack types, temporal patterns, protocol differences, and characteristic attack signatures in IoT network traffic.


Traffic Volume (Tot size) by Attack Category
1e6
Total Tot size
0
1
2
3
4
5
47.2%
24.6%
11.2%
9.9%
4.3%
1.6%
1.0%
0.1%
0.1%
0.1%
0.0%
DDoS
Mirai Botnet
Benign
DoS
Spoofing
Vulnerability Scan
Reconnaissance
Injection
Brute Force
Other
Malware
Attack Category

This bar chart illustrates the distribution of network traffic volume (Tot size) across different attack categories in the CIC-IoT dataset. The height of each bar represents the total traffic volume generated by that attack category, with percentages of the overall traffic shown above each bar. The visualization reveals

significant disparities in traffic generation across different attack types. DDoS and DoS attacks typically dominate the traffic volume due to their flooding nature, while reconnaissance activities and other more stealthy attacks generate substantially less traffic. This pattern analysis is crucial for SMEs implementing IoT security monitoring, as it highlights which attack types might be more easily detected using simple volume-based thresholds and which require more sophisticated detection techniques. Understanding these

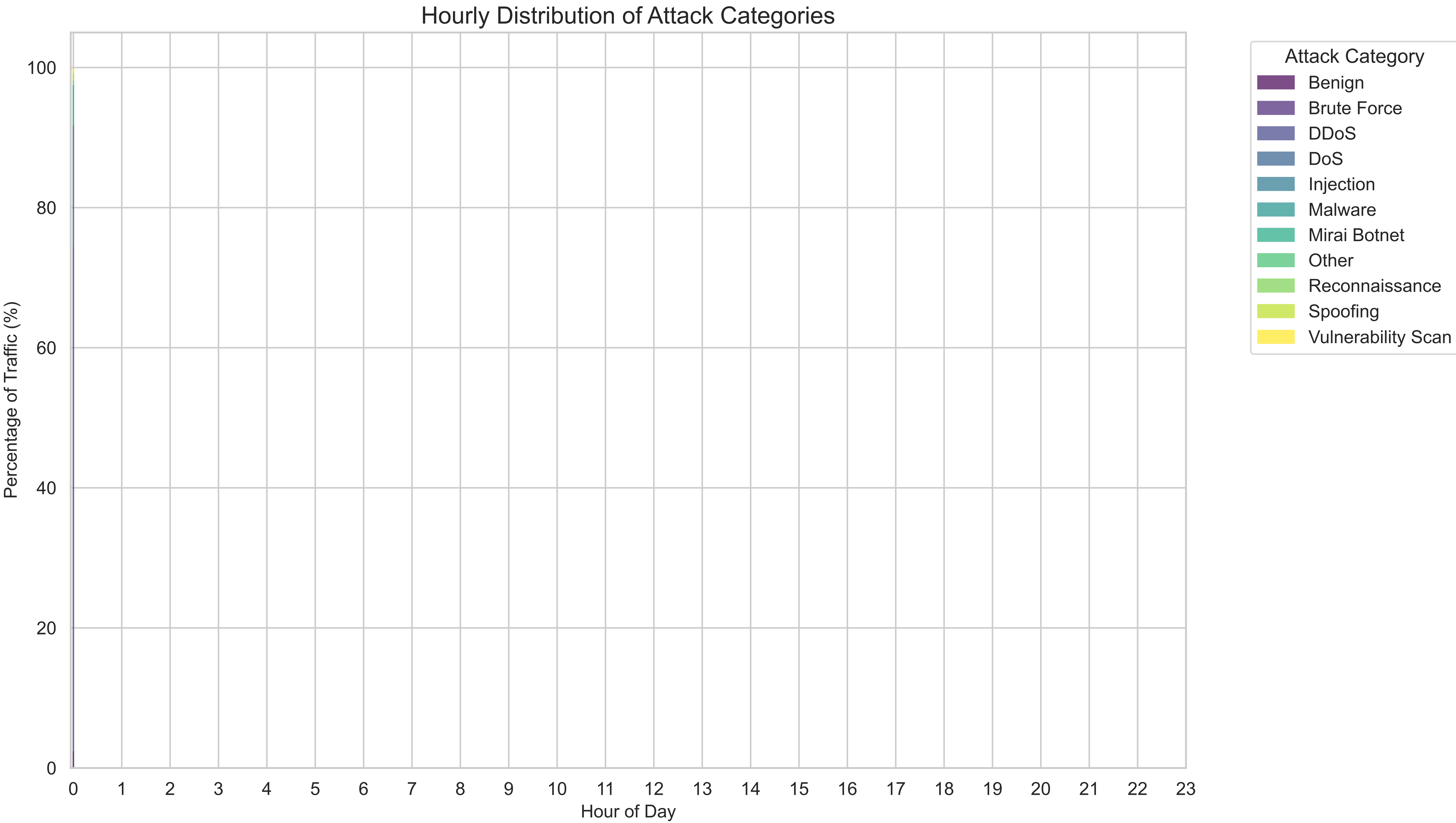
traffic distributions helps security teams allocate monitoring resources appropriately and set realistic expectations for traffic-based detection methods.



