Adaptive Sparse-Aware Quantized Deep Learning for Scalable and Privacy-Preserving IoT Networks

DR. ERIC GYAMFI

SCHOOL OF COMPUTING

Presentation Outline

- Introduction and Research Background
- Proposed Method
- Some Early Results
- Conclusion

The Growing IoT Security Challenge

A Pervasive Threat Surface

The widespread adoption of IoT across Smart Cities, Grids, and Healthcare has transformed contemporary digital ecosystems but significantly enlarged the cyber-attack surface.

- Massive scale of heterogeneous devices (over 72B Connected Devices in 2028)
- Exposure to Cyberattacks such as: DDoS, Botnets, and Spoofing, etc.
- Critical risk to life-critical infrastructure.

Limitations of Traditional IDS

Power Constraints

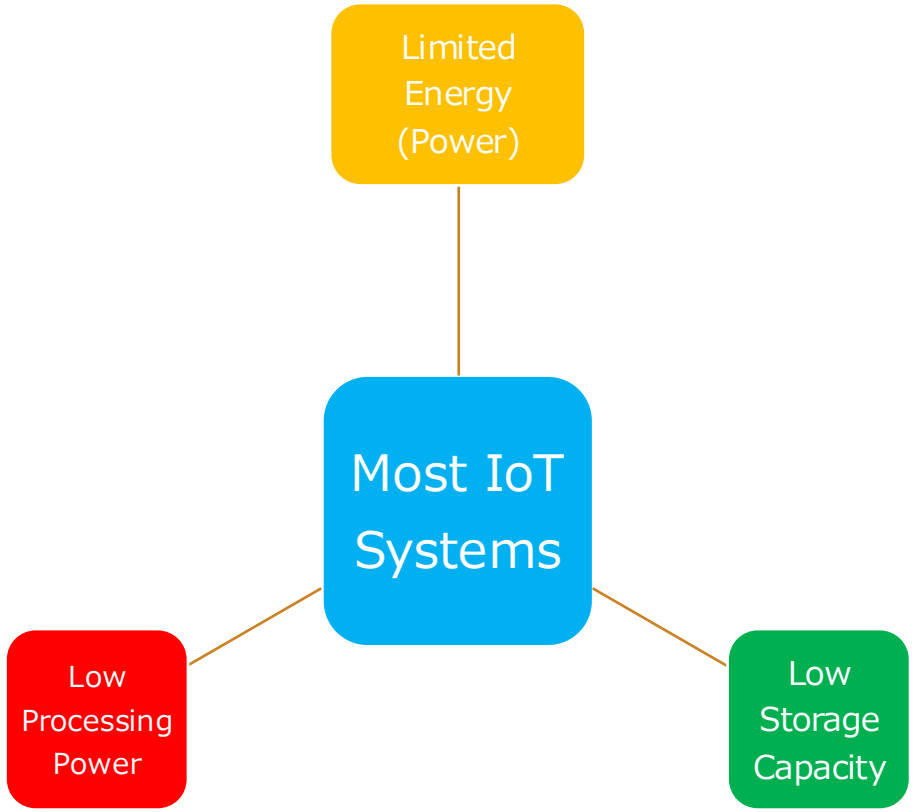
Stringent energy budgets for battery-powered or energy-harvesting IoT nodes.

Memory Footprint

Standard over-parameterized DL models exceed hardware memory limits (RAM/Flash).

Latency Deficit

High inference latency prevents real-time response to zero-day threats.



