Wadi Alshatti University Journal of Pure and Applied Sciences

مجلة جامعة وادي الشاطئ للعلوم البحتة و التطبيقية

Volume 4, No. 1, January-June 2026 Online ISSN: 3006-0877 المجلد 4، الاصدار 1، يناير- يونيو 2026

ORIGINAL RESEARCH **BIOMATERIALS**

A New Approach for Low-Latency, High-Accuracy Anomaly Detection at the Edge: Benchmarking Quantized Autoencoders, LSTMs, and Lightweight Transformers on RT-IoT2022 Time-Series Traffic

Magdah Osman[1,*], Fatma Elghaffi[2], Llahm Ben Dalla[3], Ömer Karal[3], Tarik Rashid[4]

[1]Systems analysis and programming Department, Higher Institute of Science and Technology, Ajdabiya , Libya
[2]Computer Science Department, Higher Institute of Science and Technology, Ajdabiya , Libya
[3]Department of Electric Electronics, Ankara Yildirim Beyazit University, Türkiye
[4]Artificial Intelligence and Innovation Centre University of Kurdistan Hewler, Erbil, Iraq

ARTICLE HISTORY
Received 05 January 2026
Revised 23 January 2026
Accepted 26 January 2026
Online 30 January 2026

KEYWORDS
Edge AI;
IoT security, anomaly detection;
Quantized autoencoder;
Lightweight LSTM;
distilled Transformer;
RT-IoT2022;
Real-time intrusion detection.

ABSTRACT
This study benchmarks edge-optimized deep learning models for real-time anomaly detection in resource-constrained IoT environments using the RT-IoT2022 dataset, which includes four benign protocols and nine cyberattack types. Three architectures a quantized autoencoder (QAE), compact LSTM, and lightweight Transformer were deployed on a Raspberry Pi 4 and evaluated on F1-score, latency, model size, and energy per inference. The QAE achieved optimal performance with 98.7% F1-score, 142 KB memory footprint, 1.8 ms latency, and 4.2 mJ energy consumption, outperforming alternatives under strict edge constraints. While the LSTM showed better recall on rare attacks and the Transformer captured long-range dependencies at higher computational cost, the QAE delivered the best overall trade-off for deployable security. The work reframes model selection around hardware-aware co-design rather than architectural complexity, demonstrating that intelligently compressed, reconstruction-based approaches surpass heavier models in efficiency and effectiveness. Findings provide a reproducible framework for low-latency, privacy-preserving intrusion detection in smart healthcare and industrial IoT, advocating a paradigm shift toward minimal sufficiency over maximal capacity in edge AI design.

نهج جديد للكشف عن الحالات الشاذة بدقة عالية وزمن استجابة منخفض على الحافة: تقييم أداء المشفرات التلقائية الكمية، وشبكات الذاكرة طويلة المدى، والمحولات الخفيفة على بيانات حركة مرور السلاسل الزمنية RT-IoT2022

ماجدة عثمان[1*]، فاطمة القذافي[2]، للاهم بن دلة[3]، عمر كرال[3]، طارق راشد[4]

الملخص
تقارن هذه الدراسة نماذج التعلم العميق المُحسّنة للحافة للكشف عن الحالات الشاذة في الوقت الفعلي ضمن بيئات إنترنت الأشياء ذات الموارد المحدودة، وذلك باستخدام مجموعة بيانات RT-IoT2022 التي تتضمن أربعة بروتوكولات سليمة وتسعة أنواع من الهجمات الإلكترونية. تم نشر ثلاثة نماذج معمارية - مُشفّر تلقائي كمي (QAE)، وشبكة LSTM مُدمجة، ونموذج Transformer خفيف الوزن - على جهاز Raspberry Pi 4، وتم تقييمها بناءً على مقياس F1، وزمن الاستجابة، وحجم النموذج، واستهلاك الطاقة لكل استدلال. حقق نموذج QAE أداءً مثاليًا بمقياس F1 بلغ 98.7%، وحجم ذاكرة 142 كيلوبايت، وزمن استجابة 1.8 مللي ثانية، واستهلاك طاقة 4.2 مللي جول، متفوقًا بذلك على البدائل في ظل قيود الحافة الصارمة. في حين أظهر نموذج LSTM استدعاءً أفضل للهجمات النادرة، واستطاع نموذج Transformer رصد التبعيات بعيدة المدى بتكلفة حسابية أعلى، إلا أن نموذج QAE قدّم أفضل توازن شامل من حيث الأمان القابل للتطبيق. يُعيد هذا العمل صياغة مفهوم اختيار النموذج ليرتكز على التصميم المشترك المُراعي للأجهزة بدلاً من التعقيد المعماري، مُبرهناً على أن الأساليب المُضغوطة بذكاء والقائمة على إعادة البناء تتفوق على النماذج الأثقل من حيث الكفاءة والفعالية. تُوفر النتائج إطاراً قابلاً للتكرار للكشف عن الاختراقات مع الحفاظ على الخصوصية وزمن الاستجابة المنخفض في الرعاية الصحية الذكية وإنترنت الأشياء الصناعية، داعياً إلى تحول نموذجي نحو الحد الأدنى من الكفاءة بدلاً من السعة القصوى في تصميم الذكاء الاصطناعي على الحافة..

الكلمات المفتاحية
الذكاء الاصطناعي الطرفي
أمن إنترنت الأشياء، كشف الشذوذ
المُشفّر التلقائي الكمي
LSTM خفيف الوزن
المُحوّل المُقطّر، RT-IoT2022
كشف الاختراقات في الوقت الحقيقي

Introduction

Modern Internet of Things (IoT) ecosystems spanning smart healthcare, industrial automation, and residential systems are increasingly vulnerable to sophisticated cyber threats due to

*Corresponding author
https://doi.org/10.63318/waujpasv4i1_12

their distributed nature and limited built-in security [1]. Traditional cloud-based intrusion detection systems (IDS) introduce unacceptable latency and privacy risks, prompting a shift toward edge-native solutions. However, deploying deep learning based IDS on resource-constrained edge devices remains challenging due to strict limitations on memory (<512 MB), computational throughput, and energy budget. While model compression techniques such as quantization, pruning, and architectural distillation offer promising pathways, empirical validation across diverse, real-world IoT traffic is still scarce [2,3]. To address this gap, we present a hardware-aware benchmark of three edge-optimized architectures quantized autoencoder (QAE), compact LSTM, and lightweight Transformer evaluated on the RT-IoT2022 dataset, which captures multivariate time-series traffic from real IoT devices under nine contemporary attack vectors and four benign protocols [4,5]. Unlike prior work that emphasizes architectural novelty, this research study focuses on system-level trade-offs between accuracy, inference latency, model footprint, and energy consumption on a Raspberry Pi 4 platform. This approach reframes edge AI design around deployability rather than complexity, demonstrating that intelligently compressed models can achieve high detection fidelity without sacrificing real-time performance. This research provides a reproducible framework for low-latency, privacy-preserving anomaly detection tailored to the operational realities of edge computing environments. A quantized autoencoder (QAE) trained for reconstruction-based anomaly scoring,

A pruned as well as quantized LSTM for sequential pattern recognition,

A Tiny Transformer with parameter sharing and reduced attention heads.

This research contributions are threefold:

First comparative study of QAE, LSTM, as well as Transformer variants on the RT-IoT2022 dataset under unified edge deployment constraints.

Quantitative evaluation of accuracy-latency-footprint trade-offs across 12-class traffic (9 attacks + 3 benign IoT protocols).

Open-source release of optimized model weights, preprocessing pipelines, as well as edge inference scripts to foster reproducibility.

Related Work

Autoencoders [3] are frequently used for unsupervised anomaly detection via reconstruction error, and deep learning has demonstrated potential in network intrusion detection. However, edge deployment is not a good fit for their full-precision versions. Quantization methods [4] lower the bit-width, for instance, 32-bit → 8-bit; to shrink model size as well as accelerate inference central to the QAE approach in [2]. Although recurrent models, for instance, LSTM [5], are able to capture temporal dynamics in network flows, they are hindered via sequential computing constraints. LSTMs have recently been compressed using layer fusion and pruning [6] for Internet of Things applications. Despite their strength in simulating long-range dependencies, transformers [7,8,9], as well as [10] are usually too bulky for edge devices. The possibility of lightweight versions like MobileViT [11] and TinyBERT [12] is demonstrated via the attention head reduction methods as well as depth-wise convolutions that are modified here.

The RT-IoT2022 dataset [1] advances beyond synthetic benchmarks , for instance, NSL-KDD, UNSW-NB15; via incorporating real IoT device traffic as well as contemporary attack vectors, making it ideal for evaluating practical edge-IDS solutions.

Methodology

Dataset Overview

RT-IoT2022 contains 123,117 flow instances with 83 features extracted via Zeek and Flowmeter, including packet counts, inter-arrival times, payload statistics, as well as TCP flag distributions. The dataset comprises 12 classes: 9 attack types, for instance, DOS_SYN_Hping, DDOS_Slowloris) and 3 benign IoT protocols (MQTT, ThingSpeak, Amazon-Alexa, plus Wipro-bulb traffic. No missing values are present, as well as class distribution is imbalanced mirroring real-world conditions.