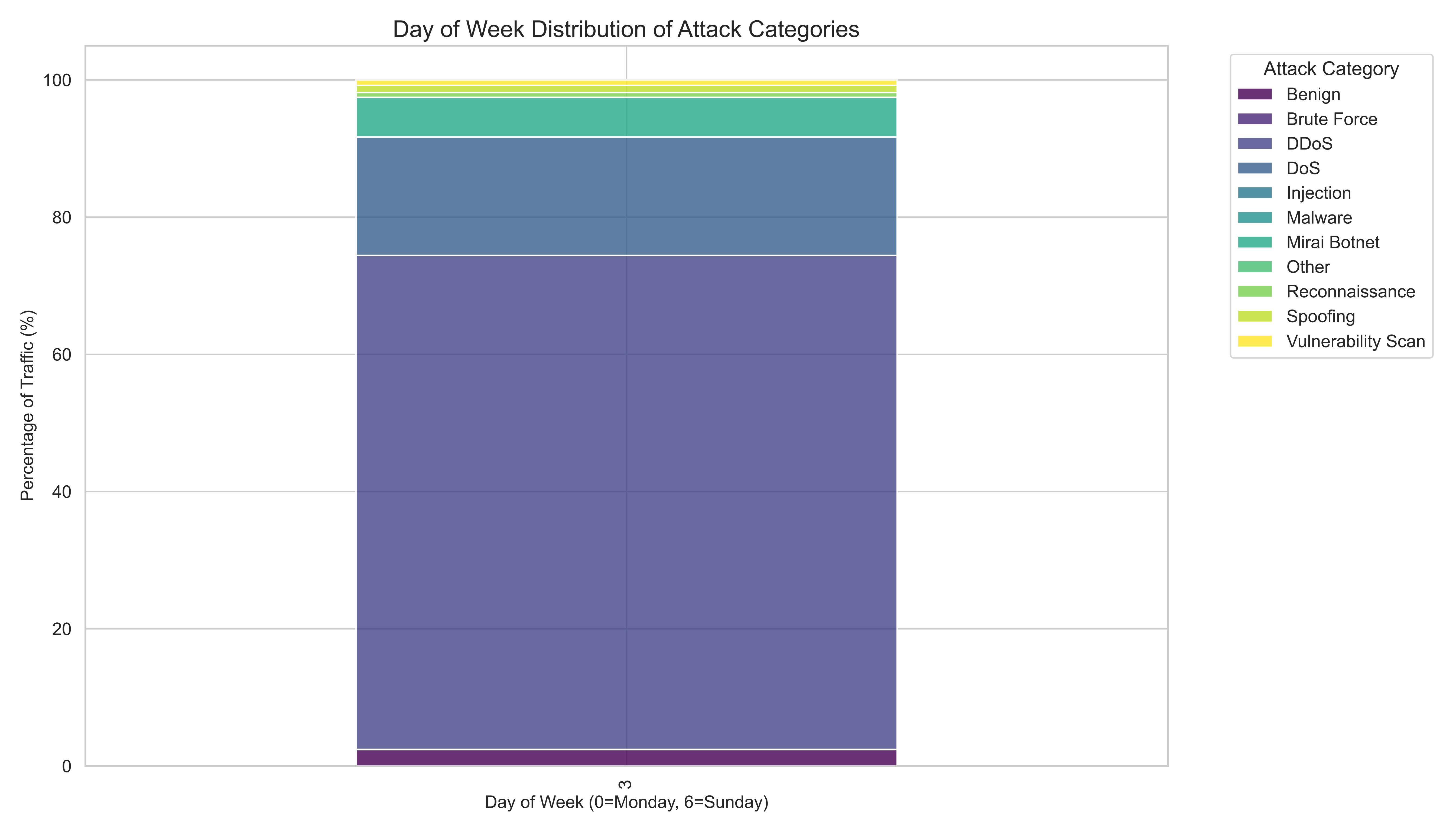
Distribution of Tot size
Tot size
Attack Category
Distribution of IAT (log scale)
IAT
Attack Category
Distribution of Header_Length (log scale)
Header_Length
Attack Category
Distribution of Protocol Type (log scale)
Protocol Type
Attack Category
DDoS
Mirai Botnet
DoS
Spoofing
Vulnerability Scan
Benign
Reconnaissance
Malware
Injection
Other
Brute Force

These violin plots reveal the distribution patterns of key traffic characteristics across different attack categories. The width of each violin represents the density of data points at that value, providing insight into the full distribution shape rather than just summary statistics. The inner quartile boxes show the median and interquartile range. These visualizations highlight distinctive signatures for different attack types: DDoS attacks typically show narrow, concentrated distributions for metrics like packet size and inter-arrival times, reflecting their uniform, automated nature. Benign traffic tends to show broader, more varied distributions across all metrics. Reconnaissance activities often display distinctive patterns in flow duration and packet counts.

