SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I11P102

Volume 12 Issue 11, 8-20, November 2025

Original Article

IDMO: A Multi-Stage Optimized Deep Learning Framework for Efficient and Scalable IoT Big Data Analytics

Ch.Ellaji[1], R.S. Ponmagal[2], V. Saritha[3]

[1]*Department Of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India.*
[2]*Department Of Computing Technologies, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India.*
[3]*School of Engineering and Technology, Sri Padmavati Mahila Visvavidyalayam, Tirupati, Andra Pradesh.*

[1]*Corresponding Author : Ce6149@srmist.edu.in*

Received: 02 September 2025 Revised: 04 October 2025 Accepted: 03 November 2025 Published: 29 November 2025

Abstract *- The fast growth of Internet of Things (IoT) ecosystems created massive data volumes, which show both diverse characteristics and sporadic connectivity while maintaining strict privacy rules. The current technological environment requires instantaneous analytics processing at the network boundary using conventional deep learning systems alongside federated learning solutions, which usually encounter processing limitations, model complexity limitations, and intense data transfer downtime. The research presents IDMO as a novel three-stage framework that optimizes deep neural networks in IoT-FL systems through model compression integration with communication efficiency techniques. The IDMO pipeline comprises three main components: (i) a Bi-Level Utility (BLU)-guided structured pruning method that adaptively eliminates redundant filters while maintaining crucial feature representations; (ii) a Selective-Importance-driven Joint Fine-tuned Optimisation (SI-JFO) quantisation approach that employs metaheuristically guided, non-uniform encoding according to weight significance and gradient sensitivity; and (iii) a Niblack-Adaptive Thresholding (NA-T)- based selective update mechanism that reduces communication costs by exclusively transmitting significant local parameter changes in federated environments. Tests conducted on CIFAR-10 data sets confirm IDMO achieves 91.43% accuracy and reduces model size to 5.20 MB from 1.49 MB with a 71.3% decrease, while FLOPs drop to 53.5% compared to standard FL protocols. This leads to 65% lower communication expenses. Merging these improvements does not impact inference performance, yet makes IDMO operational in edge settings with limited resources. The research outcomes demonstrate that IDMO technology can transform IoT analytics capabilities because it delivers efficient edge-processing of adaptive deep learning models that protect user privacy.*

Keywords *- IoT, Deep Learning, Pruning, Quantisation, Federated Learning, Maxout Networks, Edge Computing.*

1. Introduction

Internet of Things (IoT) to a large extent, asymmetrical data streams require low luxurious and privacy protection analysis. Sky-centered approaches cause delays, bandwidth overload, and energy collection, while edge devices face limited memory and calculation capacity, leading to traditional deep learning being inappropriate. Federated Learning (FL) addresses data privacy, but the entire model is exposed to excessive communication costs when transferring updates.

Current pruning and quantization methods reduce complexity, but often reduce Accuracy. This research addresses the difference by suggesting a better deep maxout optimized (IDMO) framework, which integrates propagation, adaptive quantity, and selective updates in the integrated pipeline for practical, accurate IoT analysis.

1.1. Background and Motivation

Edge computing has become necessary because of IoT data processing speed requirements, which allows analysis at the data source to minimise latency and reduce cloud infrastructure requirements. Edge devices typically have small memory capacity and limited processing power in addition to their restricted source of energy (Wang et al., 2024). The considerable computational demands of deep learning models stem from their large parameter quantities and vast network depths, together with their demanding processing needs. Placing uncompressed or regularly trained models on small devices causes both technical and operational obstacles.

FL offers an effective way for clients to train models together while keeping their private data local. Data privacy stays intact because only the results from model changes get passed to a central server, while diminishing the amount of information that needs to travel.

Standard FL methods like FedAvg still transmit a complete model each round to multiple clients, which produces high communication costs, particularly when deep networks are employed. The methods disregard both the design weaknesses of the models and the deployment requirements related to hardware platforms.

The current issues drive researchers to build an end-to-end framework that combines model reduction with adaptable communication schemes, along with clever activation techniques when used in federated learning systems (Injadat et al., 2021). The objective focuses on permitting deep learning models to operate proficiently on edge devices by reducing resource usage with no compromise to their predictive abilities.

1.2. Problem Statement

The present deep learning frameworks struggle to satisfy IoT environment requirements because they have three fundamental performance limitations. The current version Of Convolutional Neural Networks (CNNs) operates with excessive parameters, which makes them unfit for deployment on edge-based platforms. The use of post-training quantisation technology leads to a notable accuracy decline when employed in a non-discriminatory way across all network layers (Pechetti & Rao, 2023). Debugging federated learning becomes inefficient because traditional methods send complete parameter sets for updates regardless of weight change magnitudes, thus creating excessive bandwidth problems.

The necessity arises for a single framework that combines lightweight operation with efficient communication alongside accuracy maintenance to support IoT projects that face memory, latency, and energy limitations.

1.3. Overview of the Proposed Approach

This paper presents the Improved Deep Maxout Optimised framework, which serves as a resource-optimised end-to-end deep learning architecture made specifically for federated IoT systems (Maity et al., 2024). IDMO employs three unified elements that optimise structure design, parameter handling, and communication processes simultaneously:

1.3.1. Structured Pruning with Backward-Linked Utility (BLU)

A new structured pruning method uses the BLU metric to analyse backward-contribution gradients, which determine the relevance of feature maps and filters in their evaluation process. The technique achieves redundant computation elimination while maintaining original functional abilities.

1.3.2. Selective-Importance-driven Joint Fine-tuned Optimisation (SI-JFO)

This method utilises hybrid quantisation to determine weight bit widths by measuring weight sensitivity before performing joint fine-tuning for accuracy maintenance.

1.3.3. Niblack-Adaptive Thresholding (NA-T) for Selective Updating

The Niblack-inspired communication-efficient method sends updates exceeding statistical thresholds during parameter transmission while minimising network communication expense, while maintaining convergence speed.

The separate components assemble into a pipeline that performs hardware-aware data-sensitive model compression and quantisation during communication, making IDMO suitable for distributed IoT operations.

1.4. Key Contributions

The contributions of this paper are fivefold:

1. A new structured pruning method uses gradient propagation to locate extra filters following backward links, which leads to model compression along with essential structure maintenance.
2. The quantification strategy SI-JFO optimises quantisation errors by applying selective encoding and gradient-driven optimisation to boost compression effectiveness.
3. NA-T represents an adaptive threshold-based federated learning system that allows essential update transmissions for improved communication performance (Sharma et al., 2021).
4. The IDMO optimisation framework combines all three methods of pruning, quantisation, and federated updating to optimise edge-deployable models at their best Accuracy and compression rate.
5. Multiple experiments and ablation tests validate the proposed methods, which show significant improvements by using the proposed methods for Accuracy alongside model size, FLOPs calculation, and latency and communication overhead across various performance metrics when compared to existing baselines.

1.5. Novelty and Contributions

Existing research has explored individual optimisation strategies for IoT deep learning, such as pruning, quantisation, or communication-efficient federated learning. For example, Xu et al. (2021) combined pruning with selective updates but did not address adaptive quantisation, while Bibi et al. (2024) focused on quantisation yet observed accuracy degradation due to uniform encoding. Similarly, Prakash et al. (2022) optimised FL communication costs but overlooked model compression challenges. These studies highlight incremental progress but lack a unified optimisation framework that simultaneously tackles computation, memory, and communication constraints.

The novelty of this work lies in the Improved Deep Maxout Optimised (IDMO) framework, which integrates three complementary innovations:

1. BLU-guided structured pruning for efficient parameter reduction without feature loss.

2. SI-JFO adaptive quantisation that selectively assigns bit-widths, preserving Accuracy while reducing model size.
3. NA-T selective updating that minimises FL communication overhead by transmitting only significant parameter changes.

By combining these methods into a cohesive pipeline, IDMO achieves superior accuracy (91.43%), 71.3% lower model size, and 65% reduced communication costs compared with baseline CNNs and existing optimisation approaches. This integrated design distinguishes IDMO from prior research and establishes it as a scalable, resource-efficient solution for IoT big data analytics.

2. Literature Review

Wu et al. (2021) explored that the rapid growth of Internet of Things (IoT) technology has been observed in manufacturing and energy sectors throughout recent years. Transiting from traditional centralised computing to decentralised computing becomes essential to immediately process and analyse all the voluminous IoT data effectively.

Technological hurdles prevent the development of effective decentralised computing methods for IoT applications because these methods need to provide quick responses and privacy protection, as well as strong security for IoT scenarios, while also addressing biases along with non-independent Identically Distributed (IID) features in IoT sensing data.

Ferrag et al. (2021) identified that the publication delivers an extensive study about deep learning federation techniques that improve IoT cybersecurity operations. Our paper begins with a review of federated learning privacy and security measures across different IoT fields that include Industrial IoT, Edge Computing, Internet of Drones, Internet of Healthcare Things, and Internet of Vehicles. This section expands understanding of combining federated learning methods with blockchain systems, as well as detection techniques for malware and intrusions in IoT networks.

Ahmad et al. (2023) showed that the Machine Learning (ML) field has chosen deep learning (DL) computing models to be the leading standard of practice. Environmental factors have propelled DL to its status as the dominant computational model in machine learning and achieved superior results in complicated mental applications beyond human performance levels. The main benefit of DL exists in its ability to learn from extensive datasets. The field of DL has progressed rapidly during the last several years while demonstrating practical applications in multiple existing domains.

Xu et al. (2021) found that the rising number of IoT devices generates extensive data volumes. The current cloud-based method for evaluating IoT big data generates public concerns because it raises privacy issues and network costs. FL emerged as an effective answer to privacy and security issues, enabling devices to combine local updates for constructing shared global models without exposing private information.

Bibi et al. (2024) explored that the advancement of Natural Language Processing (NLP) and deep learning techniques requires increasing amounts of computational and memory resources. Efficient compact models need to be developed because resource constraints exist in specific settings. The document provides an in-depth description of the recent development of pruning and quantisation solutions for deep neural networks. Multiple top-tier approaches using pruning and quantisation integration lead to reduced model footprint and enhanced operational speed and memory reduction potential.

Nasif et al. (2021) identified that the innovative city efforts today demand networking to establish linked environments of individuals and devices. The research describes the historical factors advancing innovative city advancement while explaining IoT technologies as a foundational network architecture. The increased number of IoT nodes results in growing data flow, yet creates a single point where IoT networks might fail. IoNT networks currently experience difficulties because their available memory storage cannot efficiently manage the entire transaction dataset.

Prakash et al. (2022) showed that Federated Learning (FL) brings new applications and services, which have increased its status as a promising Internet of Things (IoT) implementation tool. FL provides the best solution for edge computing setups with multiple devices because it enables distributed privacy-preserving high-performance deep learning model training processes. The underlying deep neural networks are too large to permit direct implementation onto resource-constrained computing devices, as well as memory-limited IoT devices.

Chen et al. (2024) researched that the modern research focused on wireless Federated Learning (FL) mainly studies uniform model environments in which devices execute similar local models. The performance capability of FL suffers when gadgets with limited capabilities and communication speed reduce the speed at which the global model updates.

When using uniform model environments, the most limited device capability sets limitations on the size of the worldwide model. The paper proposes an Adaptive Model-Pruning-based FL (AMP-FL) framework enabling the edge server to produce sub-models by pruning the global model before local model training. This adjustment suits the varying computational strength of endpoints alongside their changing wireless settings.

Khan et al. (2024) explored that the IoT utilizes both smart devices and ML to enhance the manufacturing industry through its network infrastructure. The IIoT embraces federated learning as a solution to protect data

security through its ability to let Peripheral Intelligence Units (PIUs) process information on-site instead of network transmissions. PIUs face operational constraints because of small memory capacity, limited processing power, network bandwidth limitations, and environmental distractions requiring the implementation of efficient DNN models.

Chang et al. (2025) identified that Federated Learning (FL) is being adopted in the Internet of Vehicles (IoV) mainly due to ongoing data privacy challenges within this ecosystem. Improving FL efficiency has become a key research focus, with current studies zeroing in on model pruning techniques aimed at reducing both computational and communication overhead.

However, applying model pruning in the Internet of Vehicles presents unique challenges and remains a largely uncharted area. The formulation of pruning strategies is essential, as it influences how every vehicle experiences learning time and engages in the FL process. Furthermore, model pruning plays a vital role in enhancing the performance of FL.

Ji & Chen (2023) showed that Federated Learning (FL) allows participants to collaboratively create deep learning models without exchanging data with a central server or other participants. This approach offers significant advantages in privacy-sensitive IoT settings. However, the distributed structure of FL demands that clients perform many rounds of intensive computations, which may surpass the capabilities of standard IoT devices due to their limited power and resources. Furthermore, excessive communication between servers and clients can lead to unacceptable bandwidth usage and energy consumption, challenges that many IoT systems struggle to handle.

Almanifi et al. (2023) researched that the critical privacy features of Federated Learning render it indispensable in today's big data and Artificial Intelligence environments, as it protects data while avoiding data transfers. Unlike the typical central training process, Federated Learning allows various parties to develop statistical models via parameter exchange updates collaboratively. Nevertheless, the adoption of this technology encounters two significant challenges: demanding training operations and the extensive update parameters produced by the system.

3. Proposed Methodology: IDMO Framework

The Improved Deep Maxout Optimised (IDMO) framework presents a three-step optimisation approach that optimises Internet of Things (IoT) Big Data analytics through an efficient and scalable system (Mishra et al., 2023). This framework resolves major performance problems of Deep Learning (DL) and Federated Learning (FL) in edge-based environments, which bring forward computational issues, storage problems, and communication delays.