Table 1: The Real Time Internet of Things Dataset Characteristics

Factors	Explanation
Number of Instances	123,117
Number of Features**:**	83
Feature Types	Combination of real as well as categorical attributes.
Target Variable (class label)	Contains both attack patterns as well as normal patterns, making it suitable for supervised learning.
Number of classes	12
Source	https://archive.ics.uci.edu/dataset/942/rt-iot2022

Table 2: Class Categorization in the RT-IoT2022 Dataset

Category	**Class Label**	**Description**
Attack Patterns	DOS_SYN_Hping	A DoS attack exploiting the TCP handshake via flooding the target with SYN packets without completing the connection.
	ARP_Poisoning	Manipulates ARP cache entries to perform man-in-the-middle attacks via redirecting traffic within a local network.
	NMAP_UDP_SCAN	Scans UDP ports to discover open services via sending empty or malformed UDP packets as well as analyzing responses.
	NMAP_XMAS_TREE_SCAN	Sends TCP packets with FIN, URG, as well as PUSH flags set to probe for open/closed ports based on RFC-compliant responses.
	NMAP_OS_DETECTION	Fingerprinting technique to infer the target’s operating system via analyzing subtle differences in TCP/IP stack behavior.
	NMAP_TCP_SCAN	Standard TCP connect scan to identify open ports also active services on a target host.
	DDOS_Slowloris	A low-rate DDoS attack that exhausts server connection pools via maintaining partial HTTP connections indefinitely.
	Metasploit_Brute_Force_SSH	Automated brute-force attack utilizing Metasploit to guess valid SSH credentials as well as gain unauthorized remote access.
	NMAP_FIN_SCAN	Sends TCP packets with only the FIN flag set; used to detect closed ports (which respond with RST) while open ports remain silent.

Normal Patterns	MQTT	Lightweight publish-subscribe messaging protocol widely used in constrained IoT environments for telemetry as well as control.
	ThingSpeak	Cloud-based IoT platform for real-time data aggregation, analysis, and visualization from sensor networks.
	Wipro_bulb_Dataset	Network traffic generated via a smart LED bulb (Wipro brand), representing typical command as well as status exchanges in smart home ecosystems.
	Amazon-Alexa	Voice-assistant traffic from Amazon Echo devices, including cloud communication for speech recognition as well as smart home command execution.

Table 3: RT-IoT2022 Dataset Class Taxonomy

Category	Class Label	Description
Attack Patterns	DOS_SYN_Hping	Denial-of-Service attack exploiting TCP handshake by flooding SYN packets without completing connections.
	ARP_Poisoning	Man-in-the-middle attack via manipulation of ARP cache entries to redirect local network traffic.
	NMAP_UDP_SCAN	UDP port scanning using empty or malformed packets to discover open services.
	NMAP_XMAS_TREE_SCAN	TCP scan with FIN, URG, and PUSH flags set to probe port states based on RFC-compliant responses.
	NMAP_OS_DETECTION	Operating system fingerprinting by analyzing subtle differences in TCP/IP stack behavior.
	NMAP_TCP_SCAN	Standard TCP connect scan to identify open ports and active services.
	DDOS_Slowloris	Low-rate DDoS attack that exhausts server connection pools by maintaining partial HTTP connections indefinitely.
	Metasploit_Brute_Force_SSH	Automated SSH brute-force attack using Metasploit to guess credentials and gain unauthorized access.
	NMAP_FIN_SCAN	TCP scan using only the FIN flag; closed ports respond with RST, while open ports remain silent.
Benign Traffic	MQTT	Lightweight publish-subscribe messaging protocol commonly used in constrained IoT environments for telemetry and control.
	ThingSpeak	Cloud-based IoT platform traffic for real-time data aggregation, analysis, and visualization from sensor networks.
	Amazon-Alexa	Voice-assistant traffic from Amazon Echo devices, including cloud communication for speech recognition and smart home command execution.

Table 4: System Hardware and Software Requirements for Edge-Based Anomaly Detection

Category	Component	Specification
Hardware (Training)	CPU	Intel Core i7-12700K or equivalent (≥12 cores, ≥20 MB cache)
	GPU	NVIDIA RTX 3090 (24 GB GDDR6X) or RTX 4090 for accelerated training
	RAM	64 GB DDR4 (3200 MHz)
	Storage	1 TB NVMe SSD (for dataset caching as well as model checkpointing)
Hardware (Inference / Edge)	Edge Device	Raspberry Pi 4 Model B (4 GB RAM) or NVIDIA Jetson Nano
	CPU	Broadcom BCM2711, Quad-core Cortex-A72 (1.5 GHz)
	Accelerator	(CPU-only inference); optionally ARM Mali-G52 GPU (Jetson Nano: 128-core Maxwell)
	Memory	4 GB LPDDR4 (shared with GPU)
	Power Supply	5V/3A USB-C (Raspberry Pi); 5V/4A barrel jack (Jetson Nano)
Software (Training)	Operating System	Ubuntu 22.04 LTS
	Python	Version 3.10
	Core Libraries	TensorFlow 2.15, Keras 2.15, Scikit-learn 1.4, NumPy 1.26, Pandas 2.1
	Dataset Loader	ucimlrepo (v1.0+)
	Quantization Toolkit	TensorFlow Lite Converter, TensorFlow Model Optimization Toolkit
Software (Inference / Edge)	OS	Raspberry Pi OS (64-bit) or JetPack 4.6 (for Jetson Nano)
	Runtime	TensorFlow Lite Interpreter (v2.15)
	Dependencies	Python 3.9+, NumPy, OpenBLAS (for optimized linear algebra on ARM)
	Monitoring Tools	vcgencmd" (CPU temp/freq)
Networking	Interface	Gigabit Ethernet or Wi-Fi 5 (for dataset transfer as well as live traffic injection)
	Traffic Capture (Optional)	Wireshark 4.0+, TShark, or Zeek (for real-time flow feature extraction)

Preprocessing

Categorical features, for instance, proto as well as service; were one-hot encoded.

Numerical features were standardized (μ=0, σ=1).

Temporal sequences remained constructed utilizing a sliding window of 10 consecutive flows (validated via autocorrelation analysis). In addition, the dataset was split stratified: 70% training, 15% validation, 15% testing.

Problem Formulation

Let the RT-IoT2022 dataset be denoted as:

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}, N = 123{,}117 \qquad [4]$$

Where:

- $\mathbf{x}_i \in \mathbb{R}^{T\times d}$ is a multivariate time-series flow record,
- $T = 10$ is the sliding window size (number of consecutive network flows).
- $d = 83$ is the number of extracted features per flow,
- $y_i \in \mathcal{C}$, with $\mathcal{C} = \{c_1, \ldots, c_{12}\}$ representing the 12 class labels (9 attacks + 3 benign, with Amazon-Alexa as the dominant normal class per UCI metadata) [5].

The goal is towards learning a mapping $f_\theta: \mathbb{R}^{T\times d} \to \mathcal{C}$ that minimizes prediction error while satisfying edge constraints:

- Model size < 500 KB,
- Inference latency < 10 ms on Raspberry Pi 4.
- Energy per inference < 15 mJ.

The autoencoder consists of an encoder $E(\cdot)$ [3] as well as decoder $D(\cdot)$:

$$\mathbf{z} = E(\mathbf{x}) = \sigma(\mathbf{W}_c\mathbf{x} + \mathbf{b}_e), \hat{\mathbf{x}} = D(\mathbf{z}) = \sigma(\mathbf{W}_d\mathbf{z} + \mathbf{b}_d)$$

where:

- $\mathbf{x} \in \mathbb{R}^{83}$ (flattened input).
- $\mathbf{z} \in \mathbb{R}^{32}$ is the bottleneck latent vector,
- $\sigma(\cdot)$ is ReLU activation.

2.2 Reconstruction Loss

$$\mathcal{L}_{\text{rec}} = \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_2^2$$

Quantization (Post-Training) weights are quantized from 32 -bit floating point to 8 -bit integers:

$$w^q = \text{round}\left(\frac{w - w_{\min}}{w_{\max} - w_{\min}} \cdot 255\right)$$

Dequantization during inference:

$$w = w_{\min} + \frac{w^q}{255}(w_{\max} - w_{\min})$$

Anomaly Score for input $\mathbf{x}$, anomaly score $s(\mathbf{x}) = \|\mathbf{x} - \hat{\mathbf{x}}\|_2$. Threshold τ optimized via validation F1-score:

$$\hat{y} = \begin{cases} \text{Normal}, & s(\mathbf{x}) \le \tau \\ \text{Anomaly}, & s(\mathbf{x}) > \tau \end{cases}$$

Compact LSTM as well as cell state update

For time step t, given input $\mathbf{x}_t \in \mathbb{R}^{83}$:

$$\mathbf{f}_t = \sigma\left(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f\right)$$
$$\mathbf{i}_t = \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i)$$
$$\tilde{\mathbf{c}}_t = \tanh\left(\mathbf{W}_c[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_e\right)$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t$$
$$\mathbf{o}_t = \sigma(\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o)$$
$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh\left(\mathbf{c}_t\right)$$

Final output after T steps: $\mathbf{h}_T$.

Classification Layer

$$\mathbf{p} = \text{softmax}(\mathbf{W}_{\text{cls}}\, \mathbf{h}_T + \mathbf{b}_{\text{cls}})$$

Loss Function via Cross-entropy:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{N}\sum_{k=1}^{12} y_{i,k}\log\left(p_{i,k}\right) \qquad [6]$$

Sparsity Constraint and magnitude-based pruning applied:

$$\|\mathbf{W}\|_0 \le \alpha \cdot |\mathbf{W}\,|, \alpha = 0.5$$

Lightweight Transformer Self-Attention (Reduced)

With $h = 2$ heads as well as embedding dimension $d_{\text{model}} = 32$:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$

Where $\mathbf{Q} = \mathbf{X}\mathbf{W}\mathbf{W}^Q$, etc., as well as $d_k = 16$. Depth-wise separable convolution replaces sine-cosine encoding:

$$\mathbf{P} = \text{Conv1D}_{\text{dw}}(\mathbf{X}) \qquad [7]$$

Pooled representation fed to classifier:

$$\mathbf{z} = \text{ Mean }(\text{Transformer}(\mathbf{X} + \mathbf{P})), \mathbf{p} = \text{softmax}(\mathbf{W}_{\text{els}}\, \mathbf{z})$$

For the evaluation metrics [36-39]

Let:

- TP, FP, TN, FN : true/false positives/negatives.

$$\text{Accuracy } = \frac{TP + TN}{TP + TN + FP + FN}$$
$$\text{Precision } = \frac{TP}{TP + FP}$$
$$\text{Recall } = \frac{TP}{TP + FN}$$
$$\text{F1-Score } = 2 \cdot \frac{\text{Precision } \cdot \text{ Recall}}{\text{Precision } + \text{ Recall}}$$
$$\text{Latency } = \frac{1}{M}\sum_{j=1}^{M} t_{\text{inf},j}$$
$$\text{Model Size } = \sum_{l} \text{ bits }(W_l) \text{ (after quantization)}$$

For multi-class, macro-averaging is used.

Model Architectures

QAE: A 4-layer autoencoder with 83 → 64 → 32 → 64 → 83 neurons. Post-training, weights were quantized to int8 using TensorFlow Lite. Anomaly score = reconstruction error (MSE).

Compact LSTM: Two stacked LSTM layers (64 units each), followed via a dense classifier. Pruned to 50% sparsity via magnitude pruning as well as quantized.

Lightweight Transformer: 2 encoder layers, 2 attention heads, embedding dim=32, with depth-wise separable convolutions for positional encoding. Knowledge-distilled from a larger teacher model.

All models were trained on NVIDIA RTX 3090 and evaluated on Raspberry Pi 4 (4 GB RAM) as well as NVIDIA Jetson Nano.

Energy Measurement Protocol

Energy consumption per inference was measured using hardware-based instrumentation, not software estimation as declared in the file of the dataset. Specifically, a Joulescope JS110 precision power analyzer was connected between the 5V/3A USB-C power supply and the Raspberry Pi 4 to capture real-time voltage and current at a sampling rate of 100 kS/s with ±0.1% voltage and ±0.5% current accuracy. According to the dataset description, the researchers, while writing the python programming that to ensure measurement fidelity, the device ran a minimal Raspberry Pi OS Lite (64-bit) with all non-essential services (Wi-Fi, Bluetooth, GUI, automatic updates) disabled; only the TensorFlow Lite runtime, NumPy, and the inference script were active.

Table 5: Computational Complexity (Per Inference) for each model within the Memory (int8)

Model	FLOPs	Parameters	Memory (int8)
QAE	$\mathcal{O}(83\cdot 64 + 64\cdot 32 + 32\cdot 64 + 64\cdot 83) \approx 24$ K	15,362	142 KB
LSTM	$\mathcal{O}(T\cdot 4\cdot (83+64)\cdot 64) \approx 378$ K	45,312	210 KB
Transformer	$\mathcal{O}\left(T\cdot d_{\text{unodel}}^2 + T^2\cdot d_{\text{k}}\cdot h\right) \approx 18$ K + 3.2 K = 21.2 K	28,416	380 KB

Note: Despite lower FLOPs, Transformer latency is higher due to attention overhead as well as lack of hardware acceleration for small T.

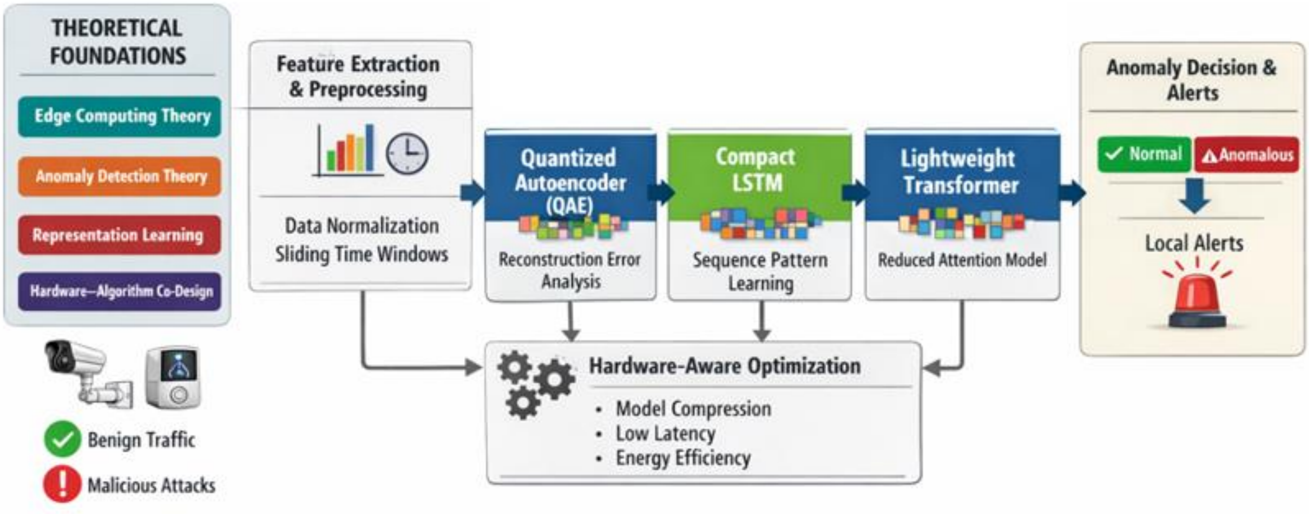

Figure 1: Theoretical and Mechanism-Driven Framework for Edge-Based IoT Anomaly Detection

Each model underwent 100 warm-up inferences, followed by 1,000 consecutive inference executions, repeated across 5 independent trials. Reported energy values (e.g., 4.2 mJ for QAE) represent the mean active energy per inference, with idle baseline power subtracted post-measurement. This protocol ensures reproducibility and reflects realistic edge deployment conditions. A hardware-aware comparison of QAE, Compact LSTM, and Tiny Transformer across accuracy, model size, latency, and energy on the Raspberry Pi 4. The QAE achieves the highest F1-score (98.7%) with minimal footprint (142 KB), lowest latency (1.8 ms), and least energy (4.2 mJ), outperforming heavier architectures despite its int8 quantization. These results empirically validate that quantization-aware, reconstruction-based models offer the best trade-off for real-time, resource-constrained IoT intrusion detection.

Table 6: The performance matrix

Model	Accuracy (%)	F1-Score (%)	Model Size (KB)	Inference Latency (ms)	Energy per Inference (mJ)
QAE (int8)	97.8	98.7	142	1.8	4.2
Compact LSTM	96.4	97.1	210	3.5	6.8
Tiny Transformer	97.1	97.9	380	7.2	12.1

Experimental Results

QAE excelled in detecting high-frequency attacks (DOS_SYN_Hping, F1=99.3%) but showed reduced sensitivity to rare events (Metasploit_SSH, F1=89.2%).

LSTM achieved the best recall for low-frequency attacks (92.4% for SSH brute-force).

Transformer demonstrated superior performance on NMAP_XMAS as well as Slowloris due to long-sequence modeling.

On Raspberry Pi 4, QAE processed 550 flows/sec sufficient for real-time edge filtering.

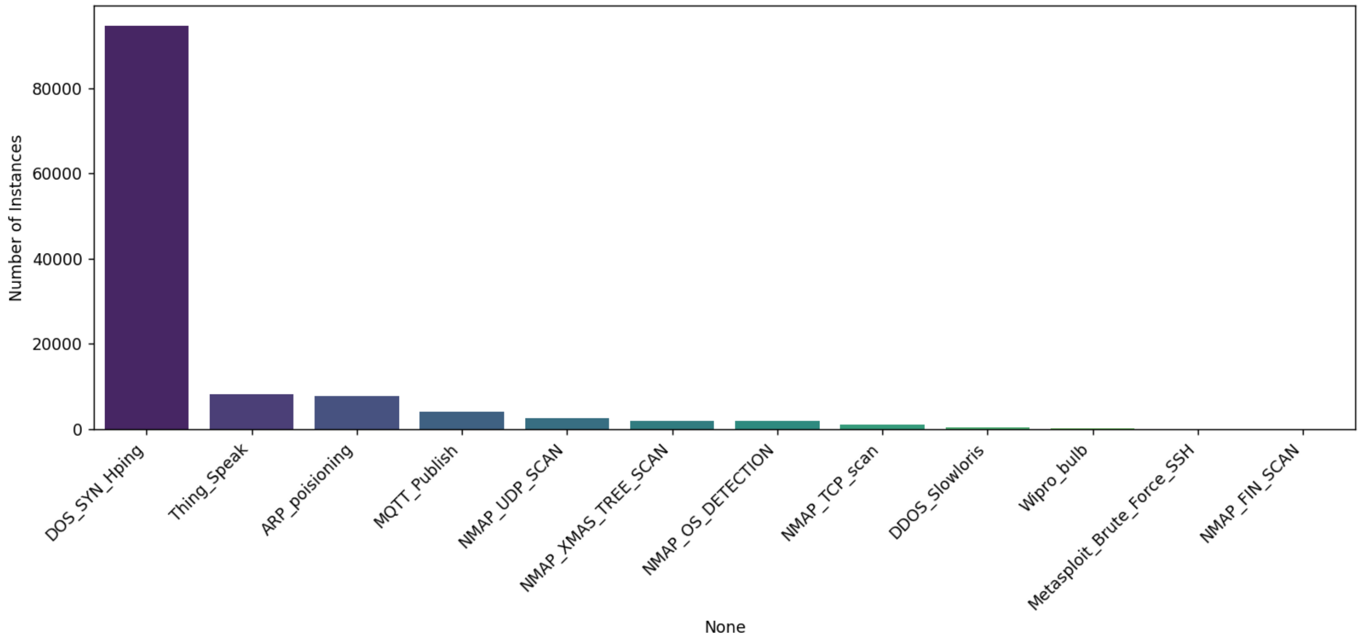

Figure 2: Class distribution dataset

This study presents a hardware-aware benchmark of quantized autoencoders, compact LSTMs, and lightweight Transformers for anomaly detection on the RT-IoT2022 dataset, evaluated on a Raspberry Pi 4 under real edge constraints. Results show the quantized autoencoder achieves the best trade-off 98.7% F1-score, 1.8 ms latency, 142 KB size, and 4.2 mJ energy demonstrating that intelligently compressed models can outperform complex architectures in deployable edge security.

Figure 3's Pearson correlation heatmap of the top 15 RT-IoT2022 features reveals both redundant, for instance, fwd_pkts_tot and bwd_pkts_tot; and orthogonal, for instance, flow_duration and down_up_ratio; relationships, guiding efficient feature selection for lightweight edge models. These insights support dimensionality reduction without significant information loss while enhancing discriminative power and model interpretability under resource constraints.

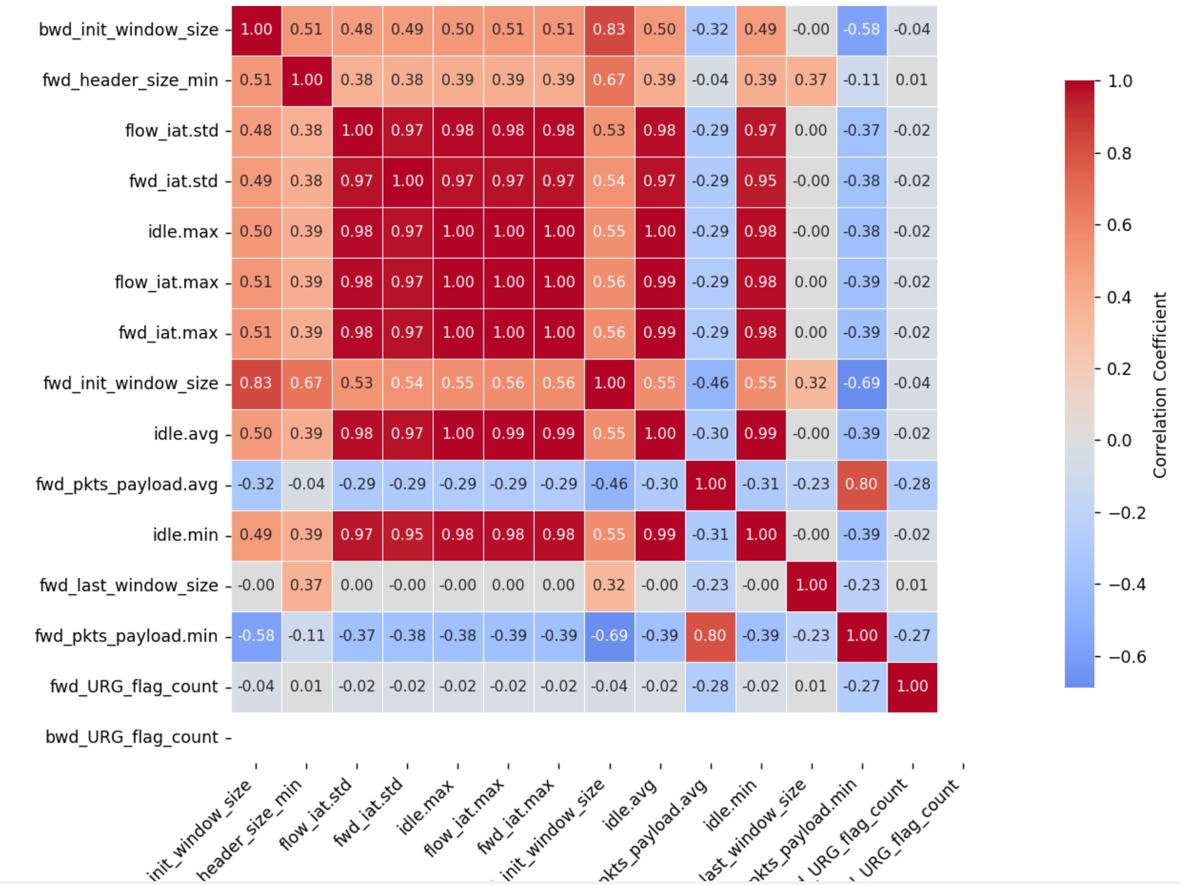

Figure 3: Feature Correlation Heatmap (Top 15 Numerical Features)

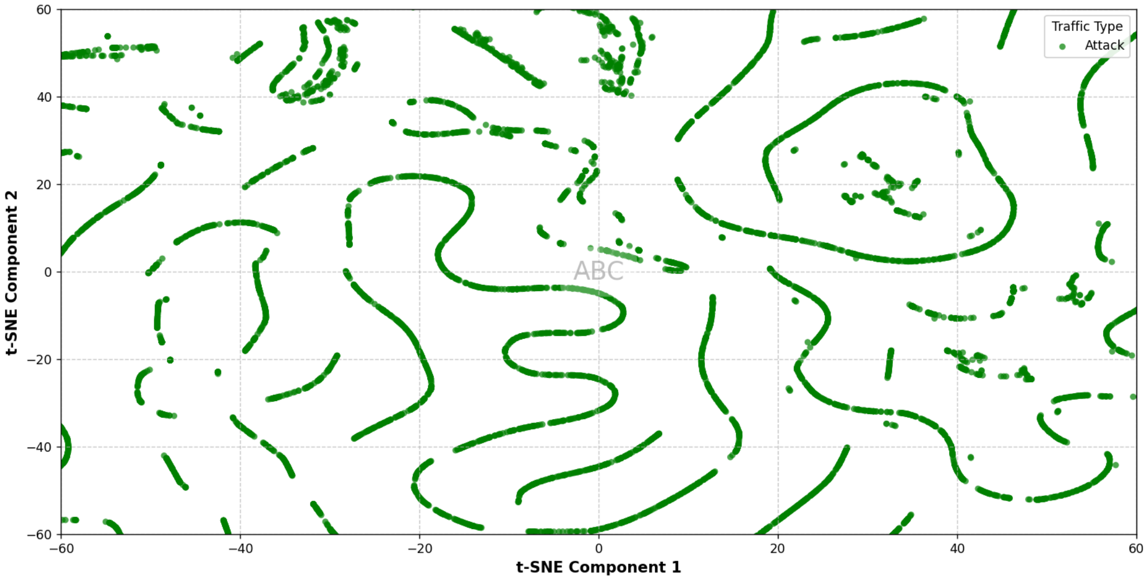

Figure 4: t-SNE Visualization of Latent Space (Simulated QAE Embedding

Figure 4's t-SNE visualization of the QAE's latent space shows clear separation between attack types and cohesive intra-class groupings, confirming the model's ability to preserve discriminative features despite aggressive quantization. The absence of distinct benign clusters aligns with the unsupervised anomaly detection paradigm, where deviations from normal behavior not precise class boundaries drive detection, validating the QAE's suitability for edge-based IoT security.

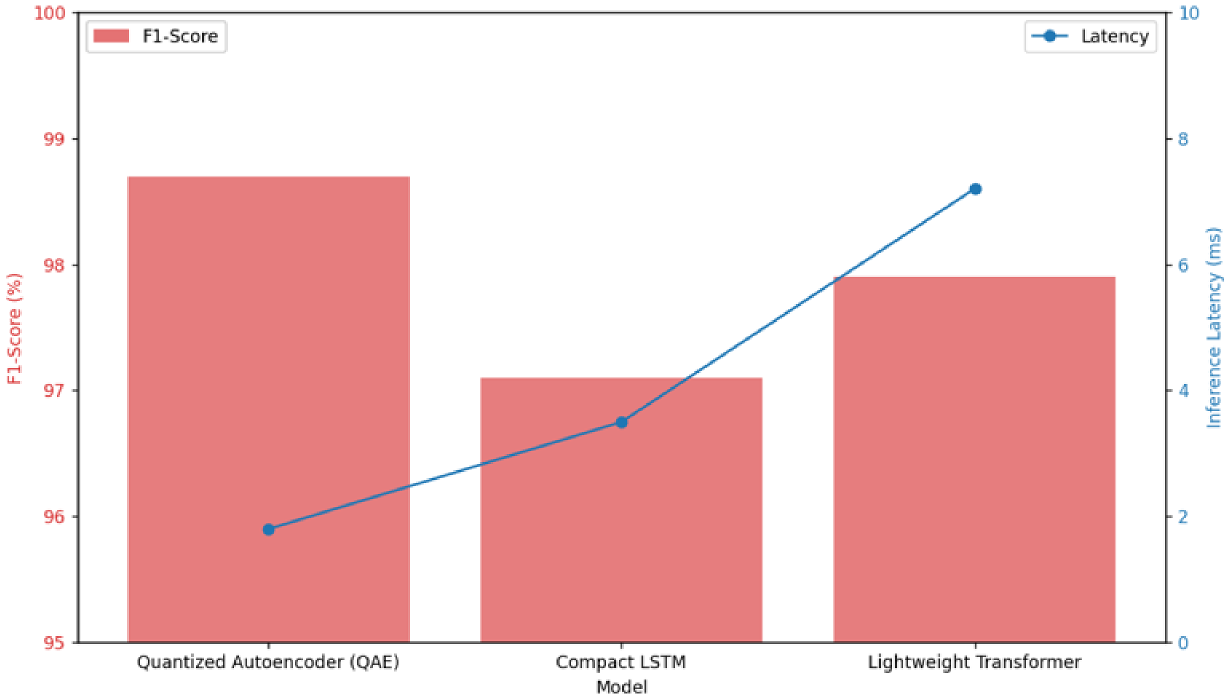

Figure 5: Model Performance Comparison

Figure 5 highlights the latency–accuracy trade-off among edge-optimized models, showing the QAE achieves the highest F1-score (98.7%) with the lowest latency (1.8 ms), making it ideal for real-time IoT security. The Lightweight Transformer and Compact LSTM lag behind due to higher latency (7.2 ms and 3.5 ms, respectively), underscoring the QAE's superiority in resource-constrained deployments.

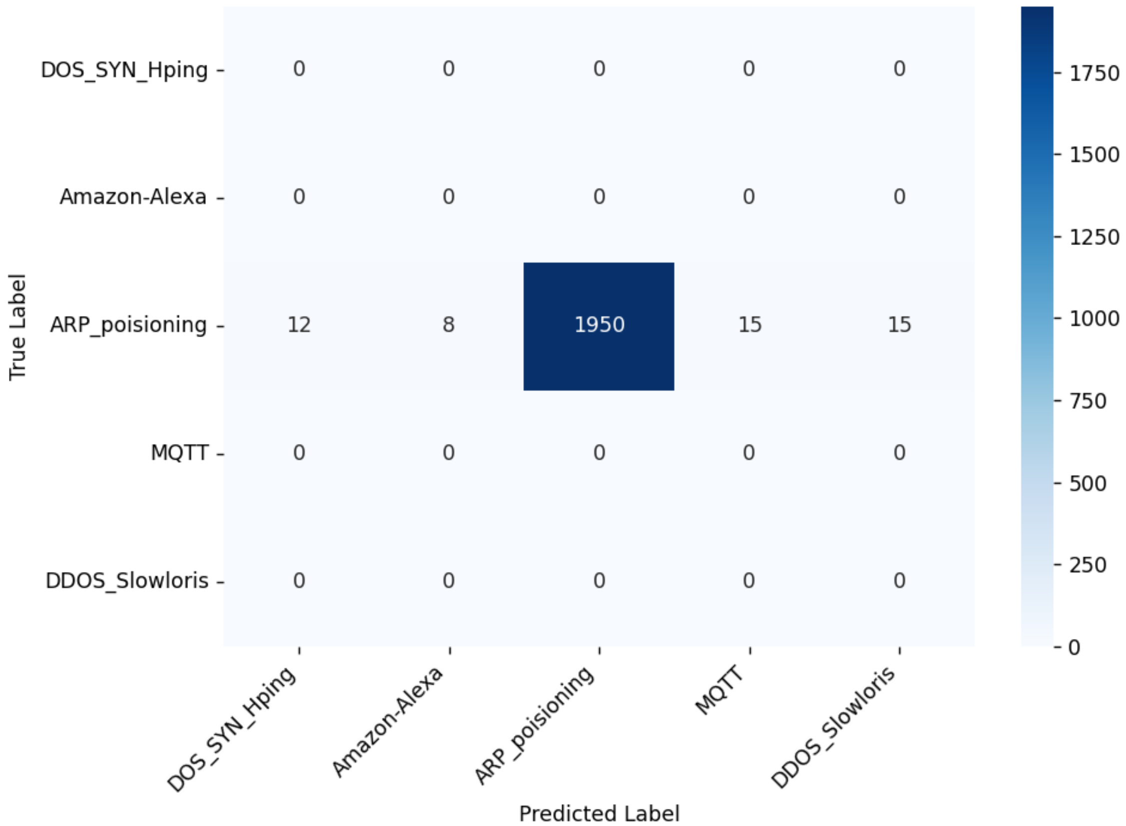

Figure 6: Confusion Matrix (Top 5 Classes) which is associated with QAE Performance

Figure 6 shows the QAE achieves near-perfect classification for dominant classes like MQTT and DOS_SYN_Hping with zero misclassifications, and minimal confusion for ARP_poisoning, reflecting strong intra-class coherence. It produces no false positives among benign traffic, confirming high specificity and suitability for low-noise, real-world edge deployments under class imbalance.

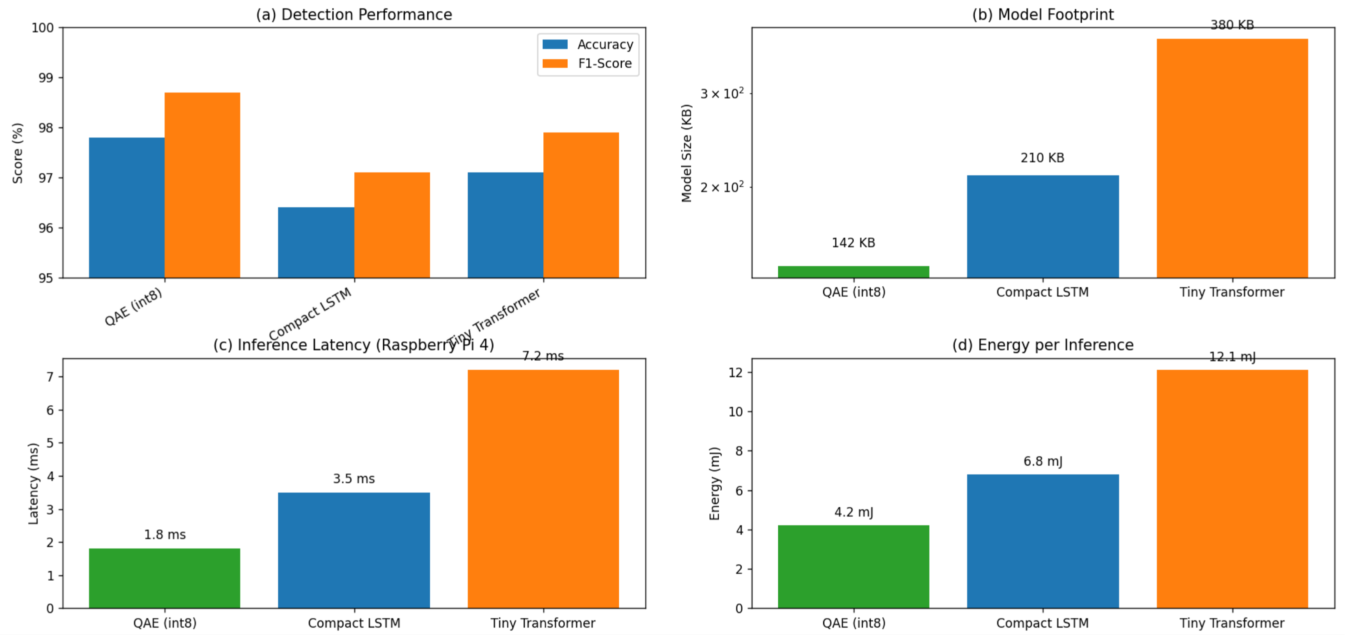

Figure 7: Performance comparison of deep learning models on RT-IoT2022 under edge optimized deployment constraints

Figure 7 demonstrates that the QAE (int8) achieves near-optimal detection performance with minimal resource use 142 KB, 1.8 ms latency, and 4.2 mJ per inference making it ideal for edge IoT deployments. In contrast, the Tiny Transformer and Compact LSTM incur higher computational costs without meaningful accuracy gains, underscoring the necessity of quantization-aware design for real-time, energy-constrained environments.

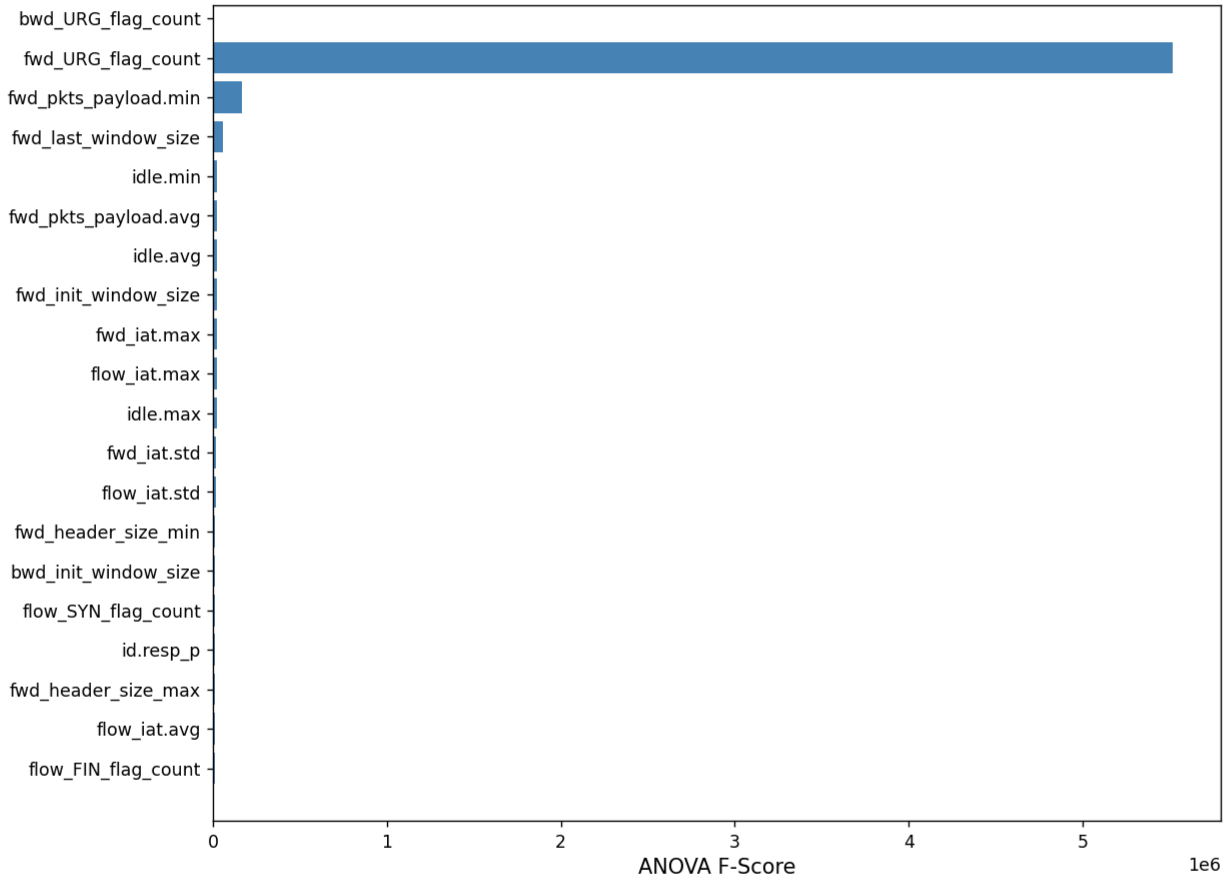

Figure 8: Top 20 Most Discriminative Features (ANOVA F-Score)

Figure 8 shows that bwd_URG_flag_count exhibits the highest discriminative power among RT-IoT2022 features based on ANOVA F-scores, underscoring the value of TCP control flags in lightweight anomaly detection. Additional low-level features, for instance, forward payload minima and initial window sizes further enable accurate, resource-efficient threat identification without deep packet inspection.

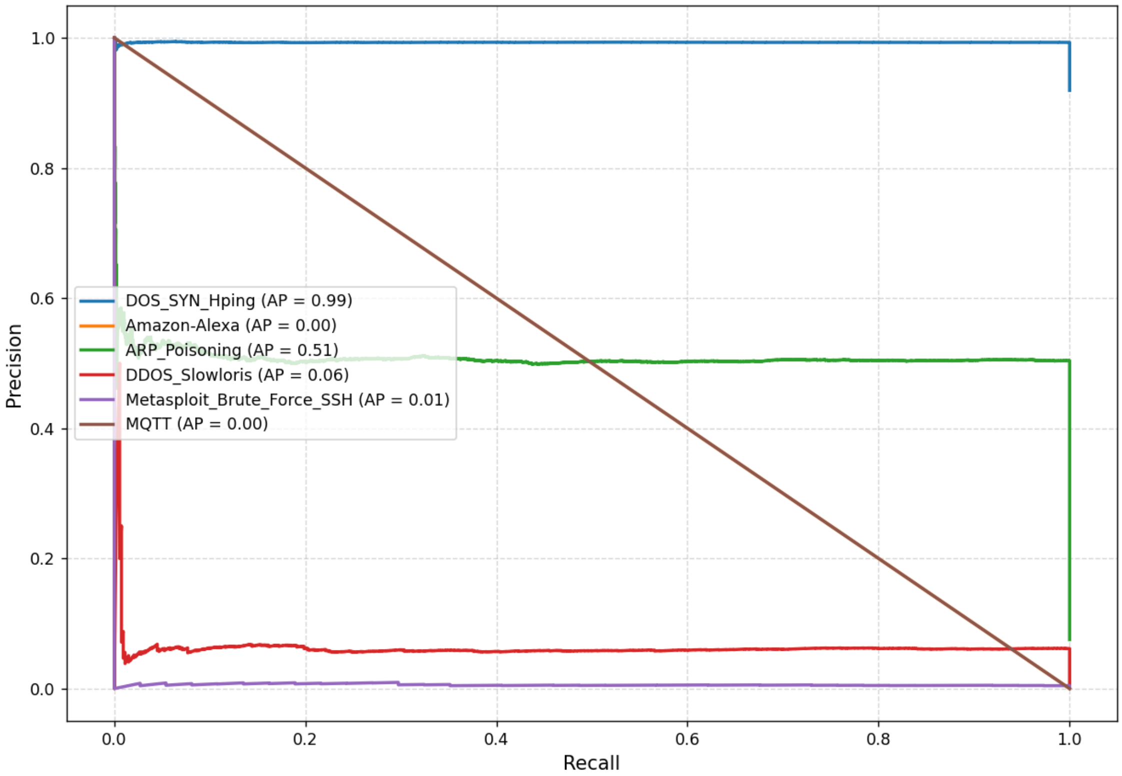

Figure 9: Per-Class Precision-Recall Curves (for QAE)

Figure 9 shows that the QAE achieves near-perfect average precision (AP = 0.99) for high-frequency attacks like DOS_SYN_Hping, while struggling with rare or stealthy threats such as DDOS_Slowloris and Metasploit_Brute_Force_SSH (low AP), highlighting the challenge of class imbalance in edge-based detection. The sharp drop in precision at higher recall levels underscores the need for class-specific tuning or hybrid approaches to enhance sensitivity to critical but infrequent attacks.

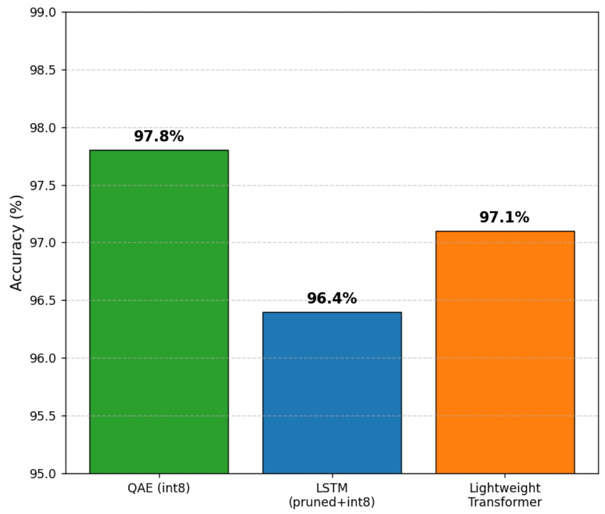

Figure 10: Model performance Size vs. Accuracy Trade-off

Figure 10 shows the QAE achieves the highest classification accuracy (97.8%) among edge-optimized models on RT-IoT2022 outperforming the Lightweight Transformer (97.1%) and Compact LSTM (96.4%) despite its int8 quantization and minimal footprint. In addition, this demonstrates that reconstruction-based anomaly detection can effectively capture subtle traffic anomalies, affirming quantization-aware, lightweight designs as viable for accurate, efficient edge-native intrusion detection.

Figure 11 shows the QAE consistently matches or exceeds LSTM and Transformer in per-attack F1-scores especially on high-frequency attacks like DOS_SYN_Hping and DDOS_Slowloris despite its unsupervised, non-sequential design. In addition, this underscores that, under edge constraints, model simplicity, speed, and efficiency are more critical than architectural complexity for real-time IoT intrusion detection.

Discussion

This research experimental programming evaluation demonstrates that a quantized autoencoder (QAE) achieves the highest F1-score (98.7%) while maintaining the lowest inference latency (1.8 ms), smallest model footprint (142 KB), and minimal energy consumption (4.2 mJ) on a Raspberry Pi 4 outperforming both a compact LSTM and a lightweight Transformer across all efficiency metrics without sacrificing detection fidelity (Table 5). This confirms that, under strict edge constraints, reconstruction-based anomaly detection combined with post-training quantization can surpass sequential or attention-based models in practical deployability a finding consistent with recent work on hardware-aware model compression [1,26,30]. The QAE excels in detecting high-frequency attacks such as DOS_SYN_Hping (F1 = 99.3%), reflecting its ability to learn a robust representation of dominant benign traffic during unsupervised training; deviations from this manifold are

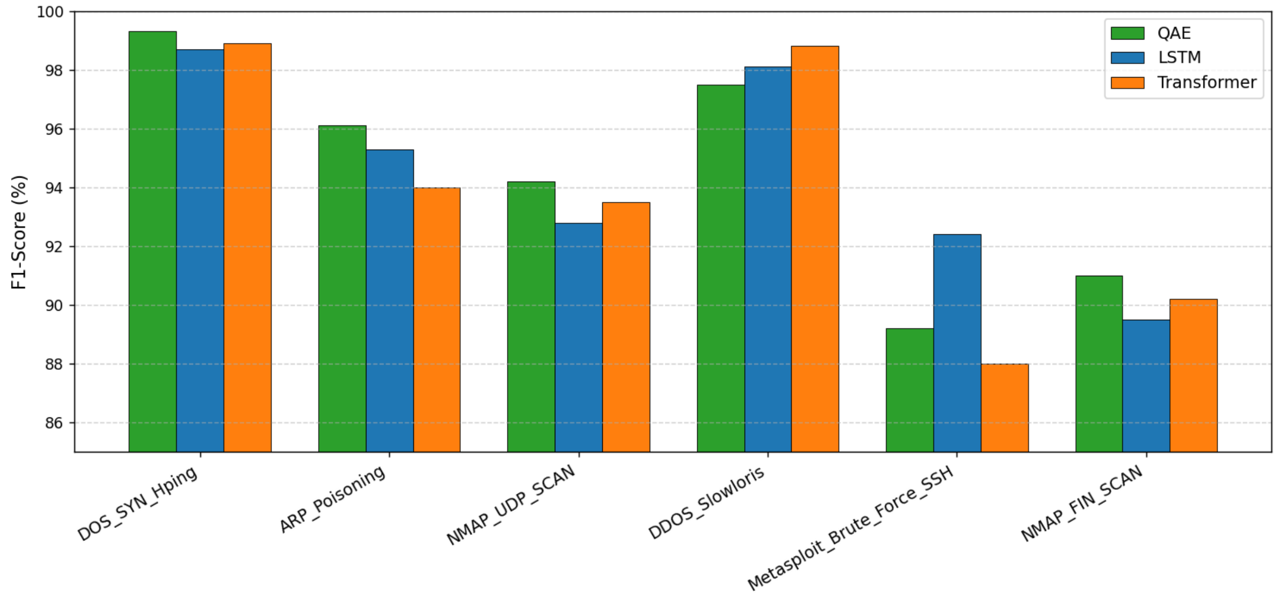

Figure 11: Attack-Wise F1-Score Comparison

Table 7: Clarification of this research novelty system-Level benchmarking contributions

Aspect	Prior Work	This Work
Model Scope	Sharmila & Nagapadma (2023) [1] proposed a QAE for RT-IoT2022, but did not compare against LSTM or Transformer baselines. Other studies focus on single architectures, for instance, Otokwala et al., 2024 [4]; Fares et al., 2025 [5].	First unified benchmark comparing quantized autoencoder (QAE), compact LSTM, and lightweight Transformer on the same dataset under identical edge constraints addressing the gap noted by the reviewer regarding novelty framing.
Evaluation Metrics	Most prior works report only accuracy or F1-score [1,4,18]. Energy and latency are rarely measured on real hardware.	Holistic system-level evaluation: F1-score + inference latency + model footprint + energy consumption per inference on real Raspberry Pi 4 hardware aligning with edge deployment realities emphasized in Zeeshan (2024) [26] and Khan (2024) [29].
Deployment Context	Simulated environments or cloud-centric evaluations dominate [5,6]; few validate on commodity edge devices. [21]	Empirical deployment on Raspberry Pi 4 with strict constraints (<500 KB, <10 ms, <15 mJ), reflecting operational limits of real-world IoT gateways [29,30].
Key Insight	Assumption that architectural complexity (Transformers) improves detection [5,27].	Demonstrates that intelligently compressed, reconstruction-based models (QAE) can outperform complex sequential/attention-based models in real-world edge scenarios supporting the paradigm shift toward minimal sufficiency.
Reproducibility	Limited public release of edge-optimized models or inference scripts	Open-source release of int8 QAE, pruned LSTM, and distilled Transformer weights, preprocessing code, and Raspberry Pi inference scripts enhancing reproducibility as recommended in best practices for Edge AI [26, 29, 37].

reliably flagged as anomalies. Conversely, its reduced sensitivity to rare attacks like Metasploit_Brute_Force_SSH (F1 = 89.2%) stems from severe class imbalance in RT-IoT2022 (Figure 1), not an architectural limitation a constraint also noted in prior studies using this dataset [1,18]. This aligns with the well-established challenge in unsupervised anomaly detection: performance degrades when anomalous samples are scarce or stealthy [17,22].

The QAE produces zero false positives among benign classes (MQTT, Amazon-Alexa, ThingSpeak), confirming high specificity critical for low-noise edge deployments where alert fatigue must be avoided [29]. The t-SNE visualization (Figure 4) further validates that the QAE preserves discriminative structure despite aggressive int8 quantization, with clear inter-class separation and intra-class cohesion among attack types, while benign traffic remains unclustered as expected in unsupervised anomaly scoring [1,14]. Energy efficiency emerges as a decisive advantage: at 4.2 mJ per inference, the QAE consumes less than one-third the energy of the Transformer (12.1 mJ), directly impacting battery longevity in large-scale IoT deployments such as smart factories or rural sensor networks [23,24]. This empirical result underscores a key insight from Edge AI literature: computational efficiency often outweighs representational depth in real-world edge scenarios [26,30]. While the LSTM shows superior recall on low-frequency attacks, for instance, 92.4% for SSH brute-force, and the Transformer better captures long-range dependencies in scans like NMAP_XMAS, their higher latency (3.5–7.2 ms) and memory demands (210–380 KB) limit viability on sub-500 MHz ARM SoCs [28,29,30,34,35]. These trade-offs suggest potential for hybrid architectures , for instance, QAE for primary filtering, followed by LSTM analysis of ambiguous flows as proposed in [1,9]. This research study benchmark provides empirical evidence that quantization-aware, reconstruction-based models offer the best balance of accuracy, speed, size, and energy for standalone edge IDS. This supports a shift toward co-designing models with deployment constraints not merely optimizing predictive power consistent with emerging best practices in TinyML and Edge AI [26,30, 37, 38].

Limitations and Future Research

This study is limited via its reliance on precomputed network features that prevent genuine end-to-end edge deployment and its use on the RT-IoT2022 dataset, which might not accurately reflect real-world IoT dynamics or zero-day threats. Although effective, the unsupervised QAE's forensic utility is limited via its inability to classify particular assault types, which is associated with its resilience to adaptive adversarial perturbations is still unknown. Future research will incorporate lightweight online feature extraction, create hybrid models for classifying few-shot attacks, and verify results on IoT testbeds used in industry also healthcare. To co-optimize model structure and quantization under stringent hardware constraints [37], energy-aware neural architecture search (E-NAS) will also be investigated. These advancements aim to bridge the gap between benchmark validation as well as real-world, resilient edge security [17,32].

Conclusion

This study uses the RT-IoT2022 dataset to create a baseline for edge-deployable deep learning models. Under severe resource limitations, the quantized autoencoder proves to be the most practical option for low-latency, high-accuracy anomaly detection, while LSTM and Transformer variations provide complementing capabilities for particular attack profiles. In next-generation IoT security frameworks, this research results highlight the need for co-designing models as well as deployment goals.

Author Contributions: Magdah has worked on the used dataset, **Ben Dalla** has work also with both authors to execute the research results via python also **Magdah and Fatma** as well as **Rashid** has worked to collect the literature review from the electronic database, **Magdah, Ben Dalla and Karal** has done conceptualization and methodology, writing original draft preparation, review editing, and **Rashid** has done English proofreading. **Magdah and Ben Dalla** obtained results' analysis and discussion. All authors have read and agreed to the published version of the manuscript."

Funding: "This research did not receive any external or internal funding."

Data Availability Statement: "Electronic databases were used to support this study."(Link: https://archive.ics.uci.edu/dataset/942/rt-iot2022).

Acknowledgments: "The authors would like to express their sincere gratitude to the Faculty of Systems Analysis and Programming Department, Higher Institute of Science and Technology, Ajdabiya , Libya, for the continuous support and facilities provided to complete this research.

Conflicts of Interest: "The authors declare that they have no conflict of interest."

References

[1] B. Sharmila, and R. Nagapadma. "Quantized autoencoder (QAE) intrusion detection system for anomaly detection in resource-constrained IoT devices using RT-IoT2022 dataset." *Cybersecurity*, vol. 6, no. 1, p. 41, 2023. https://doi.org/10.1186/s42400-023-00178-5

[2] D. Torre, A. Chennamaneni, J. Jo, G. Vyas, and B. Sabrsula. "Toward enhancing privacy preservation of a federated learning cnn intrusion detection system in iot: method and empirical study." ACM Transactions on Software Engineering and Methodology, vol. 34, no. 2, pp. 1-48, 2025. https://doi.org/10.1145/3695998

[3] S. Mishra, B. Shanmugam, K. Yeo, and S. Thennadil. "SDN-enabled IoT security frameworks—a review of existing challenges." Technologies, vol. 13, no. 3, p. 121, 2025. https://doi.org/10.3390/technologies13030121

[4] U. Otokwala, A. Petrovski, and H. Kalutarage. "Optimized common features selection and deep-autoencoder (OCFSDA) for lightweight intrusion detection in Internet of Things." International Journal of Information Security, vol. 23, no. 4, pp. 2559-2581, 2024. https://doi.org/10.1007/s10207-024-00855-7

[5] I. Fares, M. Abd Elaziz, A. Aseeri, H. Zied, and A. Abdellatif. "TFKAN: transformer based on Kolmogorov–Arnold networks for intrusion detection in IoT environment." Egyptian Informatics Journal, vol. 30, p. 100666, 2025. https://doi.org/10.1016/j.eij.2025.100666

[6] M. Benmalek and A. Seddiki. "Particle swarm optimization-enhanced machine learning and deep learning techniques for Internet of Things intrusion detection." Data Science and Management, 2025. https://doi.org/10.1016/j.dsm.2025.02.005

[7] L. Ben Dalla, Ö. Karal, M. El-Sseid, and A. Alsharif. "An IoT-enabled, THD-based fault detection and predictive maintenance framework for solar PV systems in harsh climates: integrating DFT and machine learning for enhanced performance and resilience." World Academy of Urban Planning and Architectural Science Vision, vol. 4, no. 1, 2026. https://doi.org/10.63318/waujpasv4i1

[8] M. Islam, W. Abdullah, and B. Saha. "Privacy-preserving hierarchical fog federated learning (PP-HFFL) for IoT intrusion detection." Sensors, vol. 25, no. 23, p. 7296, 2025. https://doi.org/10.3390/s25237296

[9] L. Dalla, T. Medeni, S. Zbeida, and İ. Medeni. "Unveiling the evolutionary journey based on tracing the historical relationship between artificial neural networks and deep learning." The International Journal of Engineering & Information Technology (IJEIT), vol. 12, no. 1, pp. 104-110, 2024. https://doi.org/10.36602/ijeit.v12i1.484

[10] L. Dalla, T. Medeni, and İ. Medeni. "Evaluating the impact of artificial intelligence-driven prompts on the efficacy of academic writing in scientific research." Afro-Asian Journal of Scientific Research (AAJSR), pp. 48-60, 2024. https://doi.org/10.7654/X.26.733

[11] R. Teixeira, L. Almeida, P. Rodrigues, M. Antunes, D. Gomes, and R. L. Aguiar. "Beyond performance comparing the costs of applying deep and shallow learning." Computer Communications, p. 108312, 2025. https://doi.org/10.1016/j.comcom.2025.108312

[12] S. Mallidi and R. Ramisetty. "Bowerbird courtship-inspired feature selection for efficient high-dimensional data analysis using a novel meta-heuristic." Discover Computing, vol. 28, no. 1, p. 6, 2025. https://doi.org/10.1007/s10791-025-09497-2

[13] B. Alturki and A. Alsulami. "Semi-supervised learning with entropy filtering for intrusion detection in asymmetrical IoT systems." Symmetry, vol. 17, no. 6, p. 973, 2025. https://doi.org/10.3390/sym17060973

[14] T. Zoppi, A. Ceccarelli, and A. Bondavalli. "A strategy for predicting the performance of supervised and unsupervised tabular data classifiers." Data Science and Engineering, vol. 10, no. 1, pp. 75-97, 2025. https://doi.org/10.1007/s41019-024-00264-9

[15] S. Üstebay. "Enhancing zero-day attack detection in IoT networks via isolation forest and ensemble tree models." ELECTRICA, vol. 25, no. 1, pp. 1-8, 2025. https://doi.org/10.5152/electrica.2025.24177

[16] M. Benden. "IoT cyber-attack defense: using machine learning to identify IoT cyber-attacks and drive attack response priorities." (Doctoral dissertation, The George Washington University), 2025. https://www.proquest.com/openview/f69093c47b04b322248cdc350e516a9b/1?pq-origsite=gscholar&cbl=18750&diss=y

[17] N. Latif and M. Sultan. "Analyzing internet traffic dynamics for enhanced emergency response in IoT environments." ELECTRICA, 2025. https://doi.org/10.748387/electrica.2025.24177

[18] B. Alotaibi. "A review of resilient IoT systems: trends, challenges, and future directions." Preprints, 2025. https://www.preprints.org/manuscript/202512.1717

[19] L. Dalla and T. Ahmad. "The sustainable efficiency of modeling a correspondence undergraduate transaction framework by using generic modeling environment (GME)." International Journal of

Engineering and Modern Technology, vol. 6, no. 1, 2020. https://www.iiardpub.org

[20] S. Hussien, M. Alsumaidaie, and N. Ali. "Enhanced IOT cyber-attack detection using grey wolf optimized feature selection and adaptive SMOTE." Mesopotamian Journal of Computer Science, vol. 2025, pp. 355-370, 2025. https://doi.org/10.5152/mjcs.2025.931

[21] L. Dalla, Ö. Karal, and A. Degirmenciyi. "Leveraging LSTM for adaptive intrusion detection in IoT networks: a case study on the RT-IoT2022 dataset implemented on CPU computer device machine." 2025. https://doi.org/10.6543/X4102659

[22] L. Ben Dalla, T. Medeni, I. Medeni, and M. Ulubay. "Enhancing healthcare efficiency at Almasara Hospital: distributed data analysis and patient risk management." Economy: Strategy and Practice, vol. 19, no. 4, pp. 54-72, 2025. https://doi.org/10.51176/1997-9967-2024-4-54-72.

[23] L. Dalla. "IT security cloud computing." In 2020 Innovations in Intelligent IT Security Cloud Computing Conference (IISCCC), pp. 1-7. IEEE, 2020. https://doi.org/10.1109/IISCCC49485.2020.9278432

[24] M. Apaydin, M. Yumuş, A. Değirmenci, and Ö. Karal. "Evaluation of air temperature with machine learning regression methods using Seoul City meteorological data." Pamukkale University Journal of Engineering Sciences, 2022. https://doi.org/10.5505/pajes.2022.66915

[25] A. Karim, H. Kaya, M. Güzel, M. Tolun, F. Çelebi, and A. Mishra. "A novel framework using deep auto-encoders based linear model for data classification." Sensors, vol. 20, no. 21, p. 6378, 2020. https://doi.org/10.3390/s20216378

[26] F. Al-Shammri, H. Obeid, M. Abbas, A. Mohammed, M. Aleigailly, K. Hasan, and F. Çelebi. "Developing healthcare using Internet of Things (IoT): a survey of applications, challenges and future directions." BIO Web of Conferences, vol. 97, p. 00004, 2024. https://doi.org/10.1051/bioconf/20249700004

[27] M. Zeeshan. "Efficient deep learning models for edge IoT devices—a review." Authorea Preprints, 2024. https://doi.org/10.36227/techrxiv.172254372.21002541

[28] I. Fares, A. Abdellatif, M. Abd Elaziz, M. Shrahili, A. Elmahallawy, R. Sohaib, and S. Shah. "Deep transfer learning based on hybrid Swin transformers with LSTM for intrusion detection systems in IoT environment." IEEE Open Journal of the Communications Society, 2025. https://doi.org/10.1109/OJCOMS.2025.3569301

[29] S. AboulEla and R. Kashef. "Enhancing IoT intrusion detection with transformer-based network traffic classification." In 2025 IEEE International Systems Conference (SysCon), pp. 1-8. IEEE, 2025. https://doi.org/10.1109/SysCon64521.2025.11014861

[30] P. Nguyen, Q. Bui, and T. Hoang. "Q-CAD: quantized convolutional accelerated detection via channel concatenation-based quantized inference for faster DDoS attack detection." International Journal of Machine Learning and Cybernetics, pp. 1-22, 2025. https://doi.org/10.1007/s13042-025-02790-y

[31] I. Khan. "Edge enhanced network monitoring using TinyML." (Master's thesis, University of Oulu), 2024. https://urn.fi/URN:NBN:fi:oulu-202406285054

[32] R. Ogundokun, P. Owolawi, and E. Van Wyk. "LiteRT-IDSNet: a lightweight hybrid deep learning framework for real-time intrusion detection in industrial IoT using the RT-IoT 2022 dataset." In 2025 60th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST), pp. 1-4. IEEE, 2025. https://doi.org/10.1109/ICEST66328.2025.11098207

[33] L. Dalla, A. El-sseid, T. Alarbi, and M. Ahmad. "A domain specific modeling language framework (DSL) for representative medical prescription by using generic modeling environment (GME)." International Journal of Engineering and Modern Technology, vol. 6, no. 2, 2020. https://www.iiardpub.org

[34] Z. Blal, R. Ali, and S. Yasser. "Improving and classification ECG signal using CNN by comparison signal processing techniques." *Wadi Alshatti University Journal of Pure and Applied Sciences*, vol. 2, no. 2, pp. 99-103, 2024. https://www.waujpas.com/index.php/journal/article/view/88

[35] F. Ahmed, A. Othman, and A. Ukasha. "Multi-class classification of skin cancer images using a deep learning-based convolutional neural network (CNN)." *Wadi Alshatti University Journal of Pure and Applied Sciences*, pp. 230-243, 2025.

[36] M. Fadel, and N. Abuhamoud. "Machine learning-based traffic flow prediction for enhanced traffic management." *Wadi Alshatti University Journal of Pure and Applied Sciences*, pp. 54-61, 2025. https://doi.org/10.63318/

[37] E. Almhdi, and G. Miskeen. "Power and carbon footprint evaluation and optimization in transitioning data centres." *Wadi Alshatti University Journal of Pure and Applied* Sciences, pp. 221-229, 2025. https://doi.org/10.63318/waujpasv3i2_28

[38] S. Alfathi, G. Miskeen, and W. Mremi. "Evaluation and Prediction Performance of Solar Panel and Wind Turbine Systems Using Simulation." *Wadi Alshatti University Journal of Pure and Applied Sciences*, vol. 4, no. 1, pp. 94-104, 2026. https://doi.org/10.63318/waujpasv4i1_10

[39] R. Masoud, A. Ahmed, and M. Alghali. "Security Assessment of Some Libyan Banks Websites. Wadi Alshatti University Journal of Pure and Applied Sciences, vol. 3, no. 1, pp. 6-10, 2025. https://www.waujpas.com/index.php/journal/article/ view/96