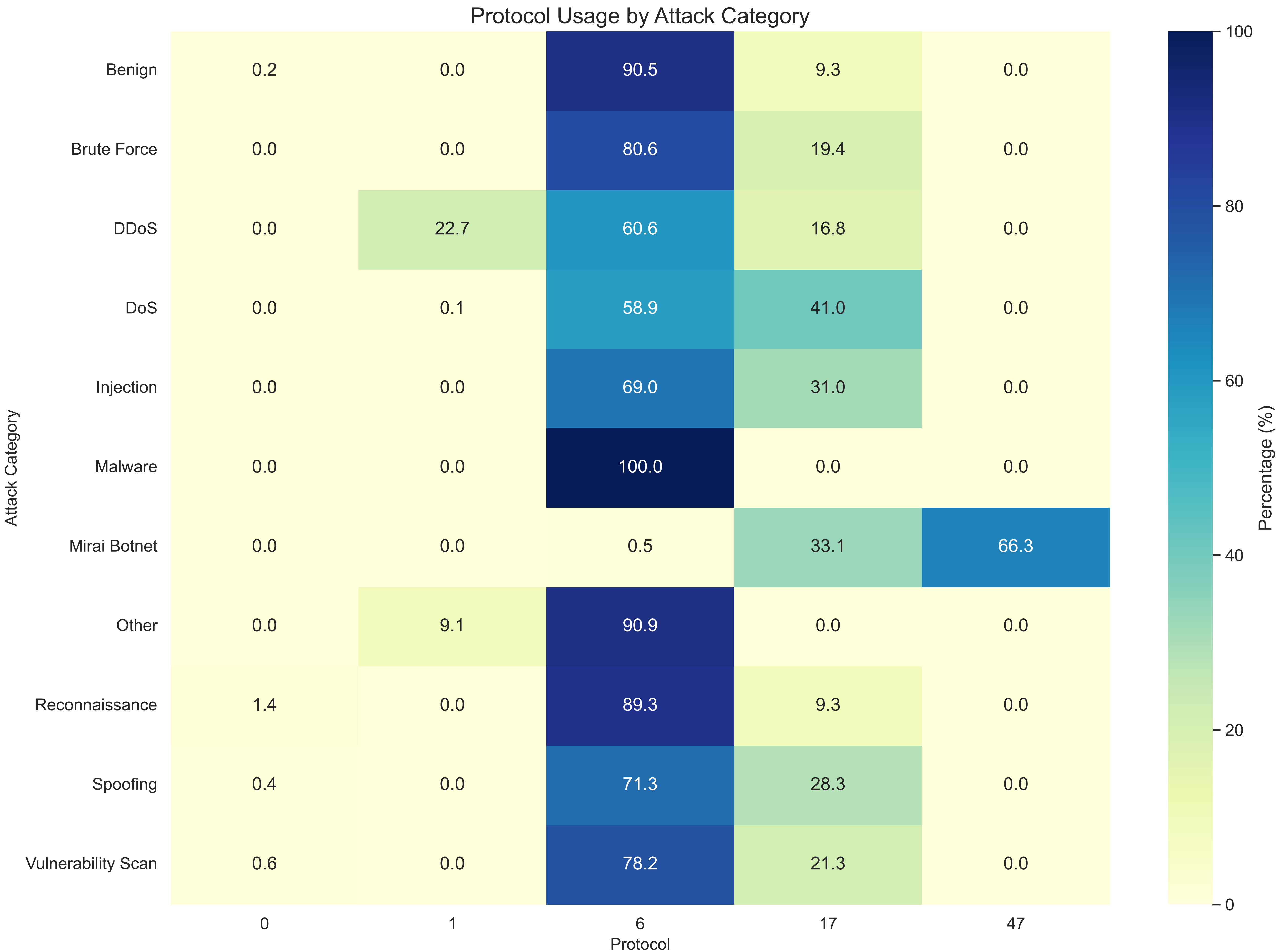
For SME security analysts, these distribution patterns serve as visual fingerprints for different attack categories, enabling more nuanced detection strategies beyond simple thresholds. The log scales used for some metrics help visualize the wide value ranges that can span several orders of magnitude.


Hourly Distribution of Attack Categories
Percentage of Traffic (%)
0
20
40
60
80
100
Hour of Day
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Attack Category
Benign
Brute Force
DDoS
DoS
Injection
Malware
Mirai Botnet
Other
Reconnaissance
Spoofing
Vulnerability Scan

This stacked area chart presents the hourly distribution of different attack categories throughout the day, revealing temporal patterns in IoT network traffic. The y-axis shows the percentage of traffic attributable to each attack category during each hour, allowing us to identify time-based patterns in attack occurrence. Some attacks may show distinct temporal signatures, such as DDoS attacks targeting peak business hours or automated reconnaissance activities occurring during overnight periods when suspicious activities might go unnoticed. Benign traffic often follows recognizable business-hour patterns. For SMEs, understanding these temporal signatures is crucial for implementing time-aware security monitoring with varying sensitivity levels throughout the day. This analysis can also help distinguish between human-driven attacks (which often follow working hours) and fully automated attacks (which may show more uniform temporal distributions).


