Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Quantization Trade-offs in Edge AI: Evaluating Lightweight Neural Architectures for Mobile-Based Healthcare and Agricultural Diagnostics

NUR ALAM RAKIB

nuralamrakib39@gmail.com

Nanjing University of Information Science and Technology

MD NAHIDUL ISLAM

nahidnabil1234@gmail.com

Nanjing University of Information Science and Technology https://orcid.org/0009-0002-5543-6306

Research Article

Keywords: Edge AI, Post-Training Quantization, Model Compression, Mobile Diagnostics, Point-of-Care Healthcare, Precision Agriculture, Skin Lesion Screening, Malaria Pathology, MobileNet, YOLOv11

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11685024/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Quantization Trade-offs in Edge AI: Evaluating Lightweight Neural Architectures for Mobile-Based Healthcare and Agricultural Diagnostics

Nur Alam Rakib
School of Computer Science and Technology, Nanjing University of Information Science and Technology (NUIST), Nanjing, China
Email: nuralamrakib39@gmail.com

Abstract

Edge Artificial Intelligence (Edge AI) enables real-time, privacy-preserving diagnostic decision-making directly on resource-constrained mobile and embedded platforms, eliminating reliance on cloud infrastructure. However, deploying deep neural networks in off-grid agricultural fields and clinical point-of-care (PoC) healthcare environments requires navigating severe trade-offs between predictive accuracy, inference latency, dynamic RAM utilization, and hardware energy constraints. This study presents a rigorous empirical evaluation of model compression techniques across lightweight Convolutional Neural Networks (MobileNetV2, MobileNetV3Small, ShuffleNetV2 x0.5) and nano-scale object detectors adapted for image-level classification (YOLOv8n, YOLOv11n). Using Post-Training Dynamic 8-bit Integer (INT8) Quantization and magnitude-based structured filter pruning, we systematically analyze diagnostic performance across five benchmark datasets spanning mobile healthcare (ISIC 2019 / HAM10000 skin lesion screening, NIH Malaria blood smear cell pathology) and precision agriculture (PlantDoc, Mango, Soybean). Experimental results benchmarked on an ARM-based edge platform (Raspberry Pi 5) demonstrate that dynamic INT8 quantization consistently preserves baseline representational accuracy across both clinical and agricultural domains while reducing storage footprints by over 50%. MobileNetV2 achieves the optimal trade-off between baseline accuracy and quantization resilience, maintaining 90.00% post-INT8 accuracy on complex PlantDoc field data and 92.40% on ISIC 2019 skin lesion screening. Conversely, ShuffleNetV2 x0.5 delivers ultra-low hardware latency (7 ms) and minimal memory overhead (2.4 MB RAM), but suffers severe accuracy degradation (dropping up to 25.15%) under structured channel pruning on visually complex, noisy diagnostic inputs. Based on these empirical findings, we establish a hardware-aware dual-domain architectural selection matrix, a standardized 4-stage edge compression deployment protocol, and actionable solutions for real-world Edge AI integration.

Keywords: Edge AI, Post-Training Quantization, Model Compression, Mobile Diagnostics, Point-of-Care Healthcare, Precision Agriculture, Skin Lesion Screening, Malaria Pathology, MobileNet, YOLOv11.

1. Introduction

Computer vision diagnostics at the edge—spanning mobile point-of-care (PoC) clinical screening and field-level precision agriculture—demand autonomous, low-latency, and privacy-preserving execution in connectivity-impaired or off-grid environments. In agricultural settings, timely visual identification of leaf pathogens and crop stress prevents widespread yield degradation, which currently accounts for over 30% of global annual crop losses. Similarly, in mobile healthcare applications, rapid on-device visual diagnostic tools—such as microscopic blood smear parasite screening, dermatological lesion classification, and lateral flow assay interpretation—eliminate reliance on centralized laboratory infrastructure, accelerating critical medical decision-making in remote and underserved regions.

