Intelligent IoT Attack Detection Design via ODLLM with Feature Ranking-based Knowledge Base

Satvik Verma[1], Qun Wang[1], E. Wes Bethel[1,2]

[1]Department of Computer Science, San Francisco State University, San Francisco, CA, 94132
[2]Lawrence Berkeley National Laboratory, Berkeley CA, 94720
sverma4@sfsu.edu, claudqunwang@ieee.org, ewbethel@sfsu.edu

Abstract

The widespread adoption of Internet of Things (IoT) devices has introduced significant cybersecurity challenges, particularly with the increasing frequency and sophistication of Distributed Denial of Service (DDoS) attacks. Traditional machine learning (ML) techniques often fall short in detecting such attacks due to the complexity of blended and evolving patterns. To address this, we propose a novel framework leveraging On-Device Large Language Models (ODLLMs) augmented with fine-tuning and knowledge base (KB) integration for intelligent IoT network attack detection. By implementing feature ranking techniques and constructing both long and short KBs tailored to model capacities, the proposed framework ensures efficient and accurate detection of DDoS attacks while overcoming computational and privacy limitations. Simulation results demonstrate that the optimized framework achieves superior accuracy across diverse attack types, especially when using compact models in edge computing environments. This work provides a scalable and secure solution for real-time IoT security, advancing the applicability of edge intelligence in cybersecurity.

Introduction

The proliferation of IoT sensors in both residential and industrial environments has led to the generation of vast amounts of data that require timely and effective processing to facilitate rapid decision-making (Zhang et al. 2023). However, IoT sensors are particularly vulnerable to various cyber-attacks, especially DoS and Distributed DDoS attacks. These attacks can cause significant losses and further harm, making the quick and accurate identification of such threats critically important (Jaton, Gyawali, and Qian 2023).

ML algorithms have been extensively used to detect abnormal DDoS traffic. The authors in (Hussain et al. 2020) proposed a methodology to convert the network traffic data into image form and trained a CNN model for DDoS detection. (Jia et al. 2020) presented a DDoS attack detection algorithm based on traffic variations and LSTM and CNN models. The authors in (Aysa, Ibrahim, and Mohammed 2020) utilized LSVM, Neural Network, and Decision tree to detect abnormal activities such as DDOS features. Traditional ML algorithms rely on large datasets for training and face numerous limitations in this context. Moreover, when multiple attack types are mixed together, these algorithms often struggle to perform effectively.

The emergence of large models has shown promise in the real-time and accurate identification of various abnormal network traffic patterns (Zhu et al. 2024). Nonetheless, these models typically require substantial computational resources for training and deployment, consuming significant amounts of computing power and electricity. Moreover, utilizing third-party models introduces data privacy and security concerns, which are particularly pertinent in network attack defense scenarios (Yao et al. 2024). In distributed, large-scale IoT networks, such models often fail to meet user needs promptly due to their resource-intensive nature and potential latency issues. This has led to increased interest in edge computing paradigms, where edge intelligence and on-device large models have garnered significant attention from researchers (Xu et al. 2024a) (Chen, Li, and Ma 2024). By employing techniques like model pruning and compression, smaller models can deliver functionalities comparable to their larger counterparts in most situations (Xu et al. 2024a) (Chen et al. 2024). Building upon this, developing applications on ODLLMs can ensure affordable intelligent decision-making and local data processing (Xu et al. 2024b). However, when applying ODLLMs to network attack detection, a lack of necessary background knowledge within the models necessitates fine-tuning and the integration of knowledge base (KB) to enhance their performance.

Therefore, we propose a novel system that leverages ODLLMs augmented with KB assistance to improve the detection accuracy of DDoS attacks in IoT environments. Our system addresses the challenges of computational resource constraints and privacy concerns by enabling on-device processing. We demonstrate that with appropriate KB support, ODLLMs can achieve performance comparable to larger models while operating within the limitations of edge devices. The contributions of this paper are as follows:

A novel DDoS attack detection system that utilizes ODLLMs for IoT environments is proposed. We first introduce a novel approach for feature prioritization using a Random Forest Regressor (RFR) to rank the most critical features for different DDoS attack types, enabling the construction of efficient and scalable KBs. We then address the limitations of smaller ODLLMs by designing a simplified KB

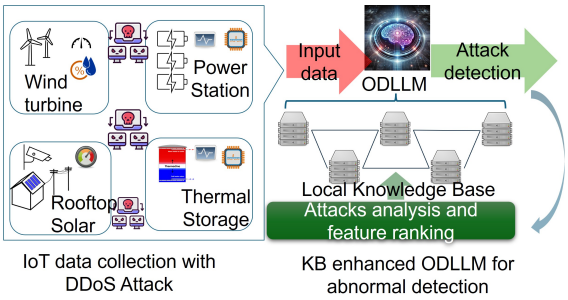

Figure 1: System model.

that retains only high-impact features, significantly improving predictive performance while reducing computational overhead. Third, we demonstrate the effectiveness of our framework through extensive experiments on the CICIoT 2023 dataset, achieving high accuracy in detecting various DDoS attack types. Our results highlight the critical role of tailored KB designs in enabling efficient and accurate attack detection on resource-constrained edge devices, paving the way for scalable and secure IoT network solutions.