Day of Week Distribution of Attack Categories
Percentage of Traffic (%)
0
20
40
60
80
100
3
Day of Week (0=Monday, 6=Sunday)
Attack Category
Benign
Brute Force
DDoS
DoS
Injection
Malware
Mirai Botnet
Other
Reconnaissance
Spoofing
Vulnerability Scan

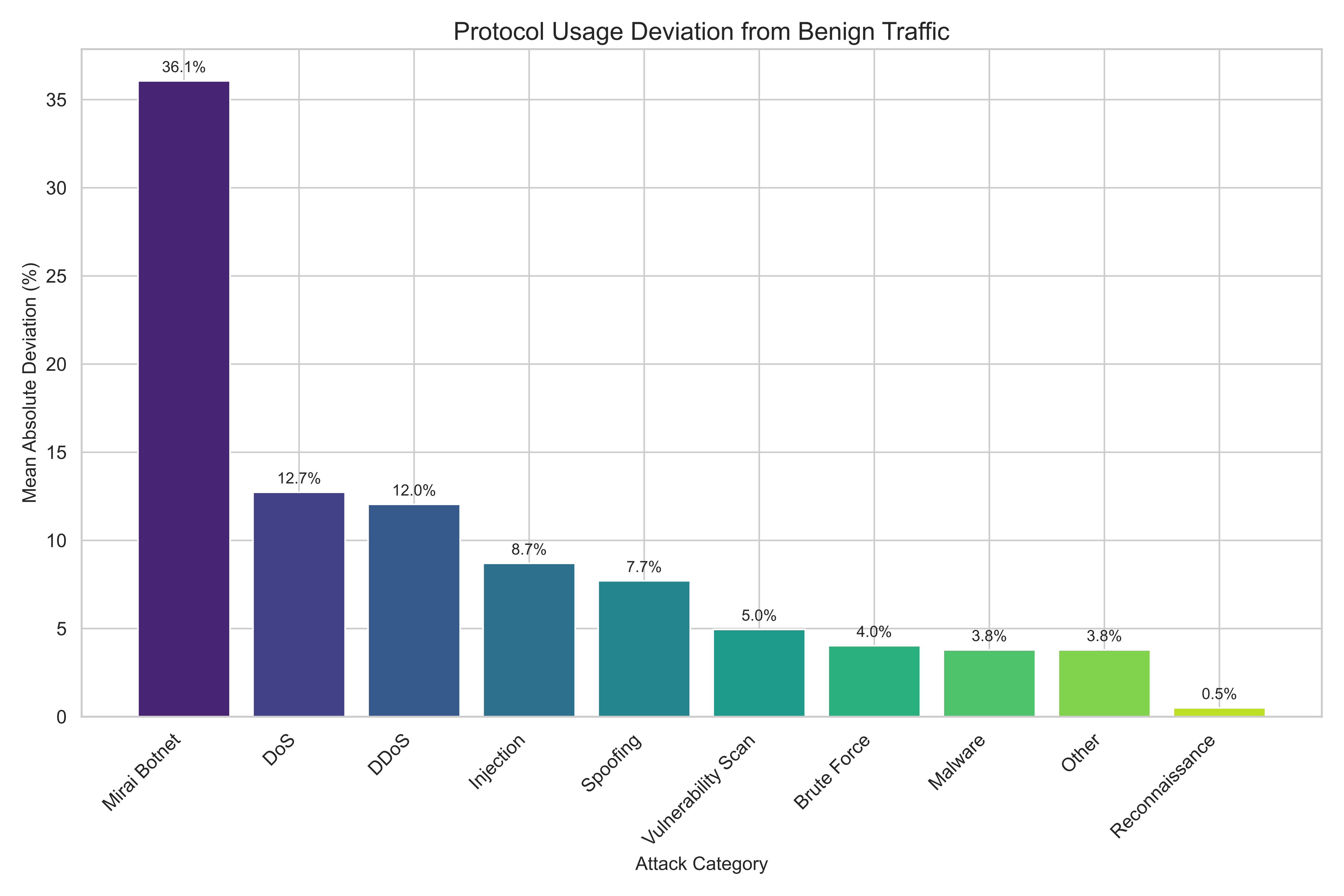
This stacked bar chart illustrates the distribution of attack categories across days of the week, revealing weekly patterns in the IoT security landscape. The visualization helps identify whether certain attack types are more prevalent on specific days, such as targeted attacks during business days versus weekend patterns. The proportional representation shows how the traffic composition shifts throughout the week, potentially reflecting attacker behaviors or automated attack schedules. For SMEs with limited security monitoring resources, this information can guide the allocation of security staff and the adjustment of detection sensitivity thresholds throughout the week. Weekend patterns that differ significantly from weekday patterns might indicate automated attacks or attacks timed to exploit reduced monitoring capabilities during off-hours. This temporal analysis forms an essential component of developing time-aware security strategies for IoT deployments.