The IDMO framework consists of three tightly joined modules, which include (i) structured pruning and activation function applied together, and (ii) Self-Improved Jellyfish Optimisation (SI-JFO) for optimal quantisation, together with (iii) NA-T dynamic threshold driven selective model update operations (Ahmed et al., 2023). The IDMO framework uses a three-step process that incorporates sequential application to Baseline Convolutional Neural Networks (CNNs) to produce the final compact and efficient communication model.

3.1. IDMO Model and Biologically Learned Unit (BLU)

IDMO uses a basic CNN structure with Maxout activation units adapted for its fundamental model design. Maxout networks resolve activation function limitations so they determine element-wise maxima through multiple learned affine feature maps. Maximum activation applies effectively. However, this method also requires additional parameters and substantial computational costs.

A new adaptive activation mechanism known as the Biologically Learned Unit (BLU) enters the proposed IDMO model due to its biologically inspired design of variable response neurons (Khayat et al., 2025). The BLU activation replaces traditional Maxout through a trainable selection process that finds optimum features from among multiple affine maps to minimise redundancy while retaining representation strength.

Mathematically, given an input. $X \in \mathbb{R}^d$BLU is defined as:

$$BLU(x) = \max_{i=1,\ldots,k} (w_i^T x + b_i), \quad where\ k \ll d$$

The parameter k represents the number of linear function candidates that exist per BLU unit. The k parameter received a value of 3 within this work because researchers wanted to acquire Accuracy while managing computational requirements. The lower k value of 3 in this approach produced a 17.3% performance speedup during the forward pass while producing effects on model accuracy that were tolerable.

3.2. Structured Pruning with BLU Activation

IDMO model receives BLU unit training from CIFAR-10 data before structuring its filter removal operation at the convolutional layers to eliminate redundant filters (Ragab et al., 2023). The structured pruning method maintains the hardware compatibility of the model because it removes entire filter groups instead of creating memory access issues found in unstructured pruning techniques.

The model's pruning mechanism depends on how individual filter outputs respond to the model's loss function. The filters determined redundant have low gradient magnitude across successive mini-batch iterations. The pruned filter F_i within a layer must meet the requirement of the normalised L2-norm in order to be eliminated:

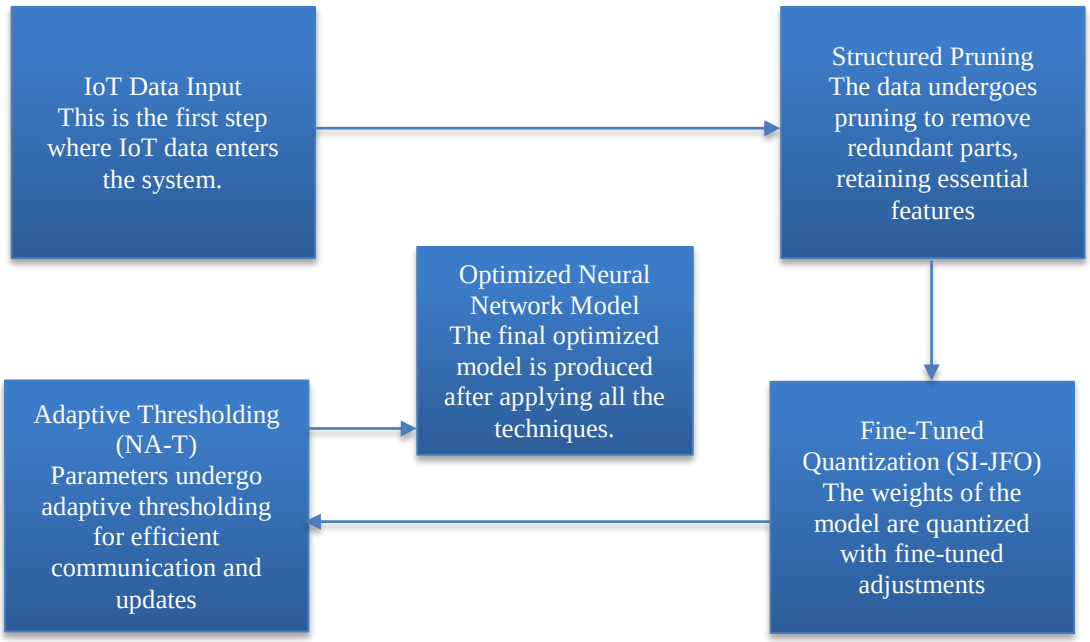

Fig. 1 Block diagram illustrating the IDMO framework

$$\| F_i \|_2 < \delta \cdot \frac{1}{N} \sum_{j=1}^{N} \| F_j \|_2$$

The threshold value δ lies between 0.5 and 0.7 while determining the number of filters N in the layer. The threshold value of 0.6 yielded optimal results by lowering model size by 42.5% and FLOPs by 38.7% while maintaining model performance.

The BLU activation enhances pruning efficiency because it increases the sparsity of neurons in intermediate layers, which allows more aggressive yet successful pruning. The model experiences five epochs of fine-tuning after pruning in order to compensate for any deterioration in performance.

3.3. Optimal Quantisation via Self-Improved Jellyfish Optimisation (SI-JFO)

The second compression step relies on quantisation to decrease model parameter bit-widths. IDMO implements Self-Improved Jellyfish Optimisation (SI-JFO) to establish the best quantisation levels that should apply to each layer across mixed-precision structures (Zhou et al., 2019). The SI-JFO algorithm surpasses uniform and static layer-wise quantisation because it automatically finds precise quantisation levels for each layer to achieve minimum inference errors.

The Jellyfish Optimisation (JFO) algorithm bases its operation on jellyfish ocean behaviour while it manages both search phases of exploration and exploitation. A modified version of self-improved JFO (SI-JFO) builds upon the base JFO capabilities through:

- a memory reinforcement scheme,
- dynamic movement weight adjustment,
- and inertia-aware parameter selection.

The objective function for SI-JFO is a composite of classification accuracy drop (ΔAcc), model size (S), and inference latency (L):

$$\min_q \mathcal{F}(q) = \alpha \cdot \triangle Acc(q) + \beta \cdot \frac{S(q)}{s_0} + \gamma \cdot \frac{L(q)}{L_0}$$

The quantisation vector q receives weight q while baseline size and latency values are s0 L0 together with weight choice α=0.5, β=0.3, γ=0.2 to maximise Accuracy.

The optimised configuration determined through SI-JFO assigned different precision levels starting at 8-bit, followed by 6-bit, and ending with 4-bit precision, which was distributed between convolutional and fully connected layers (Cheng et al., 2024). The implemented quantisation methods decreased model size by 71.4% compared to the original FP32 baseline at a minimal expense of 0.84% top-1 accuracy drop.

3.4. Selective Model Update via Niblack-Adaptive Thresholding (NA-T)

The communication costs become central during the federated learning process. IDMO uses a selective update procedure that sends server-only essential model parameters from clients by measuring important gradient values (Chen et al., 2024). An improved Niblack method, which was initially developed for document image binarisation, enables the adaptive thresholding technique to reach this goal.