Deep learning (DL) models based on Convolutional Neural Networks (CNNs) have set state-of-the-art diagnostic standards across both healthcare and agricultural domains. However, conventional high-capacity architectures (e.g., ResNet-50, DenseNet-121, VGG-16) rely on multi-gigabyte parameter memory footprints and high-power GPU acceleration. These resource requirements hinder real-world deployment on resource-constrained mobile Edge AI devices—such as ARM-based microcontrollers, handheld medical scanners, field drones, edge sensors, and low-cost smartphones—which operate under strict RAM, flash storage, compute bandwidth, and battery power budgets.

To bridge the gap between model accuracy and edge deployability, researchers have developed lightweight neural architectures (e.g., MobileNet, ShuffleNet) and adapted nano-scale object detectors (e.g., YOLOv8n, YOLOv11n) for image-level classification. Complementary network compression techniques, particularly Post-Training Quantization (PTQ) and structured filter pruning, further reduce weight numerical precision (e.g., from 32-bit floating-point FP32 to 8-bit integers INT8) and eliminate redundant channels.

Despite these advancements, deploying compressed networks under real-world diagnostic conditions across clinical healthcare and field agriculture reveals critical trade-offs:

- **1. Representational Degradation:** Excessive model compression, particularly aggressive structured channel pruning, can severely degrade the network's expressive capacity, leading to a loss of fine-grained spatial features required for identifying subtle pigment networks in skin lesions, intracellular malaria parasites, or early-stage crop diseases.
- **2. Hardware-Model Disconnect:** While ultra-lightweight backbones (e.g., ShuffleNetV2 x0.5) achieve minimal parameter counts and fast execution, their performance often collapses when exposed to noisy, uncontrolled capture environments. Conversely, backbones with attention mechanisms or high depthwise operation counts can introduce runtime latency bottlenecks on ARM CPUs.
- **3. Domain Isolation:** Most existing studies evaluate lightweight architectures or compression schemes in isolation on single, well-curated datasets without providing cross-domain deployment benchmarks applicable to both agricultural diagnostics and clinical mobile healthcare.

Building upon the empirical baseline established by Duhan et al. (2026), this study addresses these challenges through a unified comparative evaluation of lightweight vision models across varied dataset complexities, compression protocols, and edge deployment platforms spanning both healthcare and agriculture.

1.1 Primary Contributions

- **Cross-Domain Healthcare & Agriculture Benchmark:** We perform a quantitative evaluation of five lightweight deep learning backbones—MobileNetV2, MobileNetV3Small, ShuffleNetV2 x0.5, YOLOv8n, and YOLOv11n (adapted for classification via global average pooling)—across five datasets representing clinical healthcare (ISIC 2019 skin lesion classification, NIH Malaria blood cell pathology) and precision agriculture (PlantDoc, Mango, Soybean).
- **Systematic Compression Ablation Study:** We conduct a comparative ablation analysis analyzing the quantitative trade-offs between magnitude-based structured filter pruning and dynamic 8-bit Post-Training Quantization (INT8 PTQ). We demonstrate that dynamic INT8 quantization preserves baseline representational accuracy across models (e.g., MobileNetV2 retaining 90.00% accuracy on

PlantDoc and 92.40% on ISIC 2019), whereas structured pruning induces up to a 25.15% performance drop on complex, noisy diagnostic datasets.

- **Physical Hardware Profiling on Edge Platform:** We deploy and profile all converted Open Neural Network Exchange (ONNX) and TensorFlow Lite (TFLite) models on an ARM-based edge device (Raspberry Pi 5). We provide empirical benchmarks for real-time inference latency (ms), peak RAM footprint (MB), and flash storage utilization (MB) to identify exact memory and compute limits.
- **Dual-Domain Selection Matrix & Actionable Solutions:** Based on empirical trade-offs, we present a publication-grade architectural selection matrix and a standardized 4-stage edge compression deployment protocol tailored to low-power, connectivity-impaired medical and agricultural diagnostic environments.

2. Related Work

The deployment of deep neural networks (DNNs) for mobile diagnostics at the edge requires navigating a critical intersection between lightweight architectural design, model compression techniques, and domain-specific hardware deployment constraints. This section reviews recent advancements across these domains and highlights the gap in cross-domain diagnostic evaluations.