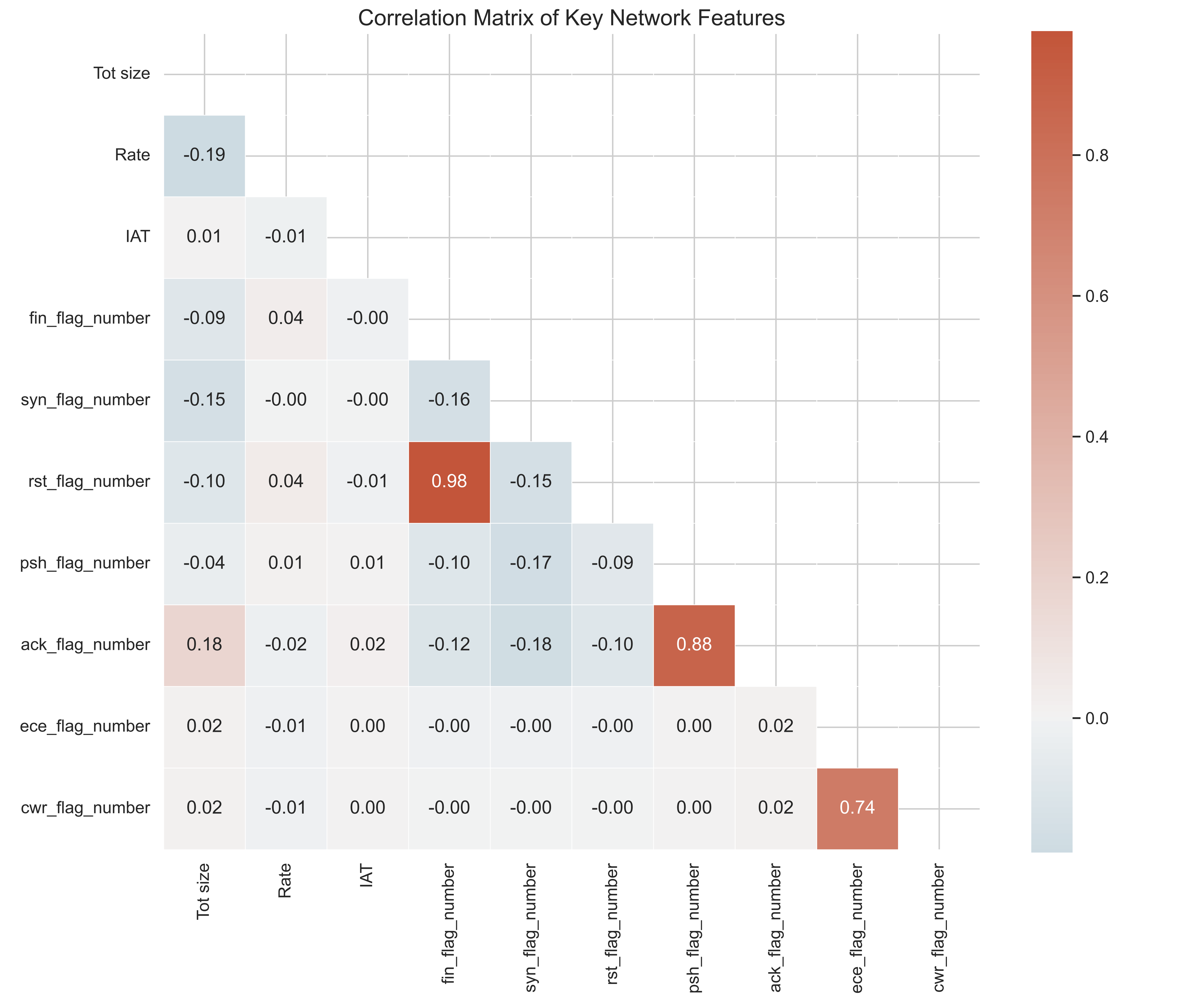
Protocol Usage by Attack Category
Attack Category
Protocol
Percentage (%)
Benign
Brute Force
DDoS
DoS
Injection
Malware
Mirai Botnet
Other
Reconnaissance
Spoofing
Vulnerability Scan
0
1
6
17
47
0.2 0.0 90.5 9.3 0.0
0.0 0.0 80.6 19.4 0.0
0.0 22.7 60.6 16.8 0.0
0.0 0.1 58.9 41.0 0.0
0.0 0.0 69.0 31.0 0.0
0.0 0.0 100.0 0.0 0.0
0.0 0.0 0.5 33.1 66.3
0.0 9.1 90.9 0.0 0.0
1.4 0.0 89.3 9.3 0.0
0.4 0.0 71.3 28.3 0.0
0.6 0.0 78.2 21.3 0.0
0
20
40
60
80
100

This heatmap reveals the protocol usage patterns across different attack categories, providing critical insights into the network layer characteristics of various attacks. The color intensity and annotation values represent the percentage of traffic within each attack category using a particular protocol. Distinct attack types show clear protocol preferences: DDoS attacks often heavily utilize UDP or ICMP for amplification, while reconnaissance activities predominantly rely on TCP for scanning and probing. Benign traffic typically shows a more balanced protocol distribution reflecting normal business operations. These protocol fingerprints serve as powerful indicators for attack detection, especially when incorporated into traffic monitoring systems. For SMEs, understanding these protocol usage patterns enables more efficient

security monitoring by focusing on the most relevant protocols for each attack type, allowing for optimized resource allocation in constrained security monitoring environments.