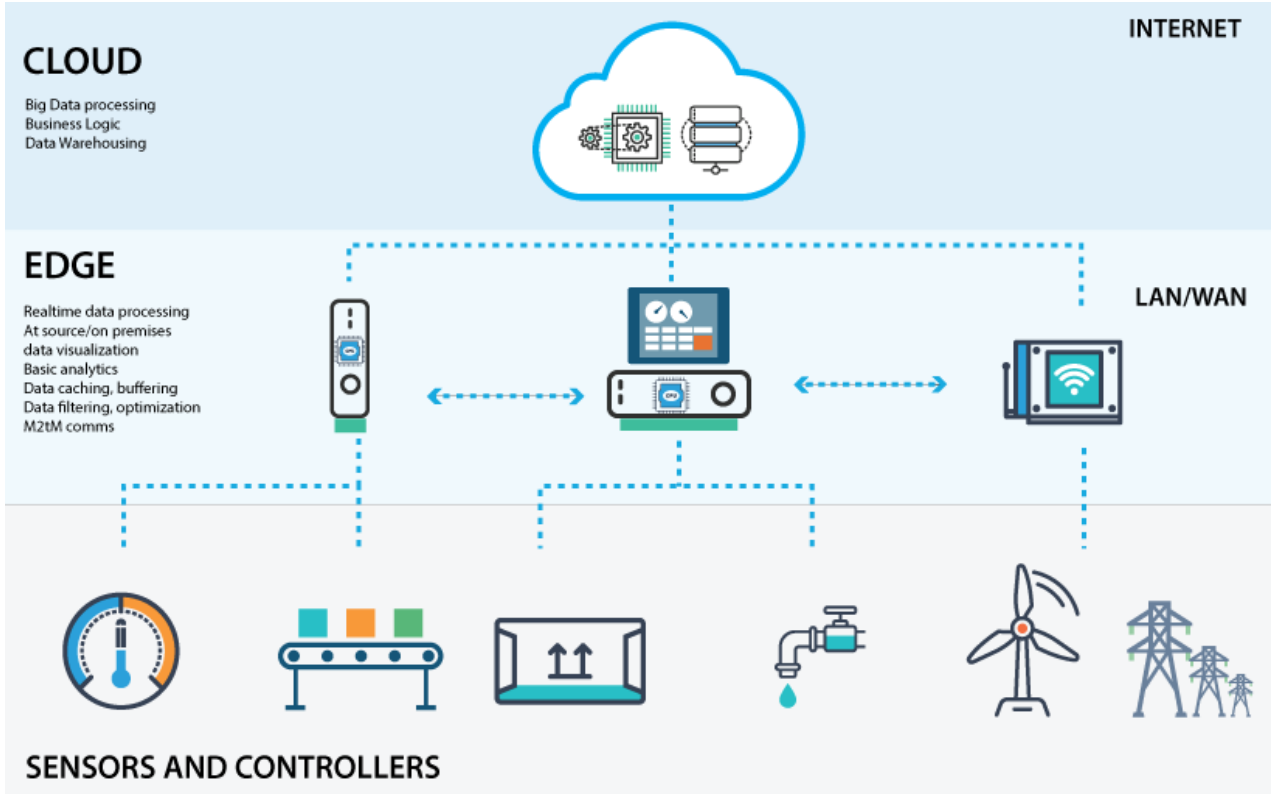
Cloud Vs. Edge Intelligence

Cloud-Based IDS

- High detection accuracy but high latency and privacy risks due to raw data transmission.

Edge-Based

- Local intelligence, real-time mitigation, and privacy-preserving processing with minimal footprints.



Three Core Implementation Approaches

Feature Reduction

Database-level approach using PCA, information gain, and autoencoders to reduce input dimensionality without losing critical signal.

Model Compression

ML-level approach involving **Structured Pruning** and **Quantization-Aware Training** to minimize SRAM footprint.

Placement Tech

Strategic deployment across Sensors (local), Gateways (edge), and MEC for a hierarchical and collaborative defense mesh.