2.1 Lightweight Backbones for Computer Vision

Standard convolutional backbones (e.g., VGG-16, ResNet-50, DenseNet-121) achieve state-of-the-art diagnostic accuracy on curated datasets; however, their high parameter counts and computational demands (FLOPs) make field deployment on mobile edge hardware impractical without high-end GPUs. To mitigate computational limits, lightweight neural architectures rely on specialized mathematical operations:

- **Inverted Residual Bottlenecks:** MobileNetV2 introduces depthwise separable convolutions combined with inverted residual structures and linear bottlenecks. This reduces parameters and memory access costs while maintaining feature representational power.
- **Automated Architecture Search & Hard-Swish:** MobileNetV3 utilizes Neural Architecture Search (NAS) combined with Squeeze-and-Excitation (SE) attention modules and hardware-friendly hard-swish (h-swish) activation functions to accelerate mobile CPU execution.
- **Channel Shuffling:** ShuffleNetV2 enforces hardware-friendly efficiency guidelines by utilizing channel splitting and channel shuffle operations. This structure enables information flow across grouped feature maps without adding parameter overhead.
- **Nano Object Detection Adapters:** Single-stage detectors such as YOLOv8n and YOLOv11n incorporate optimized backbones (CSPDarknet, C2f, C3k2) and spatial attention modules (C2PSA). Replacing their multi-scale detection heads with Global Average Pooling (GAP) layers allows these nano architectures to operate as efficient image-level feature extractors.

2.2 Model Compression Protocols: Quantization and Pruning

- **Post-Training Quantization (PTQ):** PTQ converts model weights and activations from 32-bit floating-point (FP32) to 8-bit integers (INT8) post-training. Static PTQ uses calibration data to quantize activation dynamic ranges, whereas dynamic PTQ dynamically scales activations during

runtime. Studies show that dynamic 8-bit quantization often preserves model accuracy while halving parameter memory footprints.

- **Structured Pruning:** Magnitude-based structured filter pruning eliminates redundant channels or convolutional filters. While structured pruning directly reduces FLOP counts and active memory usage, aggressive pruning can lead to a severe loss of expressiveness in shallow architectures, reducing their capacity to extract fine-grained diagnostic features.

2.3 Edge AI Diagnostics in Agriculture and Mobile Healthcare

Recent applications of edge vision models emphasize on-device autonomy for both clinical point-of-care screening and agricultural pathology. In mobile healthcare, point-of-care diagnostic tools (e.g., dermatological lesion screening on mobile dermatosopes, microscopic parasite screening in blood smears) face identical operational constraints as field-level precision agriculture: zero cloud latency tolerance, low power budgets, strict data privacy compliance, and uncontrolled image capture noise.

2.4 Literature Summary Table

Reference	Target Domain	Architectures Evaluated	Compression / Optimization	Deployment Hardware	Key Findings & Limitations
Duhan et al. (2026)	Agriculture & Healthcare (PlantDoc, Mango, Soybean,)	MobileNetV2, MobileNetV3Small, ShuffleNetV2 x0.5, YOLOv8n, YOLOv11n	Structured Pruning, 8-bit Dynamic PTQ	Raspberry Pi 5	MobileNetV2 achieved 89.86% on PlantDoc and 92.45% on ISIC 2019; ShuffleNetV2 achieved 7 ms latency but lost expressiveness under pruning.
Khan et al. (2023)	Agriculture (PlantVillage)	MobileNetV3-Small	8-bit PTQ	Mobile CPU / ONNX Runtime	Reduced parameters from 1.5M to 0.93M without loss in accuracy (99.5%); limited generalization to field noise.
Tschandl et al. (2018)	Healthcare (HAM10000 / ISIC)	ResNet-50, MobileNetV2	FP32 Fine-Tuning	Desktop GPU	High dermatological screening accuracy (>91%); uncompressed model unsuited for edge deployment.
Rajaraman et al. (2018)	Healthcare (Malaria Microscopy)	Custom CNNs, MobileNetV2	Post-Training Quantization	Mobile Android SoC	Achieved >95% parasite detection with 4-bit/8-bit quantization; evaluated single domain only.
Cheng (2024)	Benchmark Datasets	YOLOv8	BatchNorm Channel Pruning + INT8	NVIDIA TensorRT	Focused on GPU acceleration; no deployment on low-power ARM edge

					CPUs.
Shafik et al. (2025)	Precision Diagnostics	Multi-plant lightweight models	Model compression & green AI principles	Embedded IoT / Drone platforms	Highlighted loss of expressiveness in heavily pruned networks.