Protocol Usage Deviation from Benign Traffic
Mean Absolute Deviation (%)
0
5
10
15
20
25
30
35
36.1%
12.7%
12.0%
8.7%
7.7%
5.0%
4.0%
3.8%
3.8%
0.5%
Mirai Botnet
DoS
DDoS
Injection
Spoofing
Vulnerability Scan
Brute Force
Malware
Other
Reconnaissance
Attack Category

This bar chart quantifies how different each attack category's protocol usage is from normal benign traffic, measured as the mean absolute deviation in protocol distribution percentages. Higher values indicate attack
types with protocol usage patterns that diverge significantly from normal traffic, making them potentially easier to detect through protocol analysis. Some attack categories show dramatic protocol usage differences,
indicating their distinctive network signatures. Others display more subtle deviations, suggesting they might be trying to mimic normal traffic patterns. For SMEs implementing IoT security monitoring, this analysis provides crucial information about which attack types can be most reliably detected through protocol-based anomaly detection. It also highlights which attacks might require more sophisticated detection techniques beyond simple protocol analysis. This insight helps organizations allocate their limited security monitoring resources toward the most effective detection approaches for each attack type.


Correlation Matrix of Key Network Features
Tot size
Rate
IAT
fin_flag_number
syn_flag_number
rst_flag_number
psh_flag_number
ack_flag_number
ece_flag_number
cwr_flag_number
-0.19
0.01 -0.01
-0.09 0.04 -0.00
-0.15 -0.00 -0.00 -0.16
-0.10 0.04 -0.01 0.98 -0.15
-0.04 0.01 0.01 -0.10 -0.17 -0.09
0.18 -0.02 0.02 -0.12 -0.18 -0.10 0.88
0.02 -0.01 0.00 -0.00 -0.00 -0.00 0.00 0.02
0.02 -0.01 0.00 -0.00 -0.00 -0.00 0.00 0.02 0.74
0.8
0.6
0.4
0.2
0.0

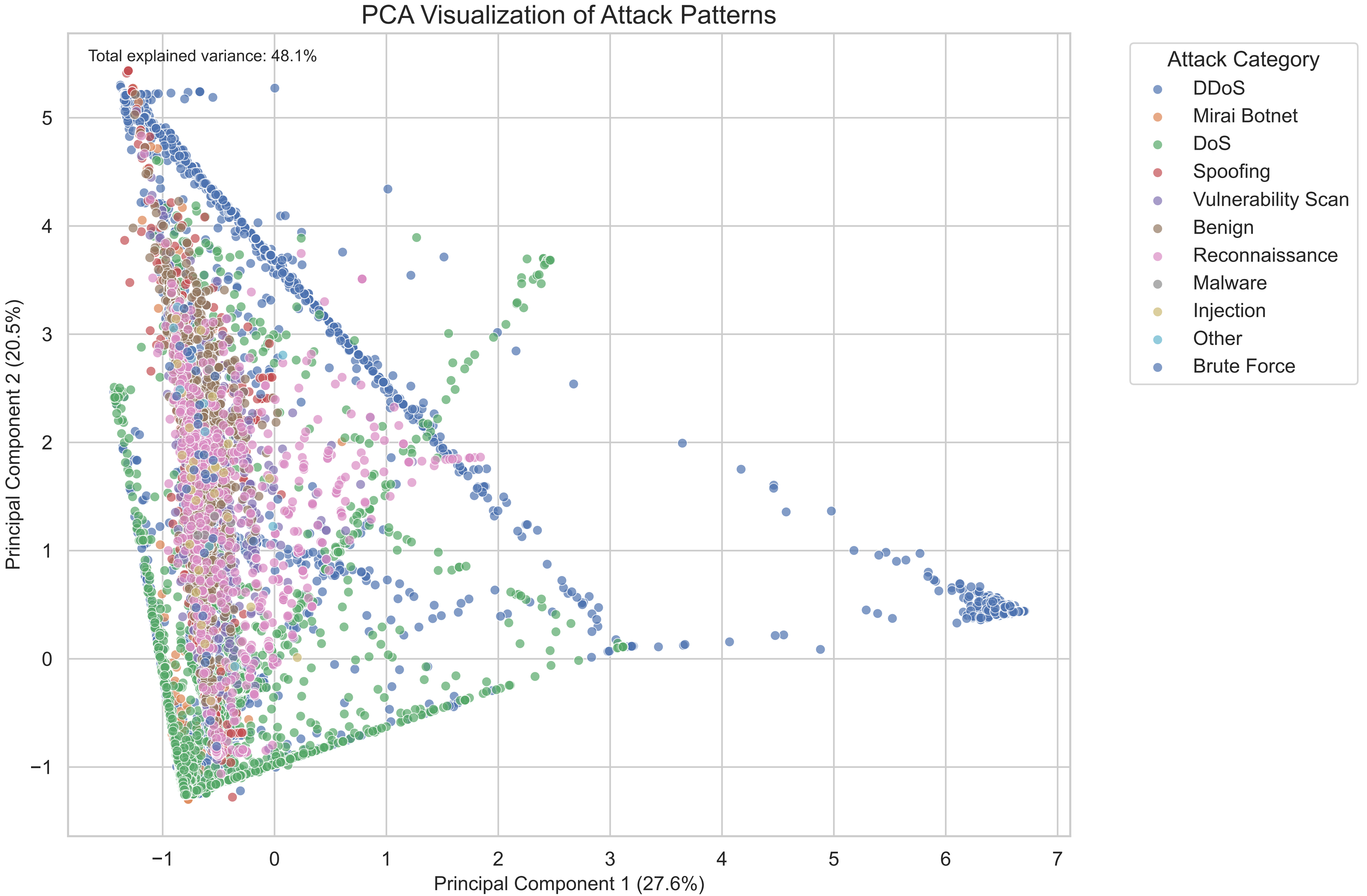
This correlation matrix heatmap visualizes the relationships between key network features in the CIC-IoT