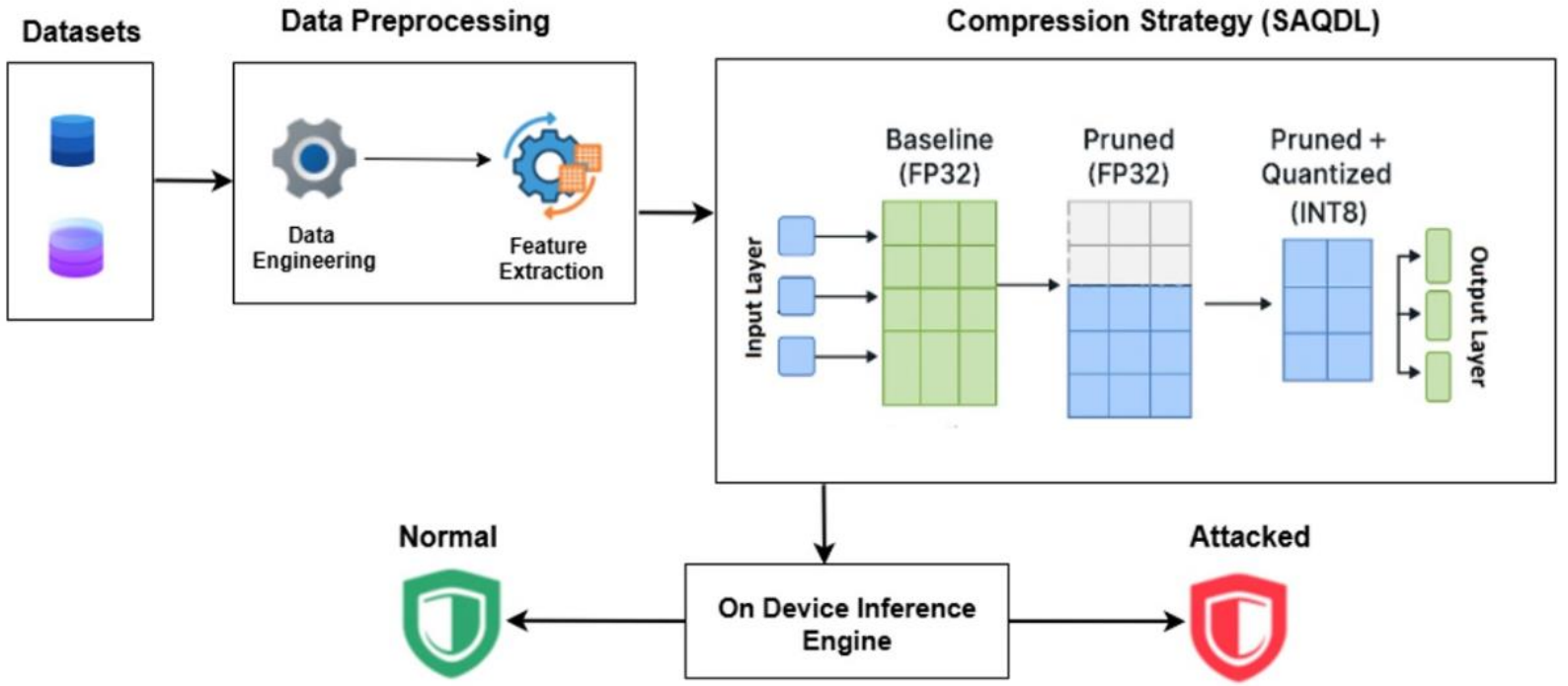
Proposed Methodology

ADAPTIVE SPARSE-AWARE QUANTIZED DEEP LEARNING (SAQDL)