The subsequent sections are organized in the following manner. The system model and problem formulation are presented in Section II. The proposed feature ranking-based KB design is developed in Section III. The findings of the simulation are presented in Section IV. Finally, Section V provides the concluding remarks for this paper.

System Model and Problem Formulation

System Model

As shown in Fig. 1, the framework operates in three primary stages:

(1) Feature Analysis and Ranking: The system begins by extracting key features from the historical records of DDoS traffic data. These features include important values such as protocol types, packet rates, inter-arrival times, TCP flags, and statistical metrics like average packet size and variance. By analyzing these features, the system ranks them based on their significance in indicating abnormal network behavior. This prioritization allows the model to focus on the most impactful indicators of potential attacks.

(2) Knowledge Base Construction: The KB serves as the intermediary layer between feature analysis and anomaly detection. It provides a structured repository of critical insights derived from the ranked features. The KB is constructed in two formats to optimize compatibility with different ODLLM model capacities. Long KBs are used for detailed analysis with medium-size ODLLM, and short KBs are used for lightweight ODLLM applications. The KB encapsulates the most distinguishing features of each attack type. These features are encoded as concise descriptors that facilitate quick comparison with incoming traffic data.

(3) Integration with LLM for Anomaly Detection: After identifying the important features and constructing KB, the system integrates them with an ODLLM to perform anomaly detection. The LLM leverages its advanced reasoning capabilities and contextual understanding to interpret the feature set comprehensively. By incorporating domain-specific knowledge, the LLM can accurately predict the type of network attack occurring.

By deploying the model on edge devices, we aim to maximize the accuracy of abnormal traffic detection.

Types of DDoS Attack

We consider four types of DDoS attacks and their characteristics:

ICMP Flood Attack overwhelms the target with a high volume of ICMP echo requests, causing the network to become congested and unresponsive. This attack usually causes increased bandwidth consumption, degraded service performance, and potential downtime.

UDP Flood Attack sends a large number of UDP packets to random ports on the target server, forcing it to process unnecessary requests. It usually has random destination ports and stateless protocol exploitation. UDP flood will increase CPU usage, and denial of legitimate service requests.

TCP SYN Flood Attack exploits the TCP handshake mechanism by sending numerous SYN packets without completing the handshake, consuming server resources. It usually has elevated SYN flags, half-open TCP connections, and spoofed IP addresses. It will exhaust connection tables, leading to an inability to establish new legitimate connections.

TCP PSH+ACK Flood Attack sends a large number of TCP packets with the PSH (Push) and ACK (Acknowledgment) flags set to overwhelm the target's processing capabilities. It usually involves high volumes of PSH and ACK packets that mimic normal traffic patterns, making them difficult to filter. This flood increases processing overhead, leads to resource depletion, and can cause potential service crashes.

Problem Formulation

Our objective is to design a DDoS attack detection model that maximizes the accuracy of identifying various attacks in IoT environments while operating efficiently on edge devices with limited computational resources. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ be the traffic dataset, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the feature vector extracted from network traffic data for the i-th sample. $y_i \in \mathcal{Y} = \{1, 2, \ldots, C\}$ is the corresponding label indicating the attack type, with C being the number of attack classes (including normal traffic). Let $f_\theta : \mathbb{R}^d \to \mathcal{Y}$ be the detection ODLLM model enhanced by KB θ, which maps input features to predicted labels.

The primary goal is to find the optimal KB θ^* that maximizes the overall accuracy on the dataset $\mathcal{D}$. The accuracy $\mathcal{A}(\theta)$ can be defined as:

$$\mathcal{A}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \delta\left(f_\theta(\mathbf{x}_i), y_i\right), \tag{1}$$

where $\delta(a, b)$ is the Kronecker delta function:

$$\delta(a, b) = \begin{cases} 1, & \text{if } a = b, \\ 0, & \text{if } a \neq b. \end{cases} \tag{2}$$

The problem can be formulated as an optimization problem:

$$\theta^* = \arg\max_{\theta} \mathcal{A}(\theta). \quad (3)$$

Considering the edge computing resource limitations and ODLLM model constraints, we need to design KB to maximize the detection accuracy.

Feature Ranking and Knowledge Base Design

To build a reliable model capable of differentiating between various types of attacks, we implemented a structured feature ranking and KB development methodology. This process involved selecting high-impact features, constructing an adaptable KB, iterating for accuracy improvements, and tailoring our approach based on model capacity.

Feature Ranking with Random Forest Regressor

Our initial step in constructing the KB involved ranking features by their importance for each attack type. We employed a Random Forest Regressor (RFR) to evaluate feature significance. The RFR assigns importance based on how well each feature helps split the data to classify attacks correctly. The more a feature contributes to reducing classification mean squared error (MSE) across multiple decision trees, the higher its importance (Firdaus, Munadi, and Purwanto 2020). The importance for feature α_i is computed as:

$$I(\alpha_i) = \frac{1}{T} \sum_{t=1}^{T} \Delta\text{MSE}_t(\alpha_i), \quad (4)$$

where T is the total number of decision trees in the forest, and $\Delta\text{MSE}_t(\alpha_i)$ represents the reduction in MSE for tree t when splitting on feature α_i. The RFR model analyzed a labeled dataset consisting of various attack types, producing an ordered list of features ranked by their importance score $I(\alpha_i)$. From this ranking, we selected the top $k = 10$ features for each attack type, focusing on characteristics with the highest predictive value for distinguishing between attack patterns. For each selected feature, we extracted descriptive statistics, including the lower bound $\min(\alpha)$, upper bound $\max(\alpha)$, and the median value $\text{med}(\alpha)$. These statistics defined a feature range:

$$\text{Range}(\alpha_i) = [\min(\alpha_i), \max(\alpha_i)], \quad (5)$$

which encapsulates typical and boundary behavior for feature α_i. The descriptive statistics for each α_i are given as:

$$\text{Statistics}(\alpha_i) = \{\min(\alpha_i), \text{med}(\alpha_i), \max(\alpha_i)\}. \quad (6)$$

This range and statistical profile provided a numerical signature for each attack type, effectively differentiating between behaviors such as protocol type, packet size, and inter-arrival time (IAT). These bounded values formed the foundation for a structured and precise KB, enabling the model to learn the unique signatures of each attack type and improve prediction accuracy.

We consider the dataset CICIOT 2023, which is a comprehensive benchmark dataset designed for evaluating intrusion detection systems in IoT environments, featuring diverse network traffic types, including normal and malicious behaviors, across multiple IoT protocols and attack scenarios (Neto et al. 2023).

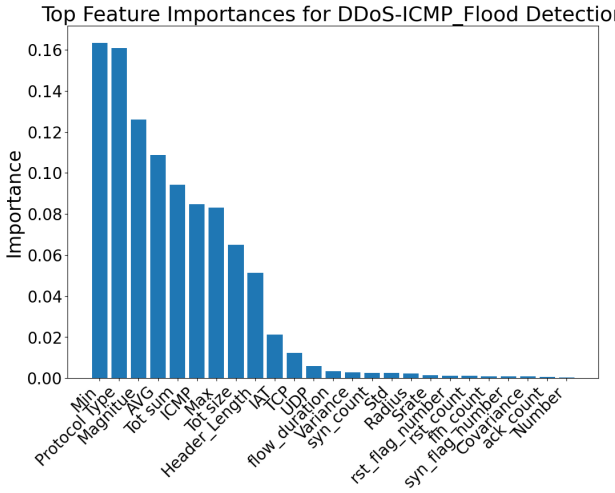

Figure 2: Ranked features for DDoS ICMP Flood attack using Random Forest Regressor on the CIC IoT 2023 Dataset.

DDoS ICMP Flood The feature ranking for DDoS-ICMP_Flood attack is shown in Fig 2. The top two features identified for DDoS-ICMP_Flood attacks in our dataset are the *MIN* value and *Protocol Type*. The *MIN* value represents the minimum size of network traffic packets and exhibits distinct patterns during ICMP flood attacks. This is because ICMP flood attacks often involve numerous small packets, with the *MIN* value ranging from 42.0 to 992.72 and a median of 42.0. In over 99% of cases, the *MIN* value is exactly 42.0, which is consistency as a signature feature for ICMP floods. The *Protocol Type* is another high-importance feature, as ICMP floods specifically utilize the Internet Control Message Protocol (ICMP). The *Protocol Type* ranges from 0.77 to 15.35, with a median of 1.0, which aligns with the expected value for ICMP packets according to standard protocol definitions.

Additionally, features like *Magnitude* and *ICMP* were observed to have moderate importance. The *Magnitude* feature, which quantifies the volume or intensity of network traffic, is particularly indicative of DDoS attacks as ICMP floods typically generate high bursts of traffic. In our dataset, the *Magnitude* ranges from 9.16 to 59.79, with a median of 9.16, demonstrating that most ICMP flood attacks are characterized by relatively high traffic intensity. Similarly, the *ICMP* feature, reflecting the presence of ICMP packets in the traffic, serves as a direct identifier for this attack type. Its values range from 0.0 to 1.0, with a median of 1.0, indicating that ICMP packets are consistently present in traffic associated with ICMP flood attacks. Protocol-specific features, such as HTTP, DNS, SSH, and flag numbers, were found to have little to no relevance for identifying ICMP flood attacks. These features consistently showed zero or near-zero importance scores, as they are unrelated to the characteristics of ICMP floods.

This feature ranking allows the model to utilize KB to focus on the most relevant characteristics for accurate detection of DDoS-ICMP_Flood attacks. The KB can be given as:

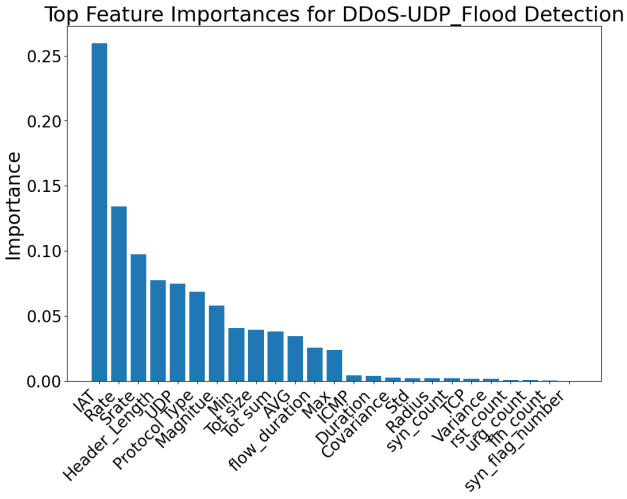

Figure 3: Ranked features for DDoS UDP Flood attack using Random Forest Regressor on the CIC IoT 2023 Dataset.

```
If the attack is DDoS ICMP flood, it
    should exhibit the following
    characteristics:
- Protocol Type: Has to be 1.0 for ICMP.
- ICMP Indicator: Has to be 1.0 for ICMP
    .
- Min Packet Size: Ranges from 42.0 to
    992.72, commonly at 42.0.
- Magnitude: Intensity ranges from 9.17
    to 59.80, with a typical value near
    9.17.
- Average Packet Size (AVG): Spans from
    42.0 to 1885.5, often around 42.0.
- Total Sum of Packets (Tot sum):
    Between 42.0 and 19764.8, commonly
    near 441.0.
- Max Packet Size: Has to be around
    42.0.
- Total Size of Packets (Tot size): Has
    to be 42.0.
- Inter-Arrival Time (IAT): Very high,
    between 0.0 and 100179851.34, with a
    median around 83128994.35.
```

DDoS UDP Flood The feature ranking of DDoS-UDP_Flood attacks is shown in Fig. 3. The two most critical features are *IAT* and *Rate*. This is because of the nature of UDP floods, which are characterized by bursts of packets with minimal inter-arrival time and a high packet rate. The *IAT* ranges from 4.3×10^{-6} to $99,748,506.4$, with a median value of $83,102,993.46$, reflecting the rapid packet generation typical of these attacks. Similarly, the *Rate* feature, which captures the volume of UDP packets sent over the network, spans from 6.0 to $1,569,352.1$, with a median of $7,480.80$. These values highlight the high-frequency, high-volume characteristics of UDP flood attacks.

Features with moderate importance include Source Rate (Srate), Header Length, *UDP*, and *Protocol Type*: *Srate* reflects the source-side packet transmission rate, ranging from 6.0 to $1,569,352.1$, aligning with the traffic burst patterns typical of UDP floods. Header Length, varying between 751.5 and $1,076,354.07$, represents packet sizes associated with UDP flood traffic. The *UDP* feature confirms the protocol type, typically close to 1.0, indicating the attack's reliance on the UDP protocol. *Protocol Type*, ranging from 4.84 to 17.0 with a median of 17.0, differentiates UDP floods from other attacks, as the value corresponds to the UDP protocol in networking standards.

Features with lower importance include *Magnitude*, *Min*, *Tot Size*, and *Tot Sum*. *Magnitude* represents the intensity of the traffic flow, reflecting moderate importance in identifying the attack's burst characteristics. It ranges from 9.97 to 41.16, with a median value of 10.0. *Min*, *Tot Size*, and *Tot Sum* play supporting roles in detecting anomalies associated with UDP floods by capturing packet-level traffic metrics, Features such as *ICMP*, *TCP*, and flag numbers have minimal or zero importance for DDoS-UDP_Flood detection.

This analysis emphasizes the role of high-importance features like *IAT* and *Rate* in distinguishing DDoS-UDP_Flood traffic. By leveraging these prioritized features, the KB can be given as:

```
If the attack is DDoS UDP flood, it
    should exhibit the following
    characteristics:
- Protocol Type: Close to 17.0,
    corresponding to the UDP protocol.
- UDP Indicator: Must be 1.0, confirming
    the presence of UDP packets."
- Inter-Arrival Time (IAT): Extremely
    varied, ranging from 4.39e-06 to
    99748506.47, with a typical value
    around 83102993.47, reflecting high-
    frequency bursts.
- Rate and Source Rate (Srate): Both
    range from 6.01 to 1569352.19, with a
     common value near 7480.80,
    indicating high packet transmission
    volumes.
- Magnitude: Represents traffic
    intensity, ranging from 9.97 to
    41.16, typically about 10.0.
- Minimum Packet Size (Min): Between
    48.74 and 468.37, commonly close to
    50.0, reflecting packet-level
    characteristics.
- Total Packet Size (Tot size): Spans
    from 49.88 to 1075.46, with a
    frequent value near 50.0.
- Total Sum of Packets (Tot sum): Ranges
     from 150.0 to 11576.45, with a
    typical value around 525.0, capturing
     the cumulative packet behavior.
```

DDoS TCP Flood The ranked features for DDoS-TCP_Flood detection are illustrated in Fig. 4. The two most critical features are *IAT* and *SYN Count*, which are essential for identifying the unique traffic patterns associated with TCP flood attacks. TCP floods often involve high-frequency traffic and repeated SYN packets aimed at exhausting server resources. The *IAT* value, which ranges from 1.3×10^{-7} to $99,691,821.6$, with a median of $83,068,279.06$, highlights the rapid and irregular timing patterns typical of high-volume TCP traffic during an attack. Similarly, the *SYN*

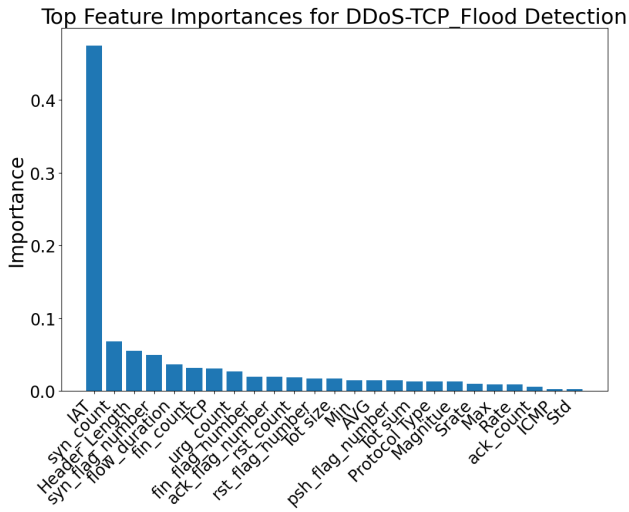

Figure 4: Ranked features for DDoS TCP Flood attack using Random Forest Regressor on the CIC IoT 2023 Dataset.

Count, which captures the frequency of SYN packets, ranges from 0.0 to 2.25, with a median of 0.00. This low median value reflects repeated connection attempts characteristic of TCP flood behavior, where malicious actors send SYN packets to initiate multiple, incomplete TCP connections.

Features with moderate importance include *Header Length*, *SYN Flag Number*, *Flow Duration*, and *FIN Count*. *Header Length*, ranging from 50.96 to $1,264,522.69$, represents the variability in packet sizes during a TCP flood. *SYN Flag Number*, with values typically close to 0.0, indicates a low presence of SYN flags in some TCP flood patterns. *Flow Duration*, spanning 0.0 to $1,270.91$ seconds, provides insights into the connection longevity and stability during the attack. *FIN Count*, ranging from 0.0 to 0.45, captures the frequency of FIN packets, shedding light on TCP session termination behaviors during an attack.

Based on the above feature ranking analysis, the KB can be constructed as:

```
If the attack is DDoS TCP flood, it should exhibit the following characteristics:
- Protocol Type: Close to 6.0, corresponding to the TCP protocol.
- PSH Flag Number: Should be 0.0, reflecting minimal push flags in typical TCP flood behavior.
- TCP Indicator: Often 1.0, confirming the use of the TCP protocol.
- URG Count: Typically 0.0, indicating no urgency flags in normal TCP traffic.
- SYN Flag Number: Typically 0.0, showing the absence or minimal use of SYN flags in regular traffic.
- Flow Duration: Ranges from 0.0 to 1270.90 seconds, often 0.0 in shorter-lived connections characteristic of flood traffic.
- FIN Count: Typically 0.0, but can reach up to 0.45 in some TCP exchanges.
- ACK Flag Number: Mostly 0.0, indicating limited acknowledgment flags in standard TCP flood traffic.
```

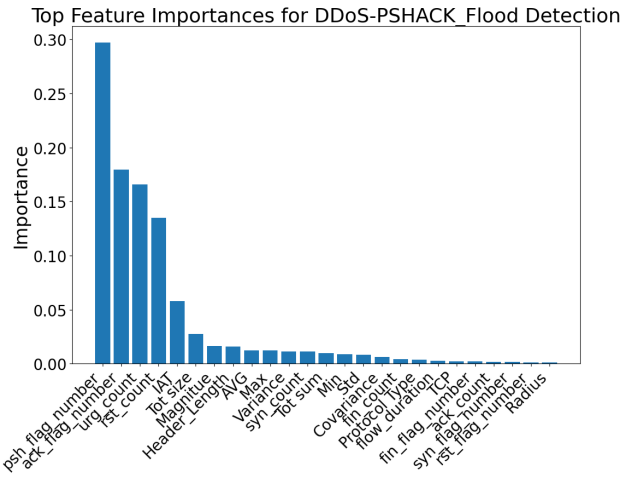

Figure 5: Ranked features for DDoS PSHACK Flood attack using Random Forest Regressor on the CIC IoT 2023 Dataset.

DDoS PSHACK Flood The ranked features for DDoS-PSHACK_Flood detection are illustrated in Fig. 5. The two most critical features are *PSH Flag Number* and *ACK Flag Number*, which are integral to the attack's mechanism. PSHACK floods heavily rely on the consistent presence of PSH and ACK flags to overwhelm the target system. The *PSH Flag Number*, ranging from 0.0 to 1.0 with a median of 1.0, reflects the frequent use of PSH flags in attack packets. This consistency highlights the attack's strategy of forcing the target system to process data packets immediately. Similarly, the *ACK Flag Number*, also ranging from 0.0 to 1.0 with a median of 1.0, underscores the importance of acknowledgment packets in the attack, which maintain the flood of connections and disrupt normal operations.

Other features with moderate importance include *URG Count*, *RST Count*, *Inter-Arrival Time (IAT)*, and *Total Packet Size (Tot Size)*: *URG Count*, with values between 0.0 and 214.22 and a median of 1.0, reflects the occasional presence of urgency flags in PSHACK packets, adding to the attack's complexity. *RST Count*, ranging from 0.0 to 472.02 with a median of 1.0, indicates the frequent use of reset flags, a common tactic in PSHACK floods to disrupt TCP sessions. *IAT*, spanning 1.5×10^{-5} to $99,998,229.54$ with a median of $83,318,215.97$, reflects the timing patterns of high-frequency bursts typical of this attack. *Tot Size*, ranging from 53.76 to 689.69 with a median of 54.0, provides additional indicators of packet size consistency in the attack.

Features with lower importance include *Magnitude*, *Header Length*, *Average Packet Size (AVG)*, and *Maximum Packet Size (Max)*: *Magnitude*, ranging from 10.34

to 31.17 with a median of 10.39, indicates the intensity of the traffic flow. *Header Length*, varying between 51.3 and $1,601,755.99$ with a median of 54.0, captures packet-level details. *AVG* and *Max* values, both with medians of 54.0, emphasize the uniformity of packet sizes during the attack.

This analysis highlights the role of high-importance features such as PSH and ACK flags, which directly correlate with the attack's strategy, supported by moderate-importance features like RST Count and Tot Size that add context to the classification. By focusing on these key features, the detection system can effectively identify and mitigate DDoS-PSHACK_Flood attacks. Thus, the KB is constructed as:

```
'DDoS-PSHACK_Flood': (
    If the attack is DDoS PSHACK flood, it should exhibit the following characteristics:
    - PSH Flag Number: Must be 1.0, indicating the presence of single push flags in the traffic.
    - ACK Flag Number: Often 1.0, but can occasionally be 0.0, distinguishing it from other TCP floods.
    - URG Count: Typically 1.0 but can reach up to 367.51, reflecting the occasional use of urgency flags.
    - RST Count: Usually 1.0, highlighting the frequent use of reset flags in the attack.
    - Inter-Arrival Time (IAT): Ranges from 1.50e-05 to 99998229.53, with a common value around 83318215.96, indicating high-frequency bursts.
    - Total Packet Size (Tot size): Between 53.76 and 1177.9, typically around 54.0, showing consistent packet sizes.
    - Magnitude: Varies in intensity from 10.33 to 40.65, with a common value near 10.39.
    - Average Packet Size (AVG): Ranges from 53.34 to 1079.47, often close to 54.0, showing consistent averages.
    - Maximum Packet Size (Max): Spans from 53.76 to 3022.11, with typical values around 54.0.
)
```

Introducing Key Features for Targeted Predictions

To enhance accuracy further, we introduced a "key feature set" alongside the KB. This set consisted of the most discriminative features for each attack type which we got while ranking the features and then comparing the different attacks, serving as a concise reference for the model during predictions. By focusing on these critical features, the model could prioritize the most relevant characteristics before consulting the broader KB. This two-tiered structure provided both a high-level guide for classification and detailed descriptions for refining distinctions between attack types.

Challenges with Smaller Models and KB Simplification

Smaller models, such as LLaMA 3.2 3B and Phi3 Mini 3.8B, presented significant challenges in utilizing the comprehensive KB. These models struggled to process the volume and complexity of multi-feature input due to their limited parameter capacity, leading to poorer predictive performance compared to scenarios where no KB was used (Xu et al. 2024b) (Shen et al. 2024).

To address these limitations, we hypothesized that smaller models were overwhelmed by the extensive KB, which included redundant and low-impact features. As a solution, we designed a simplified KB tailored specifically for smaller models. This version focused exclusively on the highest-impact features for each attack type, retaining only the most distinctive characteristics and omitting secondary details. This streamlined approach allowed smaller models to process the KB more effectively and significantly improved predictive accuracy. For example, a simplified KB for an ICMP Flood attack can be given as:

```
DDoS-ICMP_Flood: Protocol: ICMP; High packet rate; Low Inter-Arrival Time (IAT).
DDoS-UDP_Flood: Protocol: UDP; High packet rate; Low IAT.
DDoS-TCP_Flood: Protocol: TCP; High packet rate; Elevated SYN flag.
DDoS-PSHACK_Flood: Elevated PSH and ACK flags.
DDoS-SYN_Flood Elevated SYN flag.
DDoS-RSTFIN_Flood: Elevated RST and FIN flags.
DDoS-SynonymousIP_Flood: Multiple source IPs; High SYN counts.
```

The simplified KB was embedded into natural language prompts to help models identify attack types more efficiently. For instance:

```
Network Traffic Data:
- Protocol Type: TCP
- Packet Rate: 450 packets/sec (High)
- Inter-Arrival Time (IAT): Low
- TCP Flags:
    - SYN: Elevated
    - PSH: Normal
    - ACK: Normal
    - RST: Normal
    - FIN: Normal

Based on the knowledge base, determine the most likely attack type. from the following list:( DDoS-ICMP_Flood, DDoS-UDP_Flood, DDoS-TCP_Flood, DDoS-PSHACK_Flood, DDoS-SYN_Flood, DDoS-RSTFIN_Flood, DDoS-SynonymousIP_Flood, Unknow, Normal.
```

This approach successfully bridged the gap in performance for smaller models, enabling them to leverage a streamlined KB and maintain classification accuracy with minimal computational overhead.

Attack Type	Llama 3.1 8B			Phi3 Medium 14B			Gemma2 9B		
	No KB	Long KB	Short KB	No KB	Long KB	Short KB	No KB	Long KB	Short KB
ICMP	97.80%	100.00%	83.80%	50.40%	42.40%	27.40%	20.40%	100.00%	20.00%
UDP	56.40%	86.60%	76.60%	39.20%	31.40%	59.80%	1.80%	100.00%	48.80%
TCP	77.40%	3.60%	77.80%	10.60%	16.80%	6.00%	0.00%	0.00%	0.00%
PSHACK	3.20%	59.40%	54.80%	10.20%	17.80%	28.60%	3.20%	35.20%	15.00%

Table 1: Accuracy of Different Models on Various DDoS Attack Types with and without KBs.

Simulation and Performance Evaluation

We use Ollama to retrieve ODLLMs with our constructed KB on Destkop with Nividia RTX 4090 and Intel I9-13900KF (Liu, Kang, and Han 2024). We test our model's performance with the CICIoT 2023 dataset (Neto et al. 2023). Our source code is released on GitHub (https://github.com/claudwq/Intelligent-IoT-Attack-Detection-Design-via-LLM-with-Feature-Ranking-Based-Knowledge-Base-Design.git). Specifically, we consider the latest small-size models as follows:

Llama 3.1 8B is the compact variant in the Llama 3.1 series developed by Meta AI with 8 billion parameters (Dubey et al. 2024). **Phi3 Medium 14B** is part of the Phi-3 series developed by Microsoft with 14 billion parameters, it offers substantial reasoning capabilities while maintaining a manageable computational footprint (Abdin et al. 2024). **Gemma2 9B** is a high-performing and efficient language model within the Gemma 2 series developed by Google DeepMind, which includes models with 2B, 9B, and 27B parameters (Team et al. 2024). **Llama 3.2 3B** is a compact variant in the Llama 3 series with 3 billion parameters, optimized for multilingual tasks and large-scale text processing (Dubey et al. 2024). **Phi3 Mini 3.8B** is also part of the Phi-3 series with 3.8 billion parameters, designed as a lightweight model optimized for chat-based interactions and reasoning tasks (Abdin et al. 2024).

Performance of Medium-size Detectors

We first evaluate the performance of medium-size models with our KB. As shown in table 1, the simulation results evaluate the performance of three ODLLM on detecting four distinct DDoS attack types under three configurations: without a KB (KB), with a long KB, and with a short KB. Accuracy metrics are reported to assess the models' ability to classify network traffic into these categories.

The long KB is most effective for DDoS-ICMP_Flood and DDoS-UDP_Flood detection across all models, particularly for Gemma2 9B, where accuracy increased to 100.00% for both attacks. However, its performance was inconsistent for other attack types, such as DDoS-TCP_Flood, where accuracy degraded for Llama 3.1 8B (3.60%). The short KB demonstrated a better balance between performance and simplicity, particularly for Phi3 Medium 14B, where accuracy improved for DDoS-UDP_Flood (59.80%) and DDoS-PSHACK_Flood (28.60%). However, it generally underperformed for Gemma2 9B.

Llama 3.1 8B exhibited the highest accuracy overall, benefiting significantly from both KBs. Its ability to leverage the long KB for DDoS-ICMP_Flood and the short KB for DDoS-TCP_Flood highlights its versatility. Phi3 Medium 14B showed moderate performance, with notable improvement for DDoS-UDP_Flood and DDoS-PSHACK_Flood when the short KB was used. This suggests that phi3 Medium 14B is more suited for concise knowledge representation. Gemma2 9B is heavily reliant on the long KB, achieving perfect accuracy for some attacks but failing entirely for others, such as DDoS-TCP_Flood.

Performance of Small-size Detectors

To evaluate the effectiveness of the KB configurations on small-size ODLLM, we conducted experiments using two smaller language models: **Llama 3.2 3B** and **Phi3 Mini 3.8B**. Table 2 presents the accuracy results for the models across the KB configurations. For Llama 3.2 3B, the model achieved low accuracy without a KB, particularly for DDoS-PSHACK_Flood, where the accuracy was only 1.60%. The inclusion of the long KB improved accuracy for DDoS-ICMP_Flood, achieving 42.00%, but led to decreased accuracy for DDoS-UDP_Flood, which dropped to 23.40%. This suggests that the long KB may have introduced unnecessary complexity, overwhelming the model's capacity. When using the short KB, however, accuracy improved significantly across all attack types. DDoS-UDP_Flood and DDoS-TCP_Flood achieved the highest accuracy at 53.80% and 53.40%, respectively, demonstrating that short KB retained critical information while reducing complexity.

For Phi3 Mini 3.8B, the model exhibited very low accuracy without a KB, with DDoS-PSHACK_Flood achieving only 0.19% accuracy and DDoS-UDP_Flood reaching just 0.97%. Adding the long KB resulted in marginal improvements, with DDoS-ICMP_Flood reaching 9.60% accuracy and DDoS-UDP_Flood improving to 4.20%. However, these results were still significantly lower compared to the short KB configuration. With the short KB, the model's accuracy increased substantially for DDoS-UDP_Flood and DDoS-TCP_Flood, both reaching 22.20%. This demonstrates short KB was far more effective for smaller models.

Both models achieved their highest accuracy for DDoS-ICMP_Flood across all KB configurations, indicating that the features for this attack type were straightforward and well-represented in the KB. For DDoS-UDP_Flood, the short KB significantly improved accuracy, particularly for Llama 3.2 3B, which achieved 53.80%. This highlights the impact of focusing on essential UDP flood features. Similarly, DDoS-TCP_Flood detection benefitted greatly from the short KB, with Llama 3.2 3B achieving 53.40% accuracy. Both models struggled to detect DDoS-

Attack Type	Llama 3.2 3B			Phi3 mini 3.8B		
	No KB	With Long KB	With Short KB	No KB	With Long KB	With Short KB
ICMP	28.20%	42.00%	52.40%	6.65%	9.60%	13.20%
UDP	38.80%	23.40%	53.80%	0.97%	4.20%	22.20%
TCP	22.60%	27.80%	53.40%	0.97%	4.20%	22.20%
PSHACK	1.60%	3.40%	38.80%	0.19%	0.00%	3.00%

Table 2: Accuracy Comparison for DDoS Attack Detection with Different Models and KB Configurations.

PSHACK_Flood, even with the short KB, with Phi3 Mini 3.8B achieving only 3.00%. This suggests that the feature set for this attack type may require further refinement to enhance detectability.

Conclusions

In this paper, we presented an intelligent IoT network attack detection framework leveraging ODLLM integrated with feature ranking-based KB designs. The proposed system addresses the challenges of computational resource constraints and data privacy in edge environments while providing a scalable and efficient solution for DDoS attack detection. Our experiments demonstrated that ODLLMs equipped with a simplified KB tailored to model capacity could achieve competitive performance even on resource-constrained devices. By ranking features using RFR and constructing long and short KBs, we successfully optimized the system's ability to detect various DDoS attack types, including DDoS-ICMP Flood, DDoS-UDP Flood, DDoS-TCP Flood, and DDoS-PSHACK Flood.

References

Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A. A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Aysa, M. H.; Ibrahim, A. A.; and Mohammed, A. H. 2020. IoT Ddos Attack Detection Using Machine Learning. In *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, 1–7.

Chen, W.; Li, Z.; and Ma, M. 2024. Octopus: On-device language model for function calling of software APIs. *arXiv preprint arXiv:2404.01549*.

Chen, W.; Li, Z.; Xin, S.; and Wang, Y. 2024. Dolphin: Long Context as a New Modality for Energy-Efficient On-Device Language Models. *arXiv e-prints*, arXiv–2408.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Firdaus, D.; Munadi, R.; and Purwanto, Y. 2020. DDoS Attack Detection in Software Defined Network using Ensemble K-means++ and Random Forest. In *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 164–169.

Hussain, F.; Abbas, S. G.; Husnain, M.; Fayyaz, U. U.; Shahzad, F.; and Shah, G. A. 2020. IoT DoS and DDoS Attack Detection using ResNet. In *2020 IEEE 23rd International Multitopic Conference (INMIC)*, 1–6.

Jaton, N.; Gyawali, S.; and Qian, Y. 2023. Distributed Neural Network-Based DDoS Detection in Vehicular Communication Systems. In *2023 16th International Conference on Signal Processing and Communication System (ICSPCS)*, 1–9.

Jia, Y.; Zhong, F.; Alrawais, A.; Gong, B.; and Cheng, X. 2020. FlowGuard: An Intelligent Edge Defense Mechanism Against IoT DDoS Attacks. *IEEE Internet of Things Journal*, 7(10): 9552–9562.

Liu, F.; Kang, Z.; and Han, X. 2024. Optimizing RAG Techniques for Automotive Industry PDF Chatbots: A Case Study with Locally Deployed Ollama Models. *arXiv preprint arXiv:2408.05933*.

Neto, E. C. P.; Dadkhah, S.; Ferreira, R.; Zohourian, A.; Lu, R.; and Ghorbani, A. A. 2023. CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment. *Sensors*, 23(13): 5941.

Shen, W.; Li, C.; Chen, H.; Yan, M.; Quan, X.; Chen, H.; Zhang, J.; and Huang, F. 2024. Small llms are weak tool learners: A multi-llm agent. *arXiv preprint arXiv:2401.07324*.

Team, G.; Riviere, M.; Pathak, S.; Sessa, P. G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Xu, J.; Li, Z.; Chen, W.; Wang, Q.; Gao, X.; Cai, Q.; and Ling, Z. 2024a. On-device language models: A comprehensive review. *arXiv preprint arXiv:2409.00088*.

Xu, J.; Wang, Q.; Cao, Y.; Zeng, B.; and Liu, S. 2024b. A General Purpose Device for Interaction with LLMs. In *Proceedings of the Future Technologies Conference*, 613–626. Springer.

Yao, Y.; Duan, J.; Xu, K.; Cai, Y.; Sun, Z.; and Zhang, Y. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 100211.

Zhang, J.; Liang, S.; Ye, F.; Hu, R. Q.; and Qian, Y. 2023. Towards Detection of Zero-Day Botnet Attack in IoT Networks Using Federated Learning. In *ICC 2023 - IEEE International Conference on Communications*, 7–12.

Zhu, J.; Cai, S.; Deng, F.; Ooi, B. C.; and Wu, J. 2024. Do LLMs Understand Visual Anomalies? Uncovering LLM's Capabilities in Zero-shot Anomaly Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 48–57. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706868.