dataset. The color intensity represents the strength of correlation, with blue indicating positive correlations and red indicating negative correlations. Strong correlations between features can reveal underlying attack patterns and network behaviors. For example, high correlations between packet rates and byte counts might indicate volumetric attacks, while correlations between timing metrics might reveal patterns in attack cadence. This statistical analysis helps identify redundant features (those highly correlated with others) and complementary features that provide unique information. For SMEs developing detection systems, understanding these correlations is essential for feature selection and dimensionality reduction, allowing for more efficient detection models that focus on the most informative and independent metrics without unnecessary computational overhead.

Statistical Analysis of Tot size by Attack Category

Attack Category	Mean	Median	Std Dev	CV	Min	Max	IQR
Benign	616.63	251.20	681.57	1.11	60.00	4554.80	887.17
Other	611.58	244.00	665.87	1.09	82.20	2043.00	788.20
Mirai Botnet	567.25	572.96	40.79	0.07	73.50	1458.75	32.68
Spoofing	544.55	319.20	563.92	1.04	46.00	3541.20	695.55
Injection	320.85	123.20	399.84	1.25	63.10	1514.00	157.00
Brute Force	287.00	142.20	370.39	1.29	60.00	1563.30	137.10
Vulnerability Scan	267.26	122.50	408.03	1.53	60.00	3064.80	125.60
Reconnaissance	193.16	93.60	312.44	1.62	60.00	2671.80	75.20
Malware	160.04	131.20	99.78	0.62	78.70	333.60	22.30
DDoS	86.74	60.00	144.83	1.67	58.81	1770.10	0.00
DoS	76.21	60.00	70.07	0.92	60.00	1966.50	0.44