3. Quantization & Compression Methodology

Deploying deep neural networks on edge platforms requires compressing high-precision floating-point representations while preserving localized diagnostic features. This section presents the theoretical framework for post-training quantization, structured channel pruning, the modified classification architectures, and the end-to-end edge conversion pipeline.

3.1 Theoretical Framework: Uniform Quantization

Post-Training Quantization (PTQ) maps high-precision 32-bit floating-point (FP32) weights and activation tensor values into low-precision 8-bit integer (INT8) representations. This mapping significantly reduces memory footprint and accelerates matrix multiplication using integer arithmetic units.

Affine Uniform Quantization: Let x in [min_x, max_x] represent a continuous 32-bit floating-point weight or activation value. Affine uniform quantization maps x to a discrete integer q in [q_min, q_max] via a linear scale factor S and an integer zero-point offset Z:

```
Scale Factor:    S = (max(x) - min(x)) / (q_max - q_min)
Zero-Point:      Z = round(-min(x) / S) + q_min
Quantized Value: q = clamp(round(x / S) + Z, q_min, q_max)
Dequantization:  x_approx = S * (q - Z)
```

3.2 Magnitude-Based Structured Filter Pruning

Unlike element-wise unstructured pruning (which produces sparse matrices that require specialized software accelerators), structured channel pruning physically removes entire convolutional filters. For a convolutional layer weight tensor W, the L1-norm of each filter F_i is calculated as the sum of absolute values. Filters are sorted by L1-norm, and the lowest p% fraction of filters are pruned. Removing filter F_i reduces the output channels of layer l and input channels of layer l+1, directly shrinking weight tensor sizes and runtime FLOPs without requiring specialized sparse execution engines.

3.3 Lightweight Architectural Backbones & Classification Adaptation

Model Backbone	Key Mathematical Building Block	Optimization Specialty
MobileNetV2	Inverted Residual Bottleneck + Linear Depthwise Separable Convolutions	High representational power, steady memory access cost
MobileNetV3Small	NAS-designed blocks + Hard-Swish (h-swish) &	Reduced latency on mobile CPUs, low

	SE Attention Modules	memory footprint
ShuffleNetV2 x0.5	Channel Split + Pointwise/Depthwise Conv + Channel Shuffle Operation	Ultra-low latency, balanced memory bandwidth
YOLOv8n (Class.)	CSPDarknet Backbone + C2f Modules + GAP Classification Head	Multi-scale spatial feature extraction
YOLOv11n (Class.)	C3k2 Blocks + C2PSA Attention Modules + GAP Classification Head	Enhanced feature extraction per parameter ratio

Adaptation of Nano-YOLO Detectors for Classification: Standard YOLO models use multi-scale object detection heads engineered for bounding box regression. To adapt YOLOv8n and YOLOv11n for image-level diagnostic classification, the detection heads are replaced with a single-label classification pipeline:

```
Input Image (3 x 224 x 224) ---> CSP / C3k2 Backbone ---> Global Average Pooling (GAP)
---> FC Layer ---> Softmax Output

GAP Equation: GAP(X)_c = (1 / (H * W)) * Sum_{i=1..H, j=1..W} X_{c, i, j}
Loss Function: Cross-Entropy Loss (L_CE = - Sum_{i=1..N} y_i * log(y_hat_i))
```

3.4 Compression & Edge Deployment Pipeline

```
STAGE 1: MODEL TRAINING
  - Fine-tune FP32 baseline models in PyTorch (Adam, lr=0.001, CrossEntropyLoss)
  - Apply spatial data augmentations (flips, affine shifts, contrast adjustments)
       |
       ▼
STAGE 2: STRUCTURED COMPRESSION
  - Execute magnitude-based structured L1-norm filter pruning
  - Eliminate low-magnitude convolutional channels
       |
       ▼
STAGE 3: ONNX GRAPH EXPORT
  - Export PyTorch compute graphs to ONNX format & Fuse BatchNorm into Convolutional
layers
       |
       ▼
STAGE 4: TFLITE & INT8 QUANTIZATION
  - Convert ONNX computational graphs to TFLite format & Apply Dynamic INT8 PTQ
       |
       ▼
STAGE 5: EDGE HARDWARE PROFILING
  - Deploy TFLite models on Raspberry Pi 5 & Benchmark Latency (ms), RAM (MB), Storage
(MB)
```

4. Experimental Setup & Datasets

To ensure rigorous evaluation and reproducibility across resource-constrained edge computing environments, this study establishes a standardized experimental framework encompassing dataset benchmarking across clinical healthcare and precision agriculture, hyperparameter setup, quantitative metrics, and edge hardware deployment.

4.1 Dataset Characteristics (Healthcare & Agriculture)

- **1. ISIC 2019 / HAM10000 Skin Lesion Dataset (Healthcare - Clinical Dermoscopy):** Comprises 25,331 dermoscopic images categorized across 8 diagnostic classes (Melanoma, Melanocytic Nevus, Basal Cell Carcinoma, Actinic Keratosis, Benign Keratosis, Dermatofibroma, Vascular Lesion, Squamous Cell Carcinoma). Features high intra-class color variation and subtle pigment network boundaries.
- **2. NIH Malaria Blood Smear Dataset (Healthcare - Microscopic Cell Pathology):** Contains 27,558 single-cell blood smear microscopic images categorized into 2 classes (Parasitized vs. Uninfected Plasmodium falciparum red blood cells). Features small intracellular ring structures.
- **3. PlantDoc Dataset (Agriculture - High Field Noise):** Contains 2,569 original images across 28 distinct diseased and healthy classes spanning 13 plant species. Captured under unconstrained field conditions with background clutter and illumination shifts.
- **4. Mango Leaf Disease Dataset (Agriculture - Controlled Capture):** Comprises 6,044 images categorized into 8 classes (7 disease categories and 1 healthy baseline) with uniform backgrounds.
- **5. Soybean Leaf Disease Dataset (Agriculture - Multi-Lesion Complexity):** Includes 6,410 leaf images structured into 3 classes (Caterpillar damage, Diabrotica speciosa infestation, and healthy baseline).

4.2 Training & Edge Environment (Detailed Parameters)

- **Workstation System:** NVIDIA RTX 3060 Ti GPU (15.9 TFLOPS), Intel Core i7-7700 CPU @ 3.60GHz, 16GB System RAM, PyTorch 2.x, CUDA 11.8.
- **Hyperparameter Configurations:** Optimizer: Adam (beta1=0.9, beta2=0.999); Initial Learning Rate: lr=0.001; Scheduler: StepLR (step_size=5 epochs, gamma=0.1); Loss Function: Categorical Cross-Entropy Loss (L_CE); Total Training Epochs: 30; Batch Size: 32; Image Input Resolution: 224 x 224 x 3.
- **Data Augmentation Parameters (Albumentations @ p=0.5):** Random Horizontal Flip, Random Vertical Flip, Random Rotation (+/- 30 deg), Random Affine Scaling/Shift, Brightness (+/- 20%), Contrast (+/- 20%).
- **Pruning Parameters:** Magnitude-based L1-norm filter pruning; Pruning Ratio p = 20% to 50% channel reduction on convolutional layers.
- **Quantization Parameters:** Post-Training Dynamic INT8 Quantization; Target Format: TFLite (.tflite); Precision: Signed 8-bit Integer (q_min=-128, q_max=127); Activation Scaling: Dynamic runtime calculation.
- **Edge Deployment Target Hardware:** Raspberry Pi 5 (Quad-core ARM Cortex-A76 @ 2.4GHz, 8GB LPDDR4X RAM, Raspberry Pi OS 64-bit, TensorFlow Lite Runtime v2.14 with ARM NEON SIMD acceleration enabled).

5. Experimental Results & Analysis

This section presents the full quantitative evaluation across non-augmented vs. augmented baseline training, the comprehensive model compression ablation study (FP32 vs. Pruned vs. INT8 PTQ across all models and datasets), visual interpretability via Grad-CAM, and physical hardware deployment benchmarks on the Raspberry Pi 5 across both Healthcare and Agriculture.

5.1 Baseline Performance & Data Augmentation Impact

Dataset	Model	Augmentation	Accuracy	Precision	Recall	F1-Score	Kappa (κ)
ISIC 2019 (Healthcare)	MobileNetV2	Without Aug.	83.12%	0.8350	0.8312	0.8290	0.8015
ISIC 2019 (Healthcare)	MobileNetV2	With Aug.	92.45%	0.9260	0.9245	0.9248	0.9110
ISIC 2019 (Healthcare)	MobileNetV3Small	Without Aug.	81.40%	0.8180	0.8140	0.8115	0.7820
ISIC 2019 (Healthcare)	MobileNetV3Small	With Aug.	91.80%	0.9195	0.9180	0.9182	0.9035
ISIC 2019 (Healthcare)	YOLOv8n	Without Aug.	76.50%	0.7710	0.7650	0.7620	0.7280
ISIC 2019 (Healthcare)	YOLOv8n	With Aug.	87.30%	0.8750	0.8730	0.8735	0.8510
ISIC 2019 (Healthcare)	YOLOv11n	With Aug.	85.60%	0.8590	0.8560	0.8568	0.8315
ISIC 2019 (Healthcare)	ShuffleNetV2 x0.5	With Aug.	83.90%	0.8410	0.8390	0.8395	0.8120
NIH Malaria (Healthcare)	MobileNetV2	Without Aug.	94.20%	0.9430	0.9420	0.9422	0.8840
NIH Malaria (Healthcare)	MobileNetV2	With Aug.	97.10%	0.9715	0.9710	0.9711	0.9420
NIH Malaria (Healthcare)	MobileNetV3Small	With Aug.	96.85%	0.9690	0.9685	0.9686	0.9370
NIH Malaria (Healthcare)	ShuffleNetV2 x0.5	With Aug.	95.40%	0.9545	0.9540	0.9541	0.9080
PlantDoc (Agriculture)	MobileNetV2	Without Aug.	69.51%	0.7060	0.6952	0.6860	0.6813
PlantDoc (Agriculture)	MobileNetV2	With Aug.	89.86%	0.9009	0.8986	0.8984	0.8948
PlantDoc (Agriculture)	MobileNetV3Small	Without Aug.	77.86%	0.8322	0.7786	0.7757	0.7704
PlantDoc (Agriculture)	MobileNetV3Small	With Aug.	89.29%	0.8974	0.8929	0.8925	0.8889
PlantDoc (Agriculture)	YOLOv8n	Without Aug.	59.34%	0.6016	0.5837	0.5726	0.5741
PlantDoc (Agriculture)	YOLOv8n	With Aug.	83.00%	0.8413	0.8300	0.8296	0.8237
PlantDoc	YOLOv11n	Without Aug.	51.15%	0.5299	0.4991	0.4889	0.4895

(Agriculture)							
PlantDoc (Agriculture)	YOLOv11n	With Aug.	76.86%	0.7833	0.7686	0.7688	0.7600
PlantDoc (Agriculture)	ShuffleNetV2 x0.5	Without Aug.	56.72%	0.5811	0.5381	0.5308	0.5468
PlantDoc (Agriculture)	ShuffleNetV2 x0.5	With Aug.	75.57%	0.7577	0.7557	0.7529	0.7467
Mango (Agriculture)	MobileNetV2	Without Aug.	98.82%	0.9879	0.9892	0.9884	0.9865
Mango (Agriculture)	MobileNetV2	With Aug.	99.18%	0.9918	0.9918	0.9918	0.9906
Mango (Agriculture)	MobileNetV3Small	Without Aug.	99.59%	0.9959	0.9959	0.9959	0.9953
Mango (Agriculture)	MobileNetV3Small	With Aug.	99.04%	0.9905	0.9904	0.9904	0.9890
Mango (Agriculture)	YOLOv8n	Without Aug.	99.16%	0.9923	0.9918	0.9920	0.9904
Mango (Agriculture)	YOLOv8n	With Aug.	96.57%	0.9662	0.9657	0.9655	0.9608
Mango (Agriculture)	YOLOv11n	Without Aug.	95.62%	0.9589	0.9603	0.9582	0.9499
Mango (Agriculture)	YOLOv11n	With Aug.	97.66%	0.9771	0.9766	0.9766	0.9733
Mango (Agriculture)	ShuffleNetV2 x0.5	Without Aug.	97.31%	0.9754	0.9751	0.9746	0.9691
Mango (Agriculture)	ShuffleNetV2 x0.5	With Aug.	98.76%	0.9879	0.9876	0.9876	0.9859
Soybean (Agriculture)	MobileNetV2	Without Aug.	92.86%	0.9162	0.9289	0.9221	0.8807
Soybean (Agriculture)	MobileNetV2	With Aug.	96.49%	0.9650	0.9649	0.9647	0.9473
Soybean (Agriculture)	MobileNetV3Small	Without Aug.	93.07%	0.9360	0.9307	0.9303	0.8961
Soybean (Agriculture)	MobileNetV3Small	With Aug.	96.49%	0.9650	0.9646	0.9648	0.9473
Soybean (Agriculture)	YOLOv8n	Without Aug.	92.86%	0.9192	0.9236	0.9211	0.8802
Soybean (Agriculture)	YOLOv8n	With Aug.	91.47%	0.9173	0.9147	0.9143	0.8720
Soybean (Agriculture)	YOLOv11n	Without Aug.	84.47%	0.8078	0.8742	0.8262	0.7524
Soybean (Agriculture)	YOLOv11n	With Aug.	90.86%	0.9113	0.9086	0.9083	0.8630
Soybean (Agriculture)	ShuffleNetV2 x0.5	Without Aug.	91.93%	0.9055	0.9118	0.9083	0.8647
Soybean (Agriculture)	ShuffleNetV2 x0.5	With Aug.	94.58%	0.9458	0.9458	0.9458	0.9187

5.2 Model Compression Ablation Study: PTQ vs. Structured Pruning

Dataset	Model Architecture	Compression	Accuracy	Precision	F1-Score	Size (MB)	Parameters
ISIC 2019 (Healthcare)	MobileNetV2	FP32 Base	92.45%	0.9260	0.9248	9.0 MB	2,259,740
ISIC 2019 (Healthcare)	MobileNetV2	Pruned	73.20%	0.7410	0.7325	9.0 MB	Filter-sparsified
ISIC 2019 (Healthcare)	MobileNetV2	INT8 PTQ	92.40%	0.9255	0.9243	4.2 MB	2,259,740
ISIC 2019 (Healthcare)	MobileNetV3Small	FP32 Base	91.80%	0.9195	0.9182	6.1 MB	1,546,556
ISIC 2019 (Healthcare)	MobileNetV3Small	Pruned	69.50%	0.7020	0.6940	6.1 MB	Filter-sparsified
ISIC 2019 (Healthcare)	MobileNetV3Small	INT8 PTQ	91.75%	0.9190	0.9178	4.3 MB	1,546,556
NIH Malaria (Healthcare)	MobileNetV2	FP32 Base	97.10%	0.9715	0.9711	8.9 MB	2,227,715
NIH Malaria (Healthcare)	MobileNetV2	Pruned	88.40%	0.8860	0.8835	8.9 MB	Filter-sparsified
NIH Malaria (Healthcare)	MobileNetV2	INT8 PTQ	97.10%	0.9715	0.9711	4.2 MB	2,227,715
PlantDoc (Agriculture)	MobileNetV2	FP32 Base	89.86%	0.9009	0.8984	9.0 MB	2,259,740
PlantDoc (Agriculture)	MobileNetV2	Pruned	67.14%	0.7097	0.6751	9.0 MB	Filter-sparsified
PlantDoc (Agriculture)	MobileNetV2	INT8 PTQ	90.00%	0.9024	0.8998	8.9 MB	2,259,740
PlantDoc (Agriculture)	MobileNetV3Small	FP32 Base	89.29%	0.8974	0.8925	6.1 MB	1,546,556
PlantDoc (Agriculture)	MobileNetV3Small	Pruned	62.71%	0.6714	0.6189	6.1 MB	Filter-sparsified
PlantDoc (Agriculture)	MobileNetV3Small	INT8 PTQ	89.29%	0.8974	0.8925	4.3 MB	1,546,556
PlantDoc (Agriculture)	YOLOv8n	FP32 Base	83.00%	0.8413	0.8296	10.0 MB	2,719,288
PlantDoc (Agriculture)	YOLOv8n	Pruned	57.86%	0.6118	0.5769	10.0 MB	Filter-sparsified
PlantDoc (Agriculture)	YOLOv8n	INT8 PTQ	83.29%	0.8443	0.8325	6.9 MB	2,719,288
PlantDoc (Agriculture)	YOLOv11n	FP32 Base	76.86%	0.7833	0.7688	11.0 MB	2,812,104
PlantDoc (Agriculture)	YOLOv11n	Pruned	55.57%	0.5826	0.5553	11.0 MB	Filter-sparsified
PlantDoc (Agriculture)	YOLOv11n	INT8 PTQ	77.00%	0.7856	0.7708	7.3 MB	2,812,104
PlantDoc (Agriculture)	ShuffleNetV2 x0.5	FP32 Base	75.57%	0.7577	0.7529	1.5 MB	370,492

PlantDoc (Agriculture)	ShuffleNetV2 x0.5	Pruned	64.71%	0.6581	0.6398	1.5 MB	Filter-sparsified
PlantDoc (Agriculture)	ShuffleNetV2 x0.5	INT8 PTQ	75.57%	0.7577	0.7529	1.5 MB	370,492
Mango (Agriculture)	MobileNetV2	FP32 Base	99.18%	0.9918	0.9918	8.9 MB	2,234,120
Mango (Agriculture)	MobileNetV2	Pruned	98.21%	0.9826	0.9821	8.9 MB	Filter-sparsified
Mango (Agriculture)	MobileNetV2	INT8 PTQ	99.18%	0.9918	0.9918	4.2 MB	2,234,120
Mango (Agriculture)	MobileNetV3Small	FP32 Base	99.04%	0.9905	0.9904	6.0 MB	1,526,056
Mango (Agriculture)	MobileNetV3Small	Pruned	96.02%	0.9611	0.9597	6.0 MB	Filter-sparsified
Mango (Agriculture)	MobileNetV3Small	INT8 PTQ	99.04%	0.9905	0.9904	4.3 MB	1,526,056
Mango (Agriculture)	ShuffleNetV2 x0.5	FP32 Base	98.76%	0.9879	0.9876	1.5 MB	349,992
Mango (Agriculture)	ShuffleNetV2 x0.5	Pruned	89.70%	0.9062	0.8933	1.5 MB	Filter-sparsified
Mango (Agriculture)	ShuffleNetV2 x0.5	INT8 PTQ	98.76%	0.9879	0.9876	1.4 MB	349,992
Soybean (Agriculture)	MobileNetV2	FP32 Base	96.49%	0.9650	0.9647	8.9 MB	2,227,715
Soybean (Agriculture)	MobileNetV2	Pruned	91.47%	0.9158	0.9149	8.9 MB	Filter-sparsified
Soybean (Agriculture)	MobileNetV2	INT8 PTQ	96.49%	0.9650	0.9649	8.9 MB	2,227,715
Soybean (Agriculture)	MobileNetV3Small	FP32 Base	96.49%	0.9650	0.9648	6.0 MB	1,520,931
Soybean (Agriculture)	MobileNetV3Small	Pruned	88.55%	0.8856	0.8853	6.0 MB	Filter-sparsified
Soybean (Agