American Journal of Neural Networks and Applications
2026, Vol. 12, No. 2, pp. 87–98
https://doi.org/10.11648/j.ajnna.20261202.14

SciencePG
Science Publishing Group

Research Article

Secure Deployment of Neural Networks in Edge and IoT Devices

Daniel Ekpah[1], Friday Oodee Philip-Kpae[2, *], Agbotiname Lucky Imoize[3]

[1]Electrical and Electronics Engineering, University of Port Harcourt, Port Harcourt, Nigeria
[2]Electrical and Electronic Engineering, Rivers State University, Port Harcourt, Nigeria
[3]Electrical and Electronic Engineering, University of Lagos, Ikeja, Nigeria

Abstract

Secure deployment of neural networks (NNs) on edge and Internet of Things (IoT) devices is essential for real-time intelligence. However, it is challenged by latency variation, energy limits, adversarial attacks, privacy risks, memory scarcity, and hardware heterogeneity. This paper presents a secure NN deployment framework for edge/IoT using MATLAB-based simulation over 200 epochs and 200 inference requests. The framework evaluates latency, energy consumption, adversarial robustness, federated learning (FL) convergence under Differential Privacy (DP), compression-based memory reduction, side-channel power leakage, and throughput variability across 100 devices. Results show stable yet noisy latency of 80-88ms, centered near 83-84ms, and consistent energy demand with a mean of 1.07197×10^6mJ. At ε=0.1, unsecured model accuracy drops from 85.51% to 35.05%, while the defended model retains 75.50%, a 40.45 percentage-point gain. With DP noise of σ=0.03, FL global loss declines from 1.92 to about -0.05 by round 200, indicating convergence. Pruning and quantization from 20% to 50% reduce memory from 410.68 MB to 255.59 MB. Peak side-channel leakage reaches 1.62 V, and throughput varies from 4.24 to 16.50 million Ops/s. Despite simulation-only evaluation and limited adaptive-threat coverage, the study offers a practical metrics-driven blueprint for balancing security, privacy, and efficiency in edge/IoT AI deployment.

Keywords

Edge Computing, IoT Security, Neural Network Deployment, Side-channel Leakage, Differential Privacy, Adversarial Robustness, Federated Learning, Resource Optimization

1. Introduction

Edge and Internet of Things (IoT) devices are increasingly expected to execute neural-network-based intelligence close to the data source in order to support real-time decision-making in smart homes, industrial automation, healthcare monitoring, transportation, and other latency-sensitive applications. Compared with cloud-centric artificial intelligence (AI), edge intelligence reduces communication delay, improves service availability, lowers bandwidth demand, and can better preserve data locality. However, deploying neural networks directly on edge/IoT devices introduces a difficult systems problem because these platforms operate under severe constraints in computation, memory, energy, and communication, while

*Correspondence: Friday Oodee Philip-Kpae (philip-kpae.fo@ust.edu.ng)

Received: 9 October 2026; **Accepted:** 20 October 2026; **Published:** 31 October 2026

Copyright: © The Author(s), 2026. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

still being required to deliver accurate and timely inference. At the same time, the distributed and exposed nature of edge/IoT environments makes deployed models vulnerable to adversarial inputs, model poisoning, privacy leakage during collaborative learning, and device-level side-channel attacks. The major issues addressed in this paper is the secure and efficient deployment of neural networks on heterogeneous edge/IoT devices under joint resource, privacy, and threat constraints. This problem is challenging because relevant design objectives are tightly coupled rather than independent. Measures introduced to strengthen privacy may affect training stability and model utility; robustness mechanisms can increase latency and energy overhead; and aggressive compression can reduce memory footprint at the cost of predictive performance. In practical deployments, a secure edge intelligence pipeline must therefore balance multiple competing goals simultaneously: low inference latency, predictable energy use, compact memory footprint, resilience to adversarial and poisoning attacks, protection of distributed updates, and tolerance to hardware heterogeneity. Solving only one of these dimensions in isolation is insufficient for real-world deployment.

Recent studies have improved selected aspects of secure edge intelligence, including blockchain-supported trust management, privacy-preserving federated learning, intrusion detection, adversarial defense, and lightweight deployment strategies. Nevertheless, these efforts remain largely disjointed, with most works emphasizing one dimension of the problem while treating the others as secondary. As a result, there is still no unified framework that jointly evaluates efficiency metrics such as latency, energy, and memory together with privacy-utility behavior and robustness against multiple attack surfaces in heterogeneous, non-independent and identically distributed edge/IoT settings. This gap motivates the need for an integrated and metrics-driven approach that can better reflect the realities of practical deployment. To address this gap, we present a secure neural-network deployment framework for edge/IoT environments that combines privacy-preserving federated learning, adversarially resilient learning, and resource-aware model optimization within a single workflow. The proposed framework integrates differentially privacy federated learning through secure weighted aggregation with Gaussian noise injection, adversarial robustness through fast gradient sign method and projected gradient decent (FGSM/PGD) based training and perturbation-aware inference, and deployment efficiency through pruning and quantization for reduced memory and computational cost. The framework is evaluated in a MATLAB-based simulation environment designed to emulate heterogeneous edge/IoT conditions, multimodal workloads, and representative threat scenarios. Unlike prior studies that consider privacy, robustness, and deployment efficiency separately, the proposed approach examines their combined impact using metrics, including latency, energy consumption, memory footprint, convergence under privacy constraints, side-channel leakage indicators, and throughput variability across devices.

1.1. Key Contributions

The key contributions of this paper are threefold.

1) Proposing a unified secure deployment framework for neural networks on edge/IoT devices that jointly captures resource constraints and multi-surface threats, including latency and energy limits, adversarial inputs, poisoning, privacy leakage, and side-channel exposure.
2) Integrating privacy, robustness, and efficiency mechanisms in one workflow by combining differentially private federated learning with secure weighted aggregation, adversarially resilient training and inference using FGSM/PGD-based defenses, and resource-aware model compression through pruning and quantization.
3) Provision of a comprehensive metrics-driven evaluation in a MATLAB-based heterogeneous edge/IoT simulation environment, using latency, energy consumption, adversarial robustness, federated convergence, memory footprint, leakage traces, and throughput variability to demonstrate the practical trade-offs involved in secure edge AI deployment.