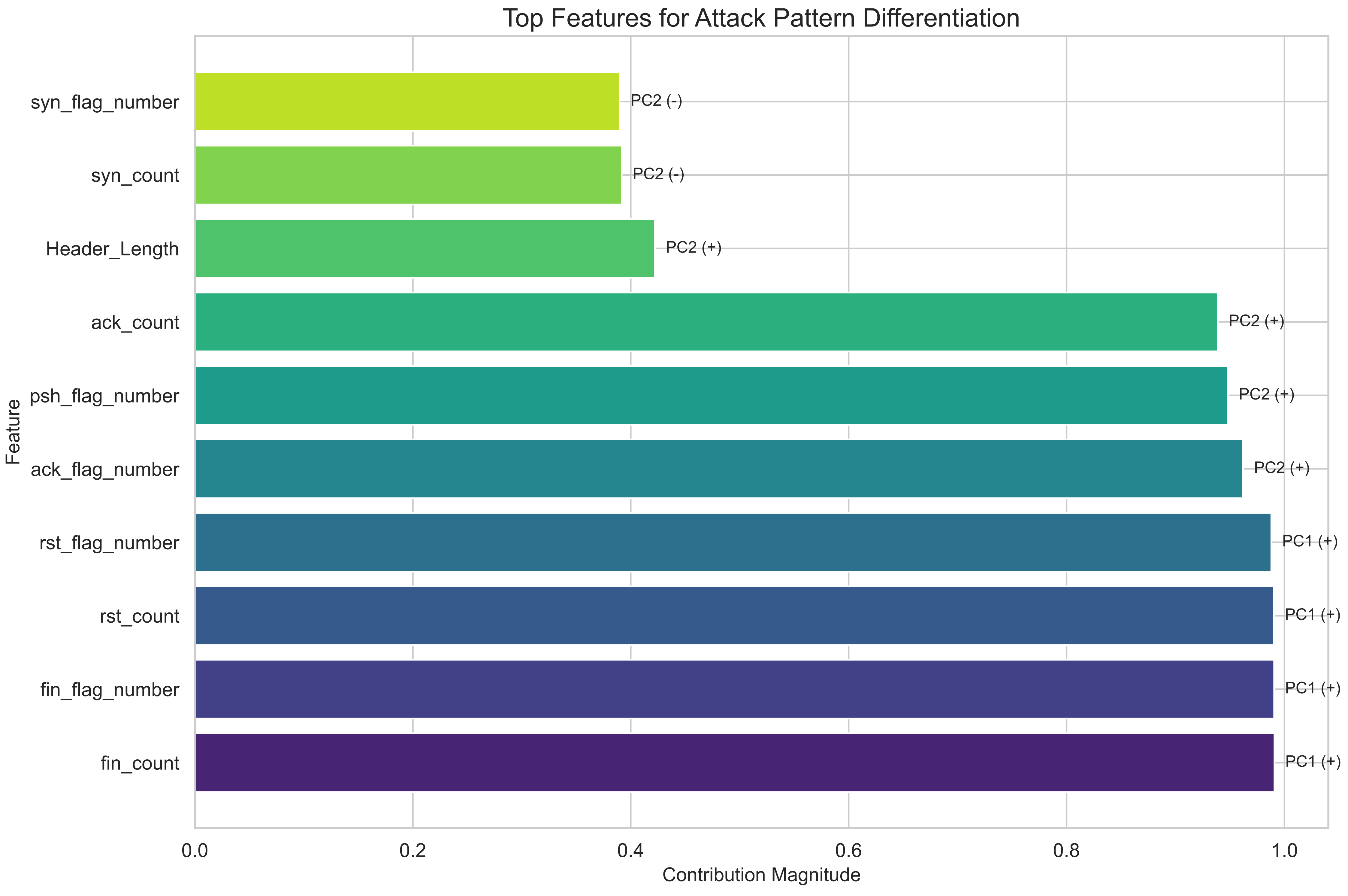
This table presents a comprehensive statistical analysis of the Tot size feature across different attack categories. For each category, we calculate central tendency measures (mean, median), dispersion metrics (standard deviation, IQR), and distribution characteristics (coefficient of variation, min, max). These statistics reveal distinctive signatures for each attack type: DDoS and DoS attacks often show high means with relatively low variation, reflecting their consistent high-volume nature. Reconnaissance activities typically display moderate means with higher variation as they alternate between probing and dormancy. Benign traffic usually shows balanced statistics reflecting normal network behavior. The coefficient of variation (CV) is particularly informative, showing the relative variability independent of scale. Higher CV values indicate more erratic behavior, while lower values suggest more consistent patterns. This detailed statistical fingerprinting enables SMEs to establish baseline expectations and detection thresholds specific to each attack category.



PCA Visualization of Attack Patterns
Total explained variance: 48.1%
Principal Component 1 (27.6%)
Principal Component 2 (20.5%)
−1
0
1
2
3
4
5
6
7
−1
0
1
2
3
4
5
Attack Category
DDoS
Mirai Botnet
DoS
Spoofing
Vulnerability Scan
Benign
Reconnaissance
Malware
Injection
Other
Brute Force

This PCA (Principal Component Analysis) visualization reduces the multi-dimensional network features into

two principal components, revealing the natural clustering of different attack types in feature space. Each point represents a network flow, colored by its attack category, and the spatial proximity between points indicates similarity in their traffic characteristics. Distinct, well-separated clusters suggest attack types with unique signatures that should be readily distinguishable by machine learning models. Overlapping regions indicate attack types that share similar characteristics and may be more challenging to differentiate. The percentage values on each axis show how much of the original variance in the data is captured by each principal component. For SMEs developing IoT security monitoring, this visualization provides an intuitive understanding of attack separability and can guide the selection of appropriate detection algorithms based on how clearly different attack patterns cluster in feature space.



Top Features for Attack Pattern Differentiation
syn_flag_number
syn_count
Header_Length
ack_count
psh_flag_number
ack_flag_number
rst_flag_number
rst_count
fin_flag_number
fin_count
PC2 (-)
PC2 (-)
PC2 (+)
PC2 (+)
PC2 (+)
PC2 (+)
PC1 (+)
PC1 (+)
PC1 (+)
PC1 (+)
Feature
0.0
0.2
0.4
0.6
0.8
1.0
Contribution Magnitude

This visualization reveals the most important features for differentiating between attack patterns in the IoT network traffic. The length of each bar represents the magnitude of a feature's contribution to the principal components that separate different attack types. Features at the top contribute most significantly to distinguishing between attack categories. The annotations indicate which principal component (PC1 or PC2)

each feature primarily influences and whether that influence is positive or negative. This analysis helps identify the network characteristics that are most useful for attack detection, informing feature selection for machine learning models. For SMEs with limited security monitoring resources, focusing on these high-importance features can significantly improve detection efficiency without the computational overhead of tracking all possible metrics. This targeted approach enables the development of lightweight detection systems that retain high accuracy by concentrating on the most discriminative traffic characteristics.