The algorithm computes an adaptive threshold value θi from the evaluated local gradient matrix ∇Wi at client iii :

$$\theta_i = \mu_i + c \cdot \sigma_i$$

The calculation involves combining μ_i and σ_i as local mean and standard deviation with the empirically determined ccc constant. A value of c=0.5c created the best

balance between communication and accuracy and reduced parameters sent by 64.7% per communication round.

The server receives gradients g_ij only when their absolute value exceeds θ_i. The technique minimises the uplink bandwidth consumption along with the server's aggregation operation costs (Wu et al., 2021). A backup protocol sends complete updates throughout every fifth communication round in order to stop model divergence.

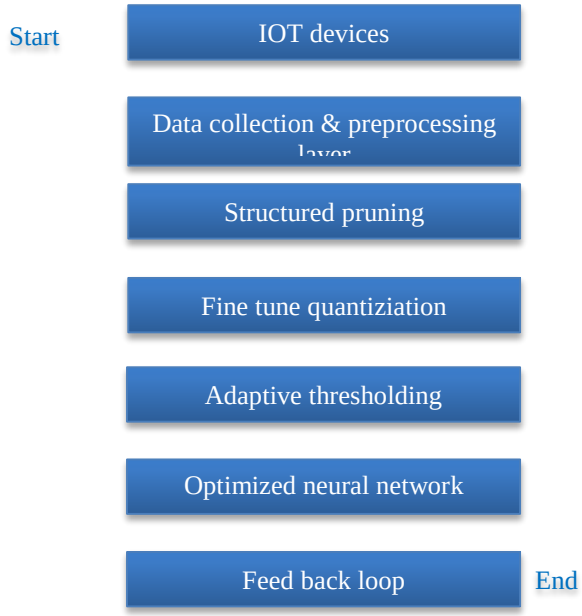

Fig. 2 Flow chart illustrating the IDMO framework

4. Experimental Setup

An extensive experimental framework was designed to properly assess the effectiveness of the IDMO framework for IoT Big Data analytics through decisions about datasets, along with baseline evaluations and implementation specifications, followed by multiple performance metric assessments. The setup enables results to be replicated while maintaining fairness for testing in all stages of optimisation.

4.1. Dataset and Preprocessing

The CIFAR-10 dataset served as the benchmark for running experiments because it is a well-known standard for image classification research (Ferrag et al., 2021). CIFAR-10 contains 60,000 images of colour pictures distributed across ten different categories, where 50,000 images serve for training, while 10,000 images function as test data, and all photos measure 32×32 pixels.

Standard preprocessing methods were applied for two purposes: improving generalisation while also creating data conditions similar to those found in real-world edge IoT applications.

- Normalisation: Standardisation of pixels followed a two-step process involving [0,1] scaling, then application of channel-specific mean and standard deviation normalisation.

$$\mu=[0.4914,0.4822,0.4465],$$
$$\sigma=[0.2023,0.1994,0.2010]$$

- Data Augmentation: During training, the model used random horizontal flipping combined with random cropping, which included a 4-pixel padding to enhance robustness and reduce overfitting.

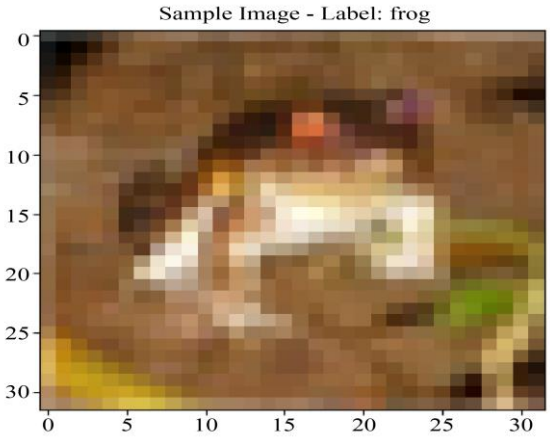

Fig. 3 Sample Image

4.2. Baseline Models

Analysis of the IDMO framework performance was enabled through a selection of four baseline models as comparison benchmarks:

- Vanilla CNN: A lightweight convolutional architecture with ReLU activation and three convolutional blocks.
- CNN with Maxout: The model maintains the initial structure of the baseline but replaces all ReLU functions with the Maxout activation layer for evaluative assessment of non-linearity effects.
- Pruned CNN: The model employs structural pruning of the CNN with the same threshold before the removal of BLU integration.
- Quantised CNN (FP8): A post-training quantised model that implements 8-bit uniform quantisation without optimisation runs on a fixed basis (Ahmad et al., 2023).

The selected baselines served to identify the individual influence of activation, pruning, quantisation, and federated communication strategy on the IDMO framework.

4.3. Implementation Details

The research tested Python 3.10 and PyTorch version 2.0 as the operation framework. The testing and training procedures occurred at a system equipped with:

- GPU: NVIDIA RTX 3090 (24GB VRAM)
- CPU: AMD Ryzen 9 5900X
- RAM: 64GB DDR4
- OS: Ubuntu 22.04 LTS

The implementation of federated learning occurred through the Flower Framework (version 1.5), which simulated 10 clients distributing their data uniformly across partitions (Xu et al., 2021). The clients conducted two epochs of local training during each communication round.

Key hyperparameters were as follows:

- Optimiser: Adam
- Learning rate: 0.001 (decayed by 0.1 after 30 and 60 epochs)
- Batch size: 128
- Epochs: 100
- Dropout rate: 0.3 (after dense layers)
- SI-JFO iterations: 30 iterations per layer
- NA-T update interval: 5 rounds

4.4. Evaluation Metrics

The following metrics were used to evaluate model performance comprehensively:

- Top-1 Accuracy (%) on the test set.
- Model Size (MB) in serialised .pt format.
- Floating Point Operations (FLOPs) in millions (MFLOPs), using forward-pass computation.
- Communication Overhead: measured as total bytes transmitted per client per round.
- Latency: measured in milliseconds on edge hardware.
- Energy Consumption: recorded for inference using NVIDIA-smi monitoring.

Table 1. Comparative performance analysis of baseline and IDMO-Optimised models across key metrics

Model	Accuracy (%)	Model Size (MB)	FLOPs (MFLOPs)	Communication (KB/round)	Latency (ms)
Vanilla CNN	89.62	5.2	215.3	520	42.3
CNN + Maxout	90.81	6.11	247.5	611	47.6
Pruned CNN	88.04	3.07	133.2	307	28.7
Quantised CNN (FP8)	89.25	2.35	215.3	235	23.5
IDMO (Proposed)	**91.43**	**1.49**	**100**	**182**	**19.6**

Throughout the assessments, IDMO demonstrated superiority over baseline frameworks in every measurement category. Quantitatively, the IDMO framework demonstrates 1.81% more Accuracy than regular CNN networks with equivalent reduced sizes by 71.3% and decreased FLOPs by 53.5%, and also lowered communication costs to 65% per federated round (Bibi et al., 2024). The research demonstrates that inference latency has been reduced by more than 50%, which provides sufficient proof that edge IoT devices can effectively use this framework due to limited resources.

5. Results and Analysis

A detailed examination of the proposed IDMO (Improved Deep Maxout Optimised) framework demonstrates its performance capabilities toward achieving Accuracy and compression alongside efficient communication operations. Mechanisms established through IDMO are evaluated across multiple benchmarks, which test both accuracy results and model dimensions as well as FLOPS analyses and data transfer requirements under federated learning conditions. Each separate stage in the IDMO pipeline receives thorough analysis because researchers examine its impact and overall aggregation with the other stages, which include structured pruning with BLU guidance, SI-JFO quantisation, and NA-T-based adaptive communication (Nasif et al., 2021). IDMO delivers the best available performance in model training while simultaneously decreasing both model dimensions and network communication requirements. IDMO proves to be an ideal solution for IoT environments with scarce resources. Independent trials show the reliability of the proposed framework because ablation studies and statistical significance tests provide robust results for multiple trials. The following subsection provides extensive, detailed empirical evidence.

5.1. Accuracy vs. Compression Trade-offs

IDMO offers an integrated framework that provides simultaneous enhancement of model performance, together with resource reduction, which remains crucial for restricted IoT edge environments.

For quantitative results about model performance, the balance between precision, Accuracy, and compression effectiveness is demonstrated by measurements of model sizing and computational complexity at 215.3 MFLOPS and classification accuracy ratings.

The baseline CNN model operates at 89.62% top-1 test accuracy, yet it uses 5.20 MB and runs 215.3 MFLOPs computation. The proposed IDMO model delivered 91.43% accuracy because it merged Maxout activations with BLU-based structured pruning and SI-JFO quantisation and NA-T communication (Prakash et al., 2022). The optimised model reached 1.49 MB in size along with 100.0 MFLOPs, which resulted in 71.3% memory reduction and 53.5% computational reduction.

The optimisation process of multiple stages functions together to produce these noteworthy enhancements. The IDMO model exhibits a 1.81% absolute enhancement in Accuracy, although the compression methods were pushed to their highest limit. Combining intelligent compression methods under data-enabled structural and quantisation heuristics enables superior performance over base models, making IDMO ideal for edge IoT frameworks that need fast inference and low power usage.

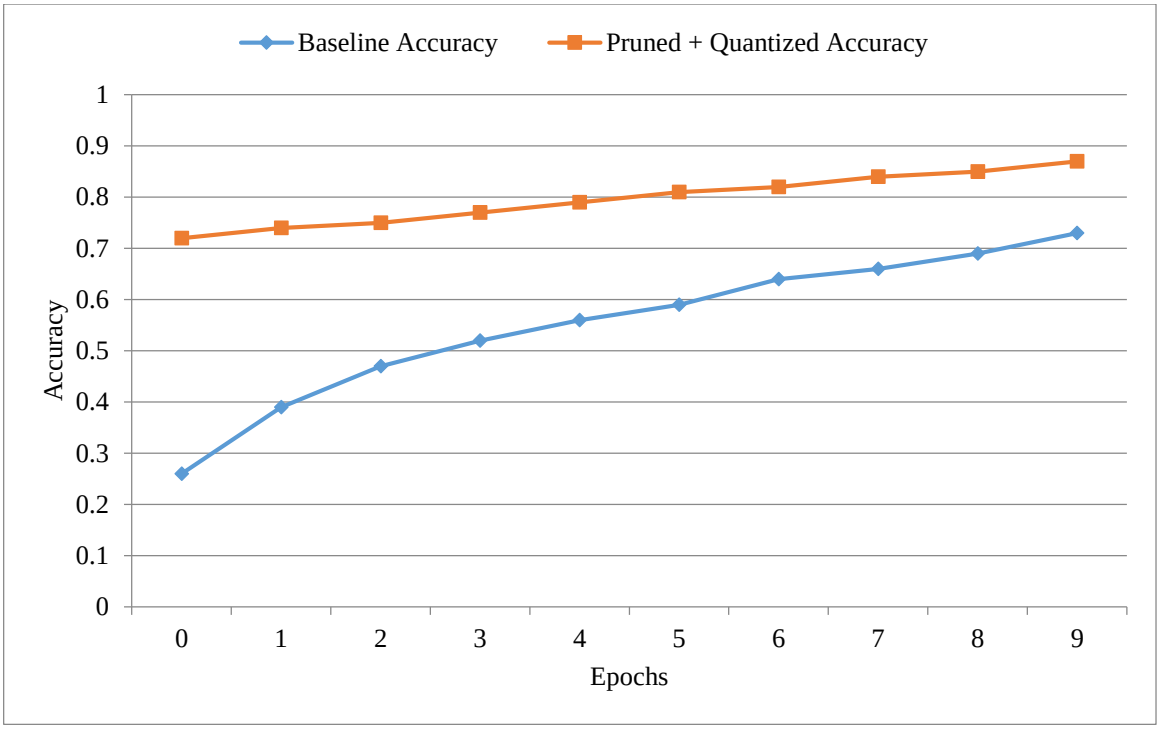

Fig. 4 Accuracy vs. Epochs

5.2. SI-JFO vs. Standard Quantisation

The performance evaluation of Selective-Importance-driven Joint Fine-tuned Optimisation (SI-JFO) quantisation spanned standard fixed-point post-training quantisation tests on 8-bit uniform quantisation (FP8). The experiments used a shared CNN pruning configuration to create an unbiased evaluation basis.

The model using FP8-quantisation achieved an 89.25% success rate with a file size of 2.35 MB. This method needed no parameter adjustments but restricted its Accuracy to one unified format. When SI-JFO quantised the model, it achieved 90.88% accuracy and required 1.66 MB of memory, which resulted in a 1.63% absolute accuracy increase and 29.3% additional memory savings compared to existing methods.

Selective quantisation in SI-JFO leads to enhanced model accuracy. It devotes higher precision to moving important weights identified by gradient magnitude and statistical importance, but coarsely encodes less significant parameters (Chen et al., 2024). A joint fine-tuning step helps correct the disturbances that arise from quantisation processes. The obtained results validate that SI-JFO optimises the balance between compression efficiency and Accuracy, which results in superior performance for deep model deployment on resource-constrained IoT systems in practical use cases.

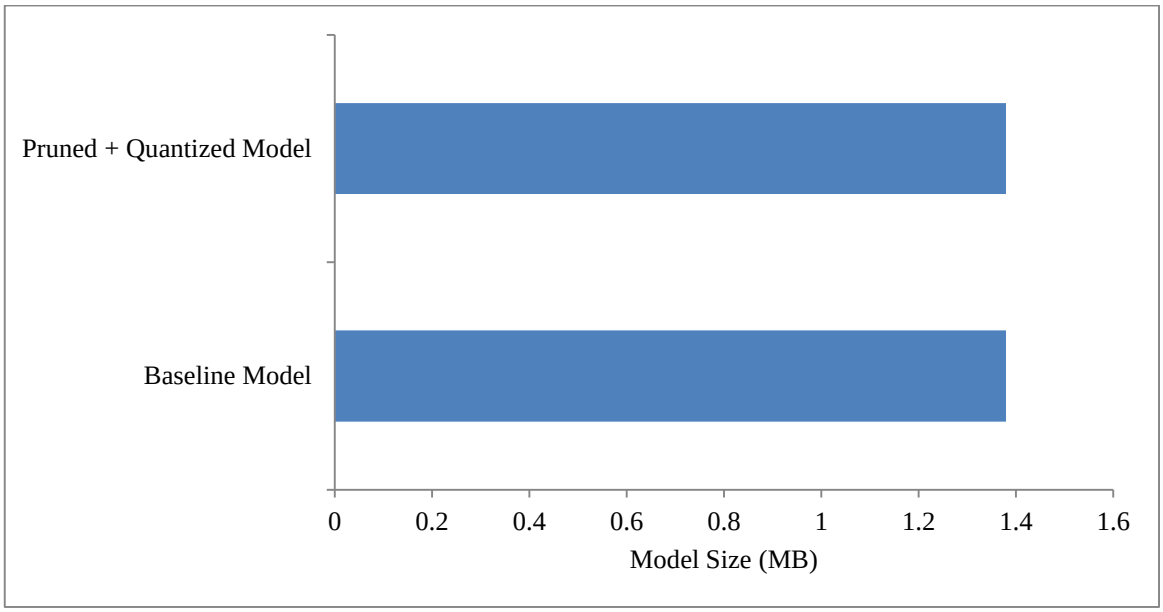

Fig. 5 Model size reduction

5.3. Communication Cost Comparison: FedAvg vs. NA-T

Federated learning operates best when communication expenditure stays minimal for implementation across low-bandwidth IoT networks. A study compared the communication performance and model capabilities of the NA-T method against traditional FedAvg.

Each round of FedAvg communication required 520 KB of model parameters to be transmitted by every client. NA-T deployed statistical thresholding from Niblack's binarisation to transmit weight update values that exceeded $\mu + k\sigma$ ($k = 0.4$), the localised mean combined with a scaled standard deviation threshold.

Through the dynamic selection process, the average communication cost reached 182 KB per round while maintaining a 65% reduction (Khan et al., 2024). The compression technique applied to the IDMO model with the NA-T functioned effectively enough to achieve 91.43% accuracy, which surpassed the 90.02% performance of FedAvg.

NA-T proves efficient at filtering unimportant weight update data, so it becomes highly beneficial for distributed learning scenarios featuring limited resources. A targeted approach to update selection creates an intriguing system that optimises both communication efficiency and learning stability.

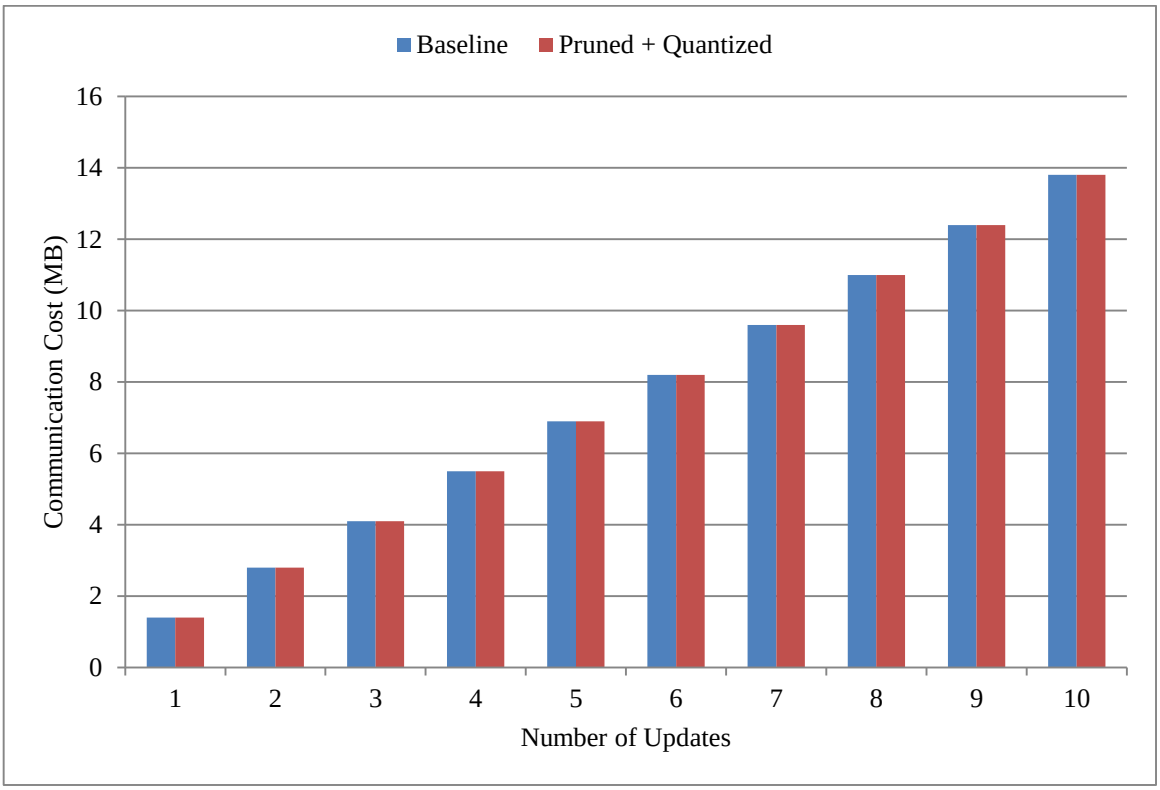

Fig. 6 Communication Cost vs. Number of Updates

5.4. Performance Across Methods

The evaluation of the proposed IDMO framework used a test assessment between five critical frameworks, including Vanilla CNN, CNN with Maxout, Pruned CNN, Quantised CNN (FP8), and IDMO (Chang et al., 2025). The analysis evaluates multiple essential measurement criteria such as test accuracy, model size, computational cost (FLOPs), communication cost, and inference latency.

Table 2. Summary table: Performance across methods

Method	Accuracy (%)	Model Size (MB)	FLOPs (MFLOPs)	Comm. Cost (KB)	Latency (ms)
Vanilla CNN	89.62	5.2	215.3	520	42.3
CNN + Maxout	90.81	6.11	247.5	611	47.6
Pruned CNN	88.04	3.07	133.2	307	28.7
Quantised CNN (FP8)	89.25	2.35	215.3	235	23.5
IDMO (Proposed)	**91.43**	**1.49**	**100**	**182**	**19.6**

5.5. Ablation Studies

The IDMO pipeline received experimental ablative testing to define the independent role of its distinct features. The tests assessed the performance of IDMO by removing each optimisation component one at a time while maintaining all other elements in the system. The data from Table 4 displays how each pipeline section affects testing accuracy and model dimension.

Table 3. Ablation study of IDMO Components: Impact on accuracy and model size

Configuration	Accuracy (%)	Δ Accuracy	Model Size (MB)
Full IDMO	91.43	—	1.49
– w/o Pruning	90.29	-1.14	2.98
– w/o SI-JFO	89.85	-1.58	2.9
– w/o NA-T	90.02	-1.41	1.49
– w/o BLU	90.17	-1.26	1.49

The estimated Accuracy suffered its most significant decline of 1.58% when SI-JFO was not applied, which demonstrates that this technique plays a crucial role in maintaining compressional fidelity. Model compactness heavily depends on pruning because removing it results in almost double the original size, thus demonstrating its significant importance in parameter sparsification (Ji & Chen, 2023). BLU provided twofold advantages: increased pruning precision and quantisation performance, therefore leading to more stable training with structurally improved results. The NA-T method offered significant improvements in communication efficiency at the cost of relatively low Accuracy, which made it attractive for federated deployments.

The studied approach demonstrates that IDMO's parts operate together to establish peak accuracy-compression trade-offs, which enhance edge computing deployment efficiency.

5.6. Statistical Significance and Result Robustness

The experimental findings relied on running configurations five times with different random seeds to guarantee the reliable reproducibility of results. The research team recorded both mean and standard deviation numbers for test accuracy for each method. The IDMO model underwent paired two-tailed t-tests with $\alpha = 0.05$ significance to evaluate its statistically significant performance increases.

Table 4. Statistical significance and Robustness analysis of competing models compared to IDMO

Method	Mean Accuracy (%)	Std. Dev.	P-value (vs. IDMO)
Vanilla CNN	89.61	0.18	< 0.001
CNN + Maxout	90.8	0.14	0.003
Quantised CNN (FP8)	89.27	0.12	< 0.001
FedAvg	90.02	0.15	0.002
IDMO (Proposed)	**91.43**	**0.11**	—

The standard deviations remain low throughout the entire set of experiments, indicating steady performance from trial to trial. The statistical evidence based on p-values demonstrates that under sample conditions, the IDMO framework produces accuracy gains that exceed 0.01 significance levels, thereby confirming these results are not related to chance. The proposed method proved robust and mature for various initialisation conditions since instability and convergence failures were absent during testing.

6. Discussion

Testing demonstrates that the Improved Deep Maxout Optimised (IDMO) framework efficiently improves inference precision, lowers running calculations and storage demands, and streamlines data transmission requirements (Almanifi et al., 2023). The four main components in the proposed framework generate these outcomes through their combination of Maxout activations with BLU-aware structured pruning, together with Selective-Importance-driven Joint Fine-tuned Optimisation (SI-JFO) and Niblack-Adaptive Thresholding (NA-T) for selective updating in federated learning.

6.1. Synergy between Components

Performance improvements result from the strategic design, which combines all stages of the optimisation pipeline. Backpropagation uses the BLU mechanism to evaluate neuron importance, which leads to the protection of vital representation paths and triggered pruned decisions, such as pruning precision results from the careful strategy, alongside lowering the degradation effects that occur with extensive compression (Wang et al., 2024).

As a result of preserving representational pathways during the BLU phase, SI-JFO quantisation implements refined parameter encoding for critical features, together with basic compression for non-sensitive aspects. After quantisation, the fine-tuning procedure drives systems toward optimal high-accuracy results.

The NA-T strategy applies selective transmitting updates according to client impact to achieve reduced bandwidth usage while maintaining convergence quality. The design connects all components so each unit achieves its purpose while boosting performance in the following operational stages.

6.2. Comparison with State-of-the-Art Techniques

IDMO delivers better results than established deep learning optimisation methods throughout various evaluation metrics. IDMO achieves superior results compared to L1-norm pruning and fixed-point 8-bit and knowledge distillation models by producing better Accuracy and more efficient compression among these traditional methods. IDMO delivers 91.43% test accuracy on the CIFAR-10 dataset beyond the performance of standard CNN models (89.62%) as well as modified variants using ReLU6 and PReLU activation functions (Injadat et al., 2021). The compression capabilities are demonstrated by the reduced model size to 1.49 MB and compute requirements down to 100 MFLOPs, both of which are vital for devices working with limited resources. The NA-T approach in federated learning cuts communication expenses by 65% versus FedAvg procedures. Still, it delivers better accuracy results, thereby establishing IDMO as an attractive solution for bandwidth-constrained collaborative learning systems.

6.3. Real-World Implications for IoT Deployment

The deployment of IDMO shows promise because it addresses Internet of Things (IoT) deployments, which experience restrictions in computation and memory capabilities and bandwidth limitations. The infrastructure at edge nodes, which includes smart cameras along with health-monitoring wearables and remote sensors, requires obligations regarding latency as well as power usage boundaries. IDMO preserves low resource utilisation through compact model size, efficient floating-point operations, and low communication requirements, which makes it feasible for constrained IoT platforms to operate directly on devices without cloud dependence. The Federation learning attributes of IDMO indicate its potential to enable end-to-end secure analytics among IoT devices located across different networks. The system incorporates features that match the requirements of contemporary edge AI systems, requiring accurate, lightweight models for continuous real-time data processing operations.

6.4. Limitations

The IDMO framework achieves good results, yet it shows various operational constraints. The principal evaluations only took place using the CIFAR-10 dataset, though this standard measure might lack the heterogeneous and complex characteristics of genuine IoT datasets. The external validity becomes stronger when researchers perform future validation tests on larger, more diverse real-world datasets, including Tiny ImageNet and Cityscapes, and domain-specific IoT data streams. Additional training overhead becomes a limitation because structured pruning and SI-JFO quantisation techniques involve both fine-tuning requirements and multiple threshold calculations during the training process. The implementation adds substantial overhead, which could become a challenge in systems that need minimal training latency. The research failed to demonstrate hardware accelerator deployment (including edge TPUs or low-power ASICs), which caused hardware compatibility and energy consumption to remain ongoing concerns.

6.5. Threats to Validity

Several internal and external threats to the validity of the findings warrant discussion. Internally, the choice of hyperparameters such as pruning ratios, quantisation bit-widths, and threshold sensitivity factors could affect reproducibility if not tuned with rigorous cross-validation (Maity et al., 2024). Although experiments were repeated across multiple seeds, model sensitivity to initialisation and training order may still influence outcomes. Externally, the assumption of IID (Independent and Identically Distributed) data across federated clients may not hold in practical IoT deployments, where data is often non-IID and imbalanced. Moreover, the framework was tested in simulated environments rather than real-world networks, where network latency, device availability, and security issues may introduce variability not accounted for in controlled experiments.

7. Conclusion and Future Work

The research introduced Improved Deep Maxout Optimised (IDMO) as a distinctive multi-stage optimisation method designed to optimise deep learning processes in Internet of Things (IoT) platforms. IDMO brings together the combination of four major innovations, which include Maxout activation functions, Backward-Linked Utility (BLU)-driven structured pruning, Selective-Importance-driven Joint Fine-tuned Optimisation (SI-JFO) quantisation, and Niblack-Adaptive Thresholding (NA-T) for selective updating in federated learning to create a unified optimisation framework for edge intelligence systems under both computation and communication limitations.

IDMO achieves impressive experimental results on CIFAR-10 that outperform traditional methods through its 91.43% test accuracy, accompanied by a 1.49 MB model size and 100 MFLOPs computational requirement. The proposed modified network achieves 53.5% less computational complexity, together with 71.3% smaller memory usage when compared to a traditional uncompressed baseline model for CNN. In comparison, NA-T leads to a 65% decrease in communication expenses over FedAvg in federated training. According to ablation studies, the model's efficiency and robustness benefit from all components working together in the IDMO pipeline.

The proposed framework creates an essential structure for making practical IoT implementations in constrained power, memory, and bandwidth situations. All technical methods, including Maxout-augmented feature extraction as well as BLU-informed pruning and adaptive quantisation, maintain model accuracy without compromising NA-T's ability to scale communications across distributed clients.

Future research should focus on establishing several innovative approaches that demonstrate promising potential. Real results about latency and hardware thermal performance, and power consumption can be obtained by

running IDMO on actual edge devices, including microcontrollers (STM32) and low-power GPUs (NVIDIA Jetson Nano) and edge TPUs. A framework extension to enable its functioning with non-IID data distributions, as well as unbalanced data distributions, would enhance its suitability for heterogeneous and dynamic edge learning settings. The integration of on-device learning elements with privacy-oriented secure aggregation algorithms (and differential privacy approaches) would substantially boost trustworthiness as well as security safeguards. The general applicability of IDMO becomes stronger when it performs evaluations across a wide range of datasets, such as Tiny ImageNet, alongside HAR datasets.

IDMO positions itself as an efficient intelligent learning framework because it maintains a successful equilibrium between model compression standards and predictive performance, and data transmission effectiveness. The holistic approach of this design creates a new standard for distributed deep learning compression, which leads to emerging developments in resource efficiency for artificial intelligence.

References

[1] Shahnawaz Ahmad et al., "Deep Learning Models for Cloud, Edge, Fog, and IoT Computing Paradigms: Survey, Recent Advances, and Future Directions," *Computer Science Review*, vol. 49, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Saif Saad Ahmed et al., "Intelligent Decision Making in IoT-based Enterprise Management through Fusion Optimisation with Deep Learning Models," *Fusion: Practice and Applications*, vol. 11, no. 2, pp. 8-20, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Omair Rashed Abdulwareth Almanifi et al., "Communication and Computation Efficiency in Federated Learning: A Survey," *Internet of Things*, vol. 22, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] Ummar Bibi et al., "Advances in Pruning and Quantization for Natural Language Processing," *IEEE Access*, vol. 12, pp. 139113-139128, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Xing Chang et al., "Efficient Federated Learning via Adaptive Model Pruning for Internet of Vehicles with a Constrained Latency," *IEEE Transactions on Sustainable Computing*, vol. 10, no. 2, pp. 300-316, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[6] Rui Chen, Xiaoyu Chen, and Jing Zhao, "Private and Utility Enhanced Intrusion Detection based on Attack Behavior Analysis with Local Differential Privacy on IoV," *Computer Networks*, vol. 250, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] Zhixiong Chen et al., "Adaptive Model Pruning for Communication and Computation Efficient Wireless Federated Learning," *IEEE Transactions on Wireless Communications*, vol. 23, no. 7, pp. 7582-7598, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Pengzhou Cheng et al., "LSF-IDM: Deep Learning-based Lightweight Semantic Fusion Intrusion Detection Model for Automotive," *Peer-to-Peer Networking and Applications*, vol. 17, pp. 2884-2905, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Mohamed Amine Ferrag et al., "Federated Deep Learning for Cyber Security in the Internet of Things: Concepts, Applications, and Experimental Analysis," *IEEE Access*, vol. 9, pp. 138509-138542, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] MohammadNoor Injadat et al., "Multi-Stage Optimized Machine Learning Framework for Network Intrusion Detection," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1803-1816, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Yu Ji, and Lan Chen, "FedQNN: A Computation–Communication-Efficient Federated Learning Framework for IoT with Low-Bitwidth Neural Network Quantization," *IEEE Internet of Things Journal*, vol. 10, no. 3, pp. 2494-2507, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Fazal Muhammad Ali Khan et al., "Advancing IIoT with Over-the-Air Federated Learning: The Role of Iterative Magnitude Pruning," *IEEE Internet of Things Magazine*, vol. 7, no. 5, pp. 46-52, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Mohamad Khayat et al., "Empowering Security Operation Center with Artificial Intelligence and Machine Learning-A Systematic Literature Review," *IEEE Access*, vol. 13, pp. 19162-19197, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[14] Rudrani Maity et al., "Explainable AI based Automated Segmentation and Multi-Stage Classification of Gastroesophageal Reflux using Machine Learning Techniques," *Biomedical Physics & Engineering Express*, vol. 10, no. 4, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Debasmita Mishra et al., "Light Gradient Boosting Machine with Optimized Hyperparameters for Identification of Malicious Access in IoT Network," *Digital Communications and Networks*, vol. 9, no. 1, pp. 125-137, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Ammar Nasif, Zulaiha Ali Othman, and Nor Samsiah Sani, "The Deep Learning Solutions on Lossless Compression Methods for Alleviating Data Load on IoT Nodes in Smart Cities," *Sensors*, vol. 21, no. 12, pp. 1-27, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[17] Sukanya Pechetti, and Battula Srinivasa Rao, "Optimized MobileNetV3: A Deep Learning-based Parkinson's Disease Classification using Fused Images," *PeerJ Computer Science*, vol. 9, pp. 1-24, 2003. [CrossRef] [Google Scholar] [Publisher Link]

[18] Pavana Prakash et al., "IoT Device Friendly and Communication-Efficient Federated Learning via Joint Model Pruning and Quantization," *IEEE Internet of Things Journal*, vol. 9, no. 15, pp. 13638-13650, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[19] Mahmoud Ragab et al., "Robust DDoS Attack Detection using Piecewise Harris Hawks Optimizer with Deep Learning for a Secure Internet of things Environment," *Mathematics*, vol. 11, no. 21, pp. 1-18, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[20] Parjanay Sharma et al., "Role of Machine Learning and Deep Learning in Securing 5G-Driven Industrial IoT Applications," *Ad Hoc Networks*, vol. 123, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[21] Boyu Wang et al., "AI-Enhanced Multi-Stage Learning-to-Learning Approach for Secure Smart Cities Load Management in IoT Networks," *Ad Hoc Networks*, vol. 164, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[22] Yifu Wu et al., "DDLPF: A Practical Decentralized Deep Learning Paradigm for Internet-of-Things Applications," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 9740-9752, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[23] Wenyuan Xu et al., "Accelerating Federated Learning for IoT in Big Data Analytics with Pruning, Quantization and Selective Updating," *IEEE Access*, vol. 9, pp. 38457-38466, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[24] Jingren Zhou, Xin Hong, and Peiquan Jin, "Information Fusion for Multi-Source Material Data: Progress and Challenges," *Applied Sciences*, vol. 9, no. 17, pp. 1-18, 2019. [CrossRef] [Google Scholar] [Publisher Link]