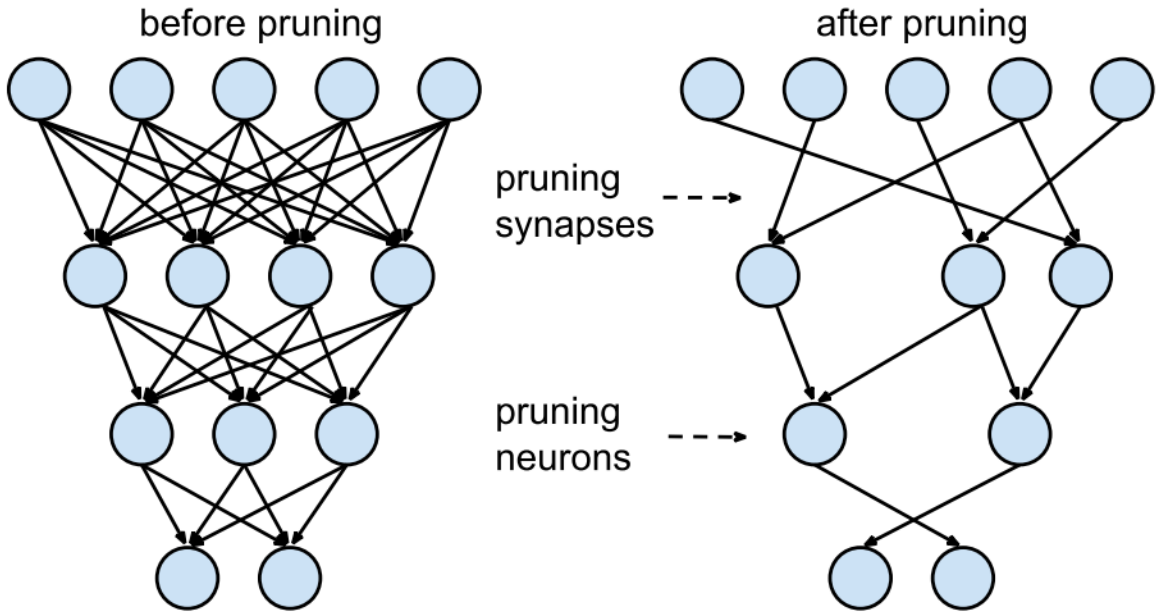
The SAQDL Framework

Design Objectives:

- Design and Implement the proposed SAQDL IDS Framework
- Embedded Optimization for Lightweight Training
- Hierarchical Collaborative Detection Architecture
- Confidence-weighted Trust and Isolation Mechanism



Structured Pruning



Eliminating Redundancy

Identifying and removing non-critical neurons and channels to optimize the network's expressive power.

- Binary masking applied to weight matrices.
- Enables hardware-level parallel acceleration.
- Maintains accuracy via sparse-aware recovery.

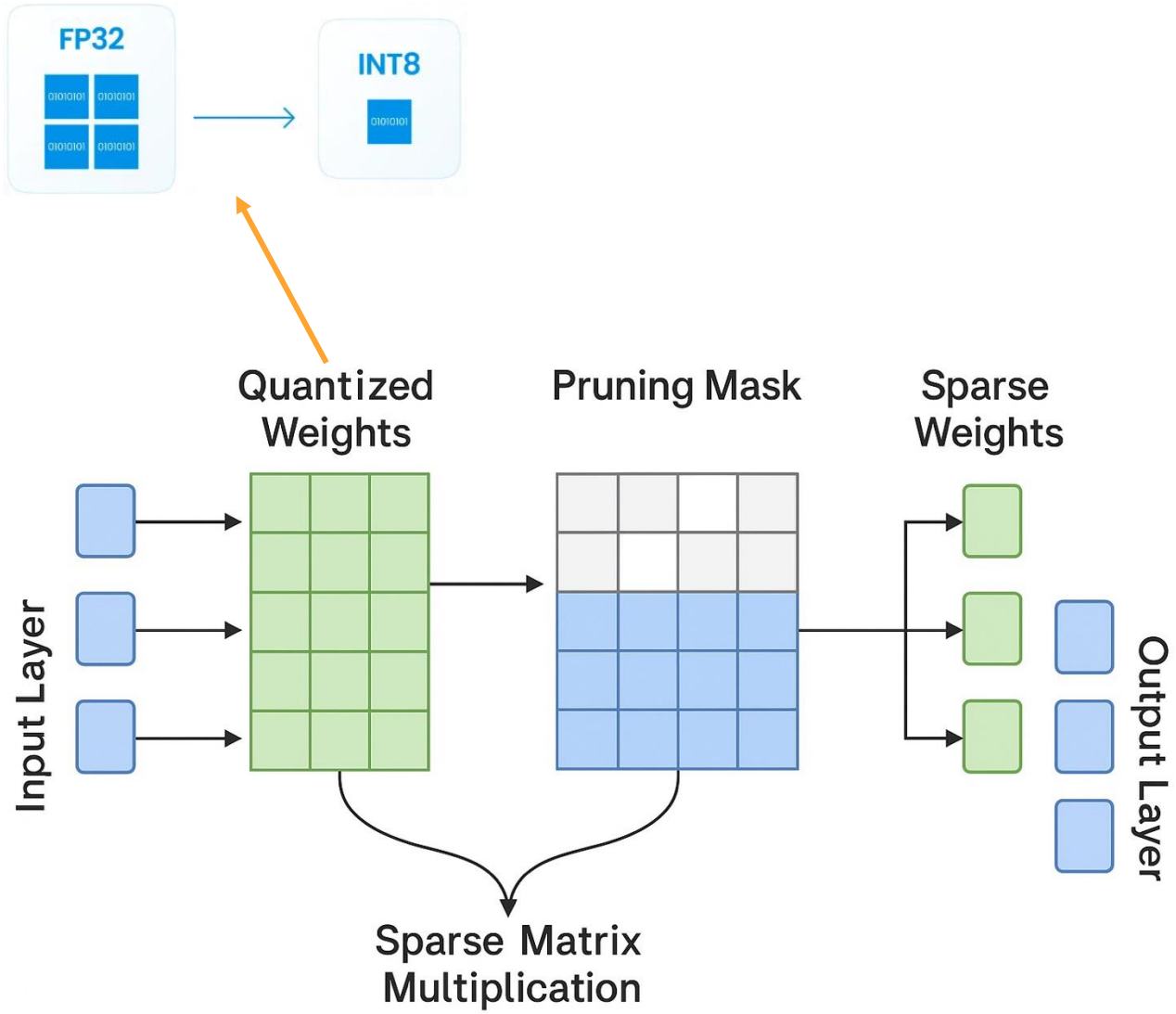
Phase 2: Quantisation (QAT)

Precision Reduction

Converts weights from 32-bit floating point to 8-bit integers (INT8). This provides a 4x reduction in model size and energy per operation.

Awareness Training

Simulates quantization noise during training to allow the network to adapt, maintaining >99% accuracy even at lower precision.



Unified Optimization Objective

97%

Model Compression

The Joint Loss Function

We formulate a customized optimization model that concurrently minimizes classification error, sparsity, and quantization noise.

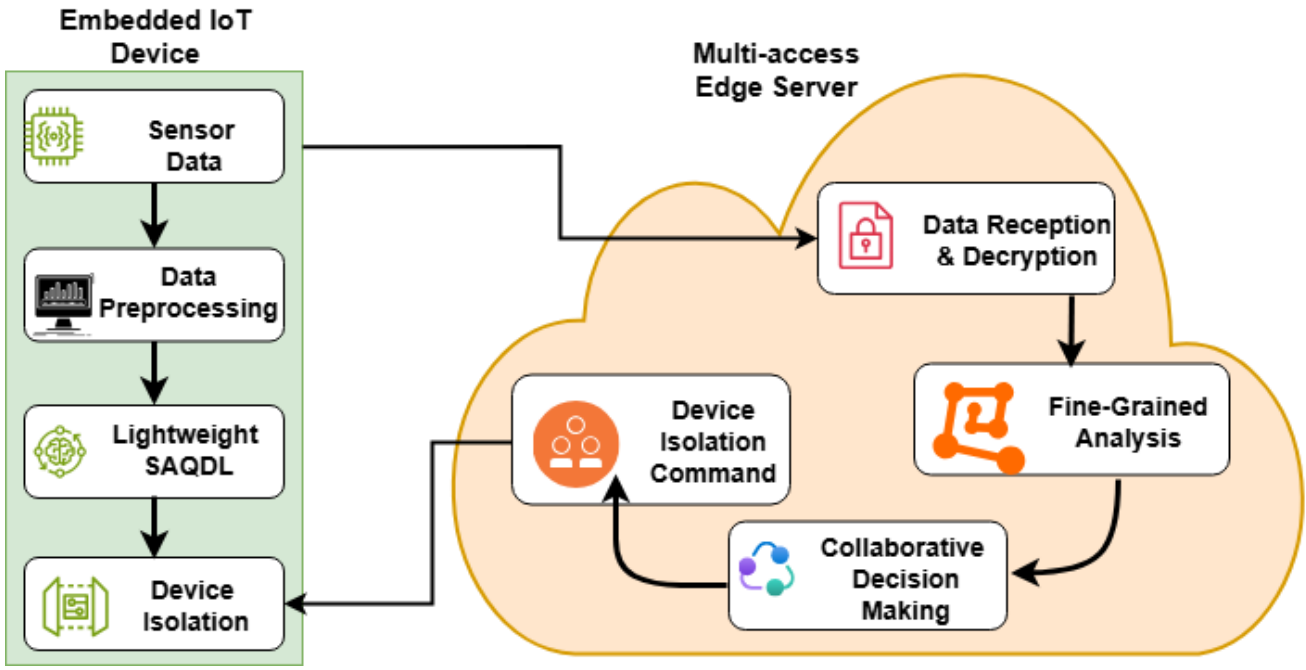
$$L_{total} = L_{CE} + \lambda_1 \sum \| M^{(l)} \|_0 + \lambda_2 R(Q_b(\theta))$$

Two-Layer Hierarchical Defense

Local vs. Edge Collaboration

The system operates at two levels to balance speed and accuracy.

- **Local Layer:** Real-time detection on IoT nodes using SAQDL.
- **MEC Layer:** Fine-grained collaborative analysis on Edge servers.
- Requests escalated only when thresholds are met.



Attack Mitigation & Isolation Flow

Local Detection

SAQDL flags anomaly on IoT device.

Edge Escalation

MEC node receives traffic metadata.

Consensus

Threat validated via collaborative analysis.

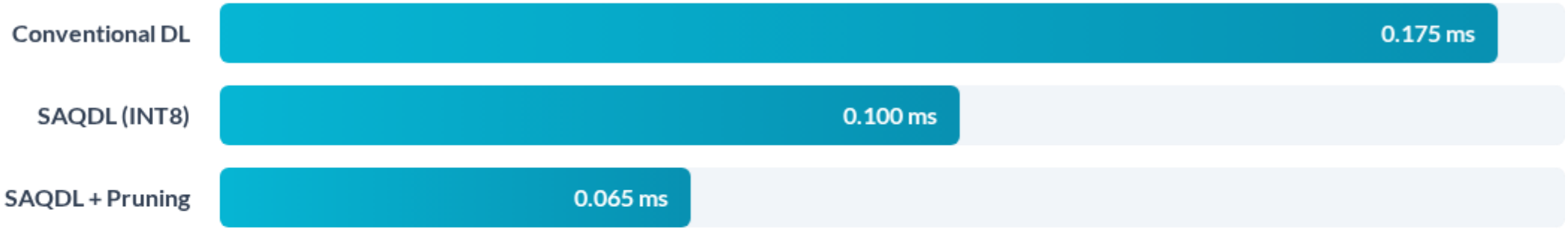
Isolation

Compromised node is logically isolated.

Experimental Setup: CICIoT2023

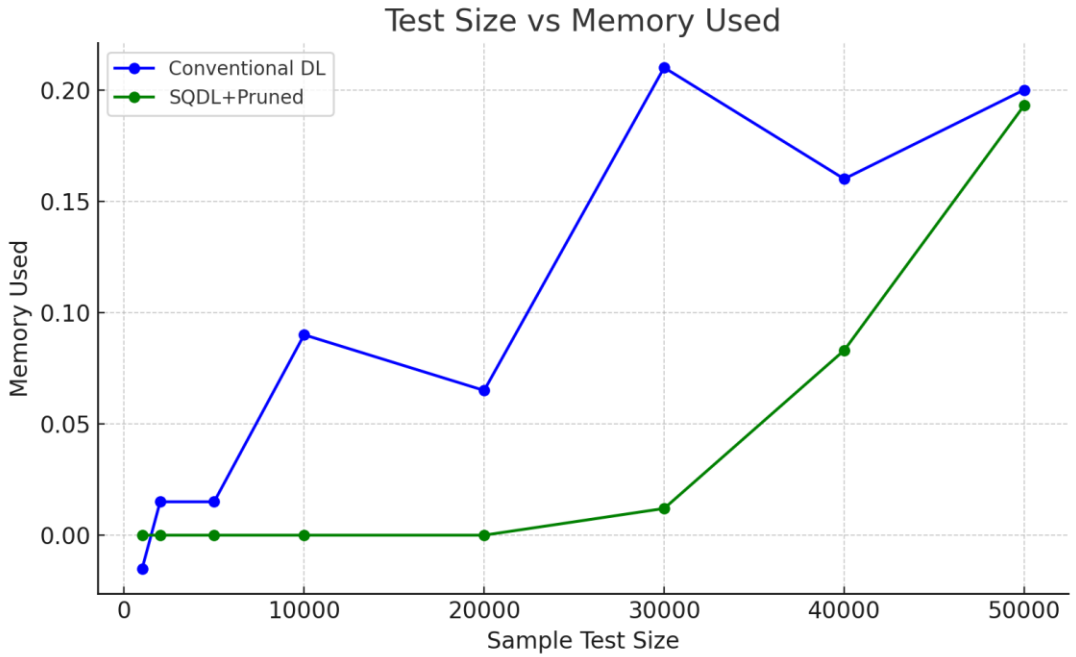
Metric Group	Deployment Device	Dataset Features	Training Epochs
IoT Device Layer	Raspberry Pi 4 (1GB)	Network Headers / Protocol Stats	100 (Early Stopping)
Edge Layer	M1-Mac Mini (MEC Node)	Min-Max Scaled / Categorical Encoded	Adam Optimizer (10^{-3})
Attack Samples	33 Multi-label Classes	DDoS, Mirai, Brute Force, etc.	80/20 Train-Val Split

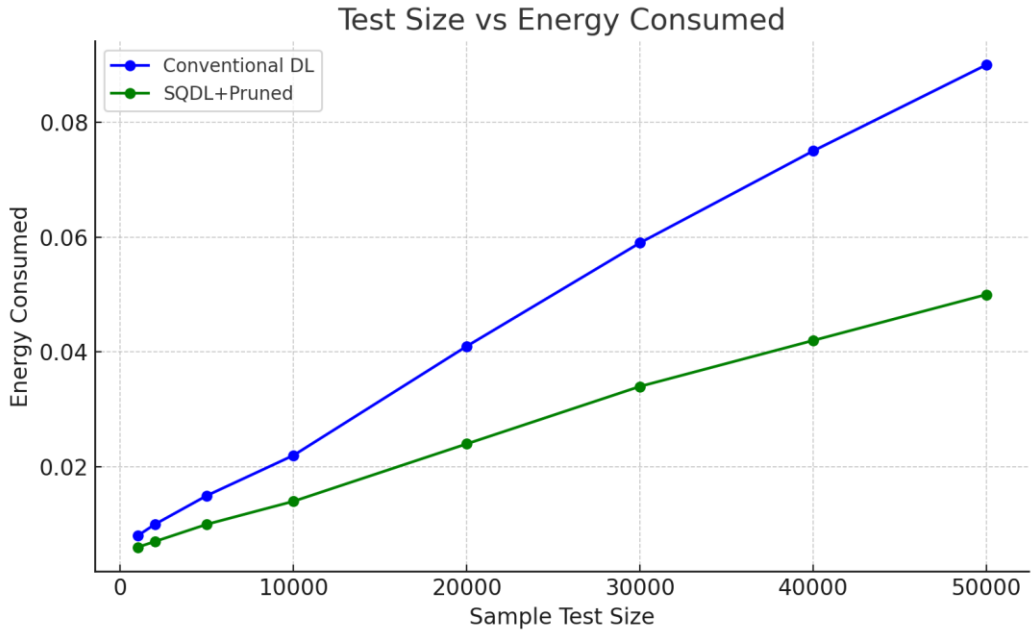
Inference Latency Performance



SAQDL achieves a ~63% reduction in CPU execution latency compared to standard deep learning architectures, enabling practical real-time deployment.

Memory Usage /Energy vs. Scale





Memory footprints remain remarkably stable and significantly below the conventional DL threshold, even as data complexity and sample sizes scale towards 50k packets.

Comparative Analysis with State-of-the-Art

Method	Accuracy (%)	Latency (s)	Energy (J)	Memory (MB)
Javaid et al. [10]	98.40	0.235	0.15	4.8
Meidan et al. [12]	97.00	0.187	0.12	3.5
Proposed SAQDL	**99.13**	**0.100**	**0.05**	**1.8**

Thank you

| ERIC.GYAMFI@NCIRL.IE|