1.2. Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews recent related work and highlights the limitations of existing approaches in jointly addressing security, privacy, robustness, and deployment efficiency for edge/IoT intelligence. Section 3 presents the system model, including the device model, neural-network inference model, threat model, federated learning with differential privacy, and the quantized/pruned deployment model. Section 4 describes the methodology, simulation environment, synthetic dataset generation, and the algorithms for adversarial-resilient training, secure deployment, and secure inference with federated updates. Section 5 presents and discusses the results on latency, energy, adversarial robustness, federated-learning convergence, memory optimization, side-channel leakage, and throughput heterogeneity. Section 6 concludes the paper with the main findings, limitations, and future research directions.

2. Related Work

Secure edge/Internet-of-Things (IoT) intelligence has progressed, but recent solutions remain fragmented. A field-programmable gate array (FPGA) system-on-chip (SoC) adaptive, multimodal Internet of Medical Things (IoMT) design prioritizes real-time efficiency, yet provides limited end-to-end security evaluation beyond system implementation [1]. Accordingly, blockchain–deep learning approaches strengthen governance and privacy, but do not tightly quantify device-level constraints such as energy, memory, and latency jitter [2]. Likewise, Advanced Encryption Standard (AES)-secured federated foundation-model learning improves communication

security, but still leaves gaps in holistic robustness to adversarial inputs, privacy leakage, and physical side-channels [3]. For the meantime, blockchain-enabled cloud–edge federated learning improves collaboration integrity, but treats deployment efficiency and multi-surface attack resilience as secondary rather than as jointly optimized outcomes [4].

Recent Internet of Things (IoT) security research improves edge/IoT intrusion detection and federated learning (FL) protection, but the contributions are still scattered across different threat models and design goals. For intrusion detection, one approach frames distributed denial-of-service (DDoS) detection in consumer IoT healthcare as an adaptive decision process using off-policy reinforcement learning (RL) with a generative adversarial network (GAN), yet it remains focused on a single attack class rather than broader intrusion coverage [5]. Moving toward realistic edge settings, a blockchain-augmented FL intrusion detection system (IDS) handles non-independent and identically distributed (non-IID) edge-IoT data with robust aggregation, but it emphasizes secure collaboration more than unified deployment constraints and resilience to diverse attacks [6]. Similarly, distributed convolutional neural network (CNN) collaboration is defended against adversarial perturbations, but the emphasis is mainly on noise resistance rather than end-to-end IDS performance under heterogeneous edge limitations [7]. A quantum reservoir–based IDS introduces modeling novelty, but practical feasibility on edge devices and integration with privacy-preserving collaboration remain open issues [8]. On the FL security side, privacy and trust mechanisms are strengthened, but often evaluated in isolation. Homomorphic-encryption (HE) FL improves confidentiality with an efficiency focus [9]. Lightweight and trustworthy FL supports decision-making on constrained edge-IoT nodes, yet does not fully unify robustness and deployment constraints [10]. Poisoning prevention is enhanced via stateful proofs of execution (PoE) with differential privacy (DP), but it can still fall short of full system-wide efficiency validation under real edge conditions [11]. Private blockchain combined with FL supports secure sensing platforms, although overhead and scalability still require unified assessment [12].

Existing solutions are disjointed, addressing efficiency, blockchain trust, federated learning (FL) privacy by means of homomorphic encryption (HE) or differential privacy (DP), intrusion detection system (IDS)/distributed denial-of-service (DDoS) detection, or poisoning/adversarial defense in isolation. This leaves no end-to-end framework that cooperatively validates efficiency (latency/energy/memory), privacy–utility, and multi-surface robustness under heterogeneous, non-independent and identically distributed (non-IID) edge–Internet of Things (IoT) conditions. Our paper bridges this by unifying privacy-preserving FL, attack-resilient learning, and edge-efficient optimization in one workflow and evaluating them with consistent metrics for real edge-IoT deployment.

According to [13] the authors deployed the trained Temporal Convolutional Network (TCN) using the TensorFlow Lite (TFLite) framework to enable efficient execution on resource-constrained edge devices. They applied TensorFlow Lite's default post-training optimization techniques, including weight quantization and computational graph simplification, and enforced full-integer (INT8) quantization across all network layers. To ensure accurate quantization, a representative subset of approximately 300 samples from each dataset was used for activation-range calibration, while the complete test datasets were exported for reproducible inference on a Raspberry Pi 4 platform. The quantization process reduced the model size by approximately 75% by converting model parameters from FP32 to INT8 precision. Furthermore, the converted models employed INT8 data types for both input and output tensors, enabling inference through integer arithmetic and thereby reducing computational and memory requirements on the target hardware. Finally, the models were serialized into TensorFlow Lite's compact FlatBuffer format to facilitate efficient storage and deployment on edge devices.

Figure 1 presents the overall secure neural-network deployment framework for edge/IoT systems. It illustrates how differential privacy, through noise injection, and security mechanisms, through adversarial training, are incorporated into the model development pipeline to produce a robust and efficient neural network. The resulting model is then pruned and quantized for lightweight deployment on federated IoT devices and used for centralized inference with low latency and stable energy consumption. Overall, the figure highlights the interaction between privacy preservation, security enhancement, model optimization, local training, and efficient inference in a unified edge/IoT AI architecture.

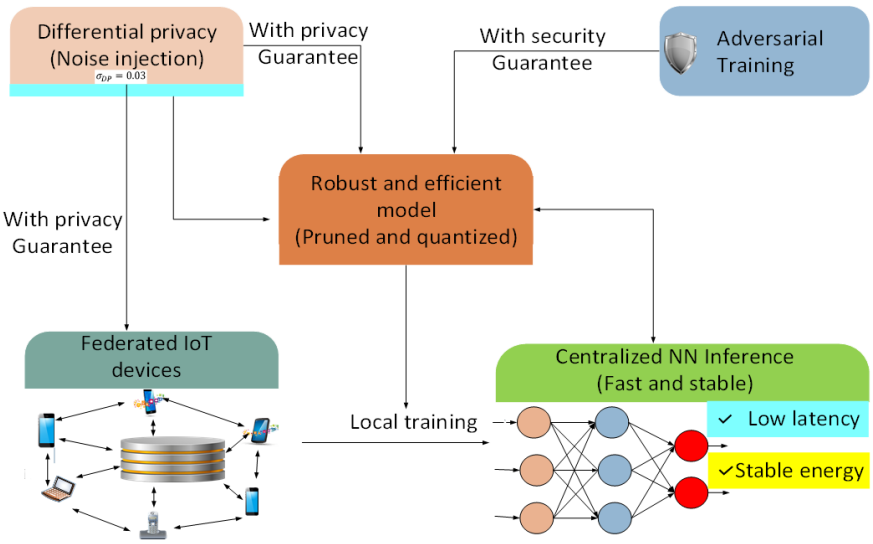


Figure 1. *Conceptual framework linking differential privacy and adversarial robustness with a pruned/quantized model, enabling efficient federated learning and fast, stable neural-network inference on IoT devices.*

Table 1 presents a summary of 12 related studies on secure edge and IoT intelligence. It includes the author and year, method used, strengths, limitations, and the way the current study addresses the identified gaps. The reviewed works span field-programmable gate array (FPGA) based edge intelligence, blockchain-supported federated learning, encrypted federated learning, adversarial defense, intrusion detection, and privacy-preserving IoT frameworks. Their major strengths include hardware efficiency, improved trust, privacy protection, robust aggregation, and effective attack detection. However, most of them are limited by narrow application scope and the lack of a unified treatment of privacy, robustness, and efficiency, as well as insufficient analysis of inference cost, memory reduction, side-channel leakage, and device heterogeneity. The current work addresses these limitations through a unified framework that jointly evaluates adversarial robustness, differential privacy in federated learning, pruning and quantization for memory optimization, inference latency, energy consumption, side-channel leakage, and throughput variation across heterogeneous edge/IoT devices.

Table 1. *Summary of recent studies on secure edge/IoT intelligence and how the present work addresses their limitations.*

Ref.	Method	Pros	Cons/Gap	How Current Work Addresses Gap
[1]	Adaptive Deep Neural Network (DNN) on Field-Programmable Gate Array (FPGA) System on Chip (SoC) for multimodal Internet of Medical Things (IoMT) edge inference	Hardware-aware, real-time, practical deployment	Focuses more on acceleration than unified security/privacy/robustness evaluation	Adds joint analysis of latency, energy, robustness, Differential Privacy Federated Learning (DP-FL), compression, and leakage
[2]	Blockchain-integrated deep learning with federated privacy-preserving edge intelligence	Strong trust, provenance, and privacy focus	Blockchain overhead; limited deployment-efficiency analysis	Evaluates security together with memory, energy, latency, and heterogeneity
[3]	AES-secured federated foundation model using Transformer-based Large Language Model (Transformer-LLM) and Long Short-Term Memory (LSTM)	Strong cryptographic protection; smart-grid relevance	Domain-specific; less focus on lightweight deployment	Provides broader edge/IoT NN deployment framework with efficiency-security trade-offs
[4]	Blockchain-supported secure FL in cloud-edge IoT	Secure collaboration and trust	Limited focus on inference cost and model compression	Adds secure inference, pruning, and quantization analysis
[5]	Proximal Policy Optimization	Strong attack-detection capability	IDS-specific, not full NN deployment pipeline	Addresses broader secure deployment beyond only detection

Ref.	Method	Pros	Cons/Gap	How Current Work Addresses Gap
	Plus Generative Adversarial Network (PPO + GAN) for Distributed Denial of Service (DDoS) detection			
[6]	Blockchain-augmented Federated Learning-based Intrusion Detection System (FL IDS) with robust aggregation	Handles Independent and Identically Distributed (non-IID) data; secure aggregation	Narrow Intrusion Detection System (IDS) scope	Extends to unified privacy, robustness, compression, and inference evaluation
[7]	Defense against adversarial noise in distributed Convolutional Neural Network (CNN) collaboration	Strong adversarial robustness	Limited privacy and deployment-efficiency analysis	Combines adversarial defense with DP-FL and resource-aware optimization
[8]	Quantum reservoir computing for intrusion detection	Innovative security method	Focused on IDS, not end-to-end deployment	Focuses on practical NN deployment trade-offs across full pipeline
[9]	FL with single-key homomorphic encryption	Strong privacy-preserving aggregation	Limited adversarial robustness and compression study	Integrates privacy with robustness, memory reduction, and deployment metrics
[10]	Lightweight trustworthy FL for constrained edge-IoT	Resource-aware and practical	Mainly FL-centric	Adds secure inference, side-channel leakage, and device heterogeneity analysis
[11]	Poisoning prevention using stateful proofs of execution	Strong poisoning defense and integrity assurance	Less emphasis on end-to-end deployment efficiency	Adds latency, energy, compression, and adversarial evaluation in one framework
[12]	Private blockchain with FL for secure urban noise monitoring	Good application-level validation	Domain-specific platform	Generalizes to broader secure NN deployment for heterogeneous edge/IoT systems

3. System Model

System model for secure neural network deployment in edge and Internet of Things (IoT) environments is presented in this section. It defines the network architecture, device roles, and communication links for inference and model updates, alongside key constraints such as computation, memory, energy, and latency as presented in Figure 2. It illustrates the secure edge/IoT neural-network deployment environment and the interaction among deployment constraints, learning mechanisms, and security risks. It shows that edge/IoT devices operate with limited computational power, memory, and energy budgets, requiring pruned and quantized models for efficient execution. Within this environment, federated learning with differential privacy supports collaborative model updates through noise injection and global aggregation, while the deployed neural network performs inference under latency and energy constraints. The figure also highlights the exposure of the system to major security threats, including adversarial attacks, model poisoning, and side-channel leakage, which collectively contribute to system vulnerability. The figure summarizes the practical relationship between resource-constrained deployment, privacy-preserving learning, inference efficiency, and threat exposure in edge/IoT AI systems.

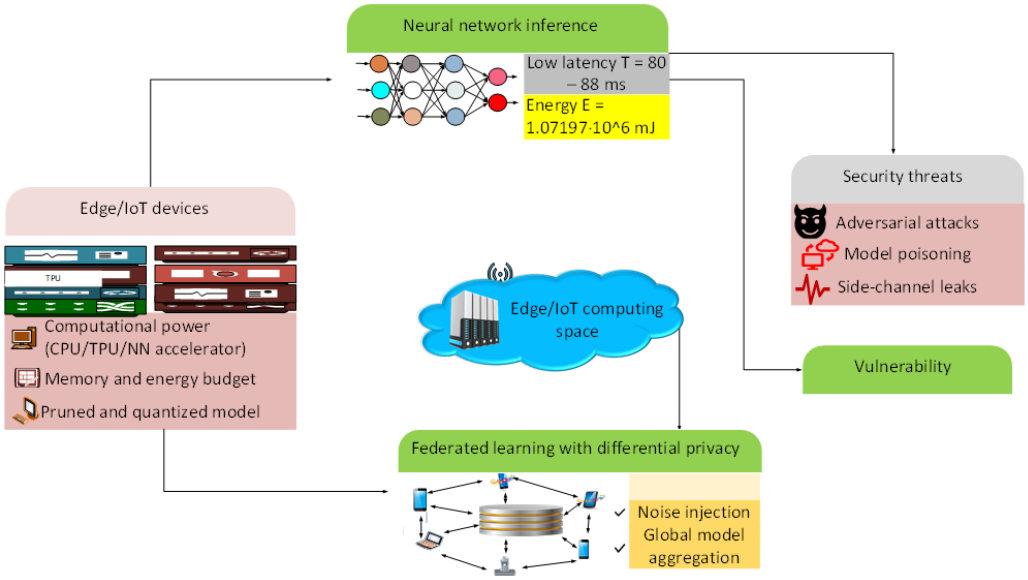


Figure 2. *System model showing edge/IoT NN inference* ($latency$ $80-88$ ms; $energy$ 1.07197×10^{6} mJ), *key security threats, and federated learning with DP* (σ_{DP} $=$ 0.03) *for privacy-preserving aggregation.*

1) Edge/IoT Device Model

Consider N IoT devices, each with computational capacity C_i, memory M_i, and energy budget E_i. Each device runs a local NN model $\mathbb{M}_i$, to process input x_i, for inference y_i. Other devices include Tensor Processing Unit/Neural Network (TPU/NN) accelerators or GPU cores to enhance performance. The device model is extended to account for: (i) CPU frequency f_i (GHz), (ii) Accelerator operations per second A_i, (iii) Quantized or pruned NN model $\mathbb{M}_i^q$ for memory and energy efficiency, and (iv) Communication interface with bandwidth BW_i and latency $t_{comm,i}$.

2) Neural Network Inference Model

The inference latency L_i is modeled as presented in (1):

$$L_i = \frac{\sum_{l=1}^{L} C_l P_l}{f_i + \alpha A_i} + t_{comm,i} \tag{1}$$

where L is the layers in the network, C_l is the computation cycles per layer, P_l is the neurons/operations per layer, f_i is the CPU frequency, A_i is the accelerator operations per second, α is the accelerator efficiency factor, and $t_{comm,i}$ is the network communication delay.

Whereas the Energy consumption E_i is expressed in (2):

$$E_i = k_i \sum_{l=1}^{L} C_l P_l (f_i + \beta A_i)^2 + E_{comm,i} \tag{2}$$

where β is energy efficiency factor for accelerator use, and $E_{comm,i}$ is the energy for data transmission.

3) Security Threat Model

We include up-to-date threats in edge/IoT NN deployment with the following assumptions as follows: (i) Adversarial examples as: $x_i^{adv} = x_i + \delta$, which may cause misclassification. (ii) Backdoor/model poisoning attacks that may occur during federated updates, (iii) Model inversion/data leakage from untrusted aggregation, (iv) Side-channel attacks exploiting power/latency information.

4) Federated Learning with Differential Privacy

We modeled the federated learning with differential privacy by considering the local model update at device i as presented in (3):

$$\theta_i^{t+1} = \theta_i^t - \alpha \nabla \mathcal{L}(\theta_i^t, x_i) \tag{3}$$

while the global aggregation with secure weighted averaging and DP noise is modeled using the expression in (4):

$$\theta^{t+1} = \sum_{i=1}^{N} w_i \left(\theta_i^{t+1} + \mathcal{N}(0, \sigma^2) \right) \tag{4}$$

where $\theta_i^{t+1} = \theta_i^t - \alpha \nabla \mathcal{L}(\theta_i^t, x_i)$ as expressed in (3), $\mathcal{N}(0, \sigma^2)$ is the Gaussian noise added to the preserve privacy.

Quantized/Pruned NN Model for Edge Devices

For the quantized/pruned NN model for edge devices, model $\mathbb{M}_i$ is compressed to $\mathbb{M}_i^q$ to reduce memory foot-print $\mathbb{M}_i$ and computation as given in (5):

$$Memory\ (\mathbb{M}_i^q) \leq \mathbb{M}_i, \text{ and } Operation\ (\mathbb{M}_i^q) \leq C_i \tag{5}$$

This allows deployment on resource-constrained IoT devices without significant accuracy loss.

4. Methodology

This paper follows a model-driven, simulation-based approach. It first develops a system and mathematical model for

secure neural network deployment on edge/Internet of Things (IoT) devices. It then designs security-aware methods for training, inference, and federated updates, combining adversarial resilience, privacy-preserving aggregation, and resource-aware optimization (e.g., pruning and quantization). The study uses a MATLAB-based simulation environment. The synthetic dataset used for the simulation was generated using the following dataset described as follows: (i) Device heterogeneity is CPU $C_i = 0.5 - 2\ \ GHz$, memory $\mathbb{M}_i = 128 - 1024\ \ MB$, energy $E_i = 1000 - 5000\ \ mAh$, (ii) Data employed includes IoT sensor signals, images, time-series data, (iii) The Adversarial perturbations is FGSM/PGD attacks, (iv) Epochs range is between $50 - 200$ for training and federated aggregation. Communication delays $t_{comm,i} = 1 - 50\ \ ms$. Model compression ratios: $20 - 50\%$ for quantization/pruning. We generate a synthetic IoT dataset in a MATLAB-based simulation environment by sampling heterogeneous device constraints (central processing unit (CPU), memory, and energy) and producing multimodal inputs (sensor, time-series, and image-like data) with task-defined labels. We also create adversarial variants using FGSM and PGD to evaluate robustness under gradient-based attacks. This setup is representative because it jointly captures device heterogeneity, multimodal IoT workloads, and adversarial conditions in a controlled and repeatable manner. The simulation emulates edge inference and federated-learning cycles under heterogeneous device constraints and network conditions. Performance is assessed using latency, energy consumption, memory footprint reduction, convergence behavior under privacy constraints, and robustness to adversarial/poisoning effects, alongside indicators of potential leakage where applicable. Algorithms 1-4 define the operational framework of the proposed secure edge/IoT neural-network deployment system. Specifically, Algorithm 1 details the adversarially robust training strategy, Algorithm 2 formalizes the secure and compressed model deployment stage, Algorithm 3 describes the protected inference procedure on resource-constrained edge devices, and Algorithm 4 presents the privacy-aware federated learning update mechanism. Taken together, these algorithms provide an integrated end-to-end description of the proposed method and clarify how the framework simultaneously addresses robustness, privacy preservation, model optimization, and efficient distributed intelligence in heterogeneous edge/IoT environments. Key symbols, parameters, and notations used in the formulation of the proposed secure neural-network deployment framework is presented in Table 2. It provides the analytical foundation for the system model by defining the principal variables associated with edge/IoT devices, neural-network inference, adversarial robustness, federated learning with differential privacy, and model compression.

Algorithm 1: Adversarial-Resilient Training: it presents the procedure for generating adversarial samples during training and updating model parameters to improve robustness against adversarial perturbations under edge/IoT constraints.

Algorithm 1: Secure NN Training with Adversarial Data Augmentation

a) *for* epoch $e = 1$ to E *do*
b) *for* each device i *do*
c) Generate adversarial examples x_i^{adv} from the security threat model using FGSM/PGD
d) Update local model θ_i using combined loss: $\mathcal{L}_{total} = \mathcal{L}(\theta_i, x_i) + \lambda\mathcal{L}(\theta_i, x_i^{adv})$
e) *end for*
f) *end for*
g) Securely aggregate θ_i at central server (federated averaging)

Algorithm 2: Secure Model Deployment Algorithm: it represents the procedure for preparing and deploying a resource-efficient, security-aware model on edge/IoT devices, including pruning/quantization, integrity checks, and secure update/aggregation steps for reliable inference and distributed learning.

Algorithm 2: Resource-Constrained Secure Deployment with Model Compression

a) *for* each device i *do*
b) Measure available resources $(C_i, \mathbb{M}_i, E_i)$
c) *if* full model $\mathbb{M}_i$ fits resources, *then* deploy $\mathbb{M}_i$
d) *else*
e) Apply pruning/quantization: $\mathbb{M}_i \to \mathbb{M}_i^q$
f) Deploy $\mathbb{M}_i^q$
g) *end if*
h) *end for*

Algorithm 3: Edge/IoT Secure Inference and Federated Update: it shows a workflow for performing secure on-device inference, detecting anomalous/adversarial inputs, and periodically executing privacy-preserving federated updates with robust aggregation under heterogeneous edge constraints.

Algorithm 3: Edge/IoT Secure Inference and Federated Update

a) *for* each inference request *do*
b) Device i receives input x_i
c) Detect adversarial perturbation using threshold τ
d) *if* adversarial perturbation is detected, *then*
e) Reject input or apply defense (input preprocessing, randomized smoothing)
f) *end if*
g) Perform inference $y_i = \mathbb{M}_i^q(x_i)$
h) Periodically update θ_i using averaging with DP and secure aggregation
i) *end for*

Simulation Parameters

Table 2. *Simulation Parameters for Secure Neural Network Deployment on Edge/Internet of Things (IoT) Devices. It shows the summary of device, network, learning, security, and workload settings used to evaluate latency, energy, resource usage, and robustness under heterogeneous edge conditions.*

Parameters	Values/Range
Devices N	$50-500$
CPU frequency f_i	$0.5-2\ GHz$
Memory $\mathbb{M}_i$	$128-1024\ MB$
Energy E_i	$1000-5000\ mAh$
Inference epochs	$50-200$
Pruning/quantization ratio	$20-50\%$
Adversarial perturbation ϵ	$0.01-0.1$
Communication delay $t_{comm,i}$	$1-50\ ms$
Federated DP noise σ	$0.01-0.05$

5. Results and Discussions

This section presents the simulation outcomes that validate the proposed secure edge/IoT neural network deployment workflow. It analyzes how adversarial-resilient training, privacy-preserving federated updates, and resource-aware optimization (pruning/quantization) jointly affect key performance metrics, including inference latency, energy consumption, memory footprint, convergence behavior, and robustness under adversarial/poisoning conditions. The section further discusses trade-offs between security, privacy, and efficiency across heterogeneous devices and network conditions, highlighting practical implications for real-time, resource-constrained edge-IoT deployments.

5.1. Inference Latency Performance

The secure neural network deployment simulation produces clear quantitative patterns across the evaluated metrics. Figure 3 shows a time-series plot of $L_i\ (ms)$ against time/epochs $(0-200)$. The curve is highly volatile, with persistent short-term fluctuations and noise, but no sustained upward or downward trend. Most values remain within a tight range of about $80-88\ ms$, with a typical level around $83-84\ ms$. Brief peaks approach $89\ ms$ around epochs $50\ and\ 65$, while the lowest points drop slightly below $80\ ms$. This shows that the behavior is consistent with latency variability (jitter) in a networked or edge-processing setting, where per-cycle delays vary but remain centered around a stable mean.

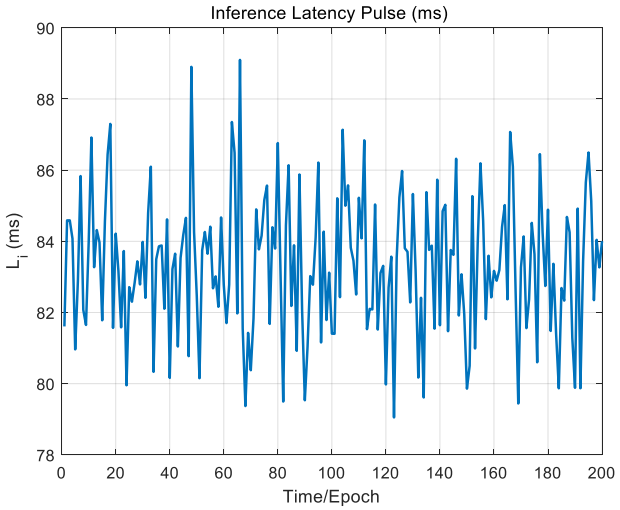


Figure 3. *Inference latency $L_i\ (ms)$ time-series across 200 epochs, showing stable latency with short-term fluctuations (jitter) and no sustained trend, consistent with realistic edge/IoT processing variability.*

5.2. Energy Consumption Profile

Figure 4 presents the energy consumption trace $E_i\ (mJ)$ per inference request, reflecting the power overhead of the deployed model across 200 requests. The curve shows rapid, high-frequency fluctuations and stochastic noise, indicating small per-cycle variations in energy use. Despite this jagged pattern, there is no sustained upward or downward drift over time, suggesting a stable computational workload throughout the simulation. Quantitatively, the signal is centered around a mean of approximately $1.07197 \times 10^6\ mJ$, with most values confined to a narrow band of about $1.07196 \times$

10^6 mJ to 1.07198×10^6 mJ. The brief peaks and dips likely reflect transient switching activity and minor execution-path differences during inference. Overall, the results indicate relatively constant energy demand, supporting the suitability of the deployment for edge/IoT scenarios where predictable power usage is critical, especially for battery-powered devices.

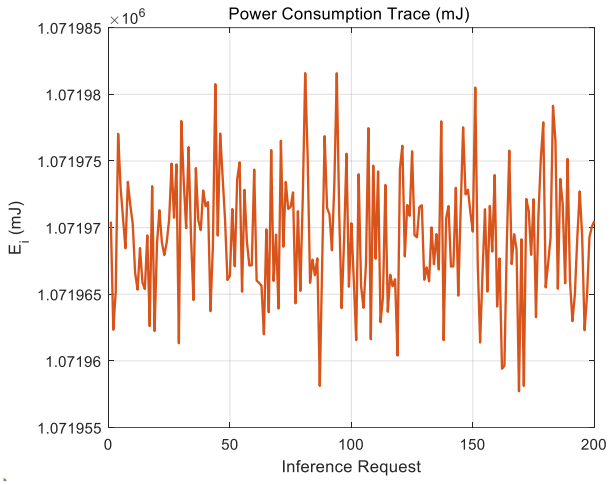


Figure 4. *Power consumption trace E_i (mJ) across 200 inference requests, showing high-frequency fluctuations around a stable mean, indicating consistent energy demand per inference cycle in the edge/IoT deployment.*

5.3. Adversarial Robustness Evaluation

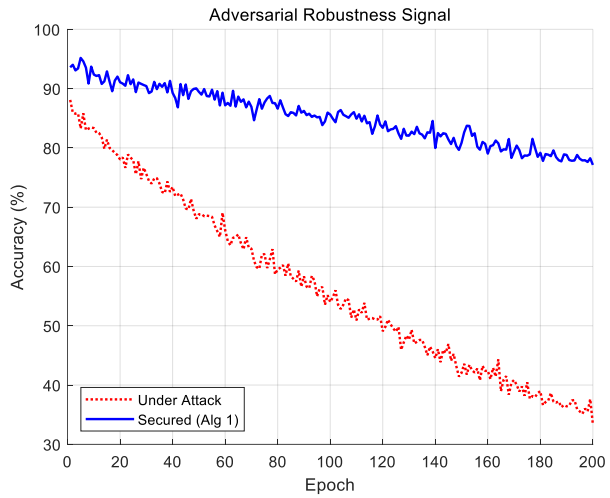


Figure 5. *Adversarial robustness signal comparing model accuracy under increasing perturbation strength (ϵ) for the unsecured and secured configurations, demonstrating improved resilience of the defended model as attack intensity increases.*

Figure 5 evaluates adversarial robustness by comparing model accuracy under increasing attack strength (ϵ) for an unsecured baseline and a secured (defended) version. Without protection, accuracy falls sharply from 85.51% to 35.05% at $\epsilon = 0.1$, indicating severe susceptibility to adversarial perturbations. With the defense enabled, accuracy degrades far more slowly, remaining at 75.50% at the same ϵ level. This represents a 40.45% $point$ robustness gain at $\epsilon = 0.1$, demonstrating that the proposed defense substantially improves resilience and reinforcing the need to embed robustness mechanisms during deployment in adversarial edge/IoT environments.

5.4. Federated Learning Convergence with Differential Privacy

Figure 6 shows the convergence of the global aggregation loss θ $over$ 200 federated learning rounds under Differential Privacy (DP) noise. The loss starts at about 1.92 and decays in an approximately exponential manner as the global model learns from distributed updates. Although DP noise is injected with standard deviation $\sigma_{DP} = 0.03$ to protect privacy, the loss still converges toward zero and settles near -0.05 by $round$ 200, reflecting small noise-induced fluctuations rather than true divergence. This trend suggest that privacy-preserving training can remain stable and convergent when the DP noise magnitude is appropriately balanced with the learning rate.

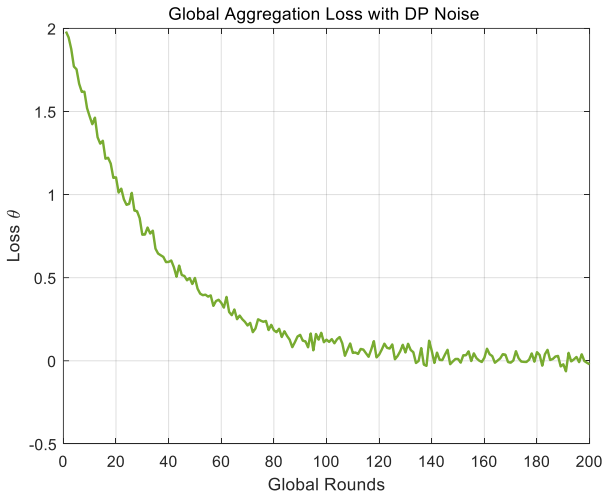


Figure 6. *Global aggregation loss (θ) over federated learning rounds under Differential Privacy (DP) noise $\sigma_{DP} = 0.03$, showing stable convergence despite privacy-preserving perturbations.*

5.5. Model Compression and Memory Optimization

Figure 7 illustrates how model compression through pruning and quantization reduces memory usage. As the pruning level increases from 20% to 50%, the quantized memory footprint $\mathbb{M}_i^q$ decreases in a near-linear manner. Memory

drops from about 410.68 MB to 255.59 MB at the highest pruning level, yielding a reduction of 155.09 MB. This highlights a substantial storage optimization for resource-constrained IoT devices. The area plot further emphasizes that more aggressive pruning can markedly lower memory requirements, although these gains should be balanced against possible accuracy degradation, reflecting the efficiency–performance trade-off in model optimization.

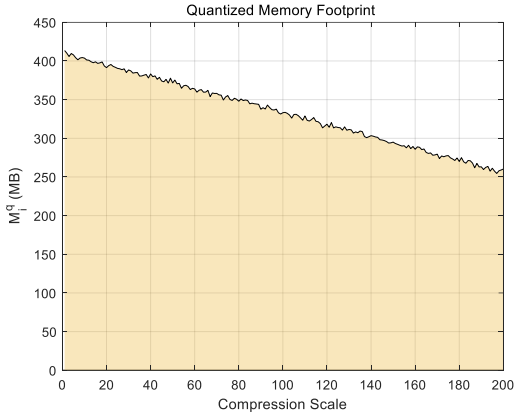


Figure 7. *Quantized memory footprint $\mathbb{M}_i^q$ versus pruning/compression level, illustrating the reduction in model memory usage as pruning intensity increases for resource-constrained edge/IoT devices.*

5.6. Side-Channel Leakage Power Trace

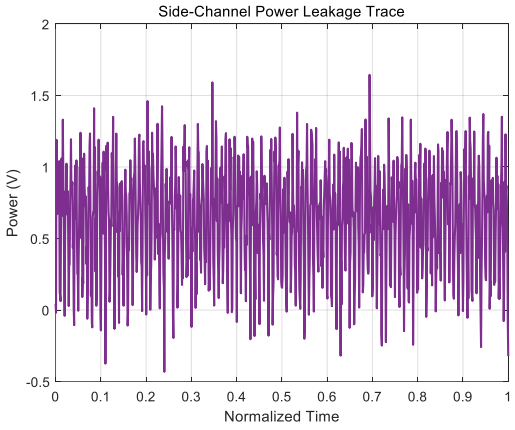


Figure 8. *Side-channel power leakage trace showing voltage variations over normalized time, highlighting periodic power signatures and noise components that may reveal data-dependent leakage in edge/IoT hardware.*

Figure 8 shows the side-channel power leakage trace used to assess physical security on the device. The plot captures voltage variations over normalized time, forming a periodic waveform with a peak of 1.62 V and a mean of 0.64 V. The oscillations, together with the observable noise floor, represent power or electromagnetic signatures that an attacker could correlate with internal computations to infer sensitive information, including cryptographic keys. These measured peaks provide a basis for designing countermeasures—such as masking and hiding—to reduce data-dependent leakage by weakening the link between processed values and observable power consumption.

5.7. Throughput Heterogeneity Across IoT Devices

Figure 9 illustrates throughput heterogeneity across $N = 100$ $devices$, emphasizing performance variation in a diverse IoT ecosystem. Throughput spans from 4.24 $million$ $operations$ per $second$ (Ops/s) to 16.50 $million$ Ops/s, reflecting underlying differences in device capability. This spread aligns with the assumed CPU frequency range of $0.5 - 2.0$ GHz across the sampled devices, where higher-frequency nodes deliver substantially higher throughput. The stem plot highlights a long-tailed distribution, with top-performing edge nodes processing nearly four times faster than the most constrained devices. Accounting for this heterogeneity is essential for designing scheduling and load-balancing strategies that maintain consistent service quality in distributed secure model deployments.

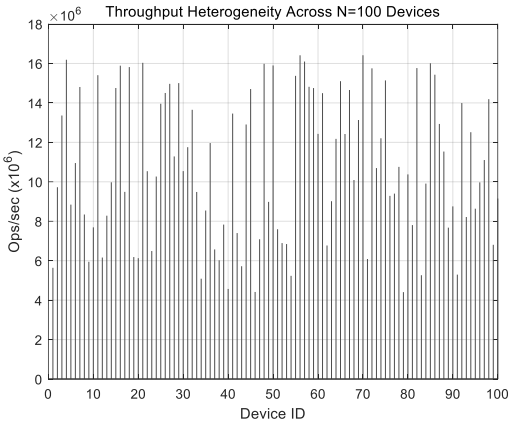


Figure 9. *Throughput heterogeneity across N edge/IoT devices, showing the distribution of operations-per-second performance under varying device capabilities and highlighting variance that impacts scheduling and load balancing.*

5.8. Summary of Results

Summary of key results from this work is presented in Table 3. It summarizes the performance gains achieved by the proposed secure neural-network deployment framework relative to the baseline model across accuracy, federated-learning convergence, and memory efficiency. First, under adversarial perturbation at $\varepsilon = 0.1$, the baseline model records an accuracy of 35.05%, whereas the secured/proposed model maintains 75.50%, corresponding to a gain of 40.45 percentage

points. This result indicates that the proposed defense mechanism substantially improves model robustness against adversarial attacks and preserves prediction reliability in hostile edge/IoT environments. Second, both the baseline and secured models achieve the same initial accuracy of 85.51%, yielding no gain at initialization. This shows that the security enhancements do not degrade the model's original clean-data performance before attack conditions are introduced. Third, for federated learning with differential privacy at $\sigma_{DP} = 0.03$, the global loss decreases from 1.92 at the start of training to approximately -0.05 by round 200, representing a total reduction of 1.97. This demonstrates that the proposed privacy-preserving learning mechanism remains stable and convergent despite the addition of DP noise. Finally, model compression through pruning from $20\%\ to\ 50\%$ reduces memory usage from $410.68\ MB\ to\ 255.59\ MB$, producing a memory saving of 155.09 MB. This confirms that the proposed framework improves storage efficiency and makes deployment more practical for resource-constrained edge and IoT devices.

Table 3. *Summary of key simulation results comparing unsecured and proposed secured deployment, suggesting possible gains in adversarial robustness, privacy-preserving federated convergence, and memory efficiency.*

Metric	Baseline	Secured/Proposed	Gain
Accuracy at $\varepsilon = 0.1$	35.05%	75.50%	$+40.45\ pp$
Initial accuracy	85.51%	85.51%	$0\ pp$
FL loss $\sigma_{DP} = 0.03$	1.92 $(start)$	-0.05 $(round\ 200)$	$reduced\ by\ 1.97$
Memory $(20\% \rightarrow 50\%\ pruning)$	$410.68\ MB$	$255.59\ MB$	$-155.09\ MB$

6. Conclusion

This study presented a unified framework for the secure deployment of neural networks in edge and Internet of Things (IoT) environments by jointly addressing adversarial robustness, privacy preservation, and deployment efficiency under practical device constraints. The results showed that the proposed approach can maintain stable inference latency and energy behavior while significantly improving robustness against adversarial perturbations, supporting convergence in federated learning under differential privacy, and reducing model memory footprint through pruning and quantization. These findings confirm that secure edge/IoT intelligence should not be treated as a single-dimensional optimization problem, but rather as a careful balance among security, privacy, accuracy, and resource efficiency. Overall, the work contributes a metrics-driven and practically oriented framework that can guide the design of resilient and lightweight AI systems for heterogeneous edge environments.

Despite these contributions, the study has some limitations. The evaluation was conducted in a MATLAB-based simulation environment rather than on physical edge/IoT hardware, which means that real-world deployment effects such as hardware-specific timing behavior, thermal constraints, communication instability, and implementation overhead may not be fully captured. In addition, although the framework considers important threats such as adversarial attacks, privacy leakage, and side-channel exposure, it does not exhaustively evaluate fully adaptive attackers or all possible real-world attack combinations. Future work should therefore focus on hardware-based validation on diverse edge platforms, broader threat modeling with stronger adaptive adversaries, and deeper investigation of mitigation strategies for side-channel leakage and heterogeneous device behavior. Further research may also explore more advanced privacy mechanisms, lightweight defense techniques, and adaptive compression strategies to improve the scalability and practicality of secure edge/IoT AI deployment.

Abbreviations

AI	Artificial Intelligence
PGD	Projected Gradient Decent
FL	Federated Learning
DP	Differential Privacy
NN	Neural Networks
IoMT	Internet of Medical Things
IoT	Internet-of-Things
FPG	Field-Programmable Gate Array

Author Contributions

Daniel Ekpah: Conceptualization
Friday Oodee Philip-Kpae: Conceptualization
Agbotiname Lucky Imoize: Writing – review & editing

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] N. B. Gaikwad, S. K. Khare, D. Mendhe, H. Mir, S. Kosta and U. R. Acharya, "FPGA SoC Implementation of Adaptive Deep Neural Network-Based Multimodal Edge Intelligence for Internet of Medical Things," in *IEEE Access*, vol. 13, pp. 134041-134056, 2025, https://doi.org/10.1109/ACCESS.2025.3592729

[2] K. Swathi, P. Durga, K. V. Prasad, A. K. Chaitanya, K. Santhi, P. Vidyullatha and S. V. A. Rao "Secure blockchain integrated deep learning framework for federated risk-adaptive and privacy-preserving IoT edge intelligence sets" *Scientific Reports* 15, 41133 (2025). https://doi.org/10.1038/s41598-025-24895-8

[3] M. K. Hasan, S. R. Kabir, S. Islam, S. Abdullah, H. S. Abbas, B. K. Pandey, "AES Cryptography Enabled Responsible Federated Foundation Model Using Transformer LLM and LSTM for Smart Grid IIoT Networks," in *IEEE Internet of Things Journal*, vol. 12, no. 23, pp. 49801-49810, 1 Dec.1, 2025, https://doi.org/10.1109/JIOT.2025.3608807

[4] M. Kumar, J. K. Samriya, G. KaurWalia, P. Verma, H. Wu and S. S. Gill, "Blockchain Empowered Secure Federated Learning for Consumer IoT Applications in Cloud-Edge Collaborative Environment," in *IEEE Transactions on Consumer Electronics*, vol. 71, no. 2, pp. 3986-3996, May 2025, https://doi.org/10.1109/TCE.2025.3532676

[5] J. Yang, V. Govindarajan, L. Y. Por, Z. A. Shaikh, Q. Xin, P. Bhattacharya, "DDoS Attack Detection in Consumer IoT-Based Healthcare Systems Using Improved Off-Policy Proximal Policy Optimization and Generative Adversarial Network," in *IEEE Transactions on Consumer Electronics*, vol. 71, no. 4, pp. 11662-11675, Nov. 2025, https://doi.org/10.1109/TCE.2025.3605098

[6] K. J. Mukisa, L. A. C. Ahakonye, D. -S. Kim and J. -M. Lee, "Blockchain-Augmented FL IDS for Non-IID Edge-IoT Data Using Adaptive Trimmed Mean Aggregation," in *IEEE Internet of Things Journal*, vol. 12, no. 21, pp. 45150-45159, 1 Nov.1, 2025, https://doi.org/10.1109/JIOT.2025.3598832

[7] M. Asif, M. K. Kazmi, M. A. Rahman, S. R. Hasan and S. Homsi, "SHEATH: Defending Horizontal Collaboration for Distributed CNNs Against Adversarial Noise," in *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 4353-4368, 2025, https://doi.org/10.1109/TIFS.2025.3556349

[8] T. Ahmad, S. Khan, N. Nigar, E. Mohamed Obaid and S. Naaz, "Quantum Reservoir-Based Intrusion Detection for Edge-IoT Systems," in *IEEE Transactions on Consumer Electronics*, vol. 71, no. 4, pp. 11269-11278, Nov. 2025, https://doi.org/10.1109/TCE.2025.3618827

[9] Y. Liu, H. Zhang, M. Wang, Q. Xie and L. Wang, "An Efficient Privacy-Enhanced Federated Learning with Single-Key Homomorphic Encryption," in *IEEE Internet of Things Journal*, vol. 12, no. 13, pp. 23598-23613, 1 July1, 2025, https://doi.org/10.1109/JIOT.2025.3553134

[10] A. Tariq, F. Sallabi, M. A. Serhani, E. S. Barka, T. Qayyum, I. Din, "AI-Driven Lightweight and Trustworthy Federated Learning Framework for Decision-Making in Resource-Constrained Edge-IoT," in *IEEE Transactions on Consumer Electronics*, vol. 71, no. 4, pp. 12166-12180, Nov. 2025, https://doi.org/10.1109/TCE.2025.3628207

[11] N. Rattanavipanon and I. De Oliveira Nunes, "SLAPP: Poisoning Prevention in Federated Learning and Differential Privacy via Stateful Proofs of Execution," in *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 4167-4182, 2025, https://doi.org/10.1109/TIFS.2025.3555179

[12] M. J. Baucas and P. Spachos, "Private Blockchain and Federated Learning-Based Edge-Iot Platform for Secure Urban Noise Monitoring," *ICC 2025 - IEEE International Conference on Communications*, Montreal, QC, Canada, 2025, pp. 1145-1150, https://doi.org/10.1109/ICC52391.2025.11160895

[13] E. Manor and S. Greenberg, "Custom hardware inference accelerator for tensorflow lite for microcontrollers," IEEE Access, vol. 10, pp. 73484–73493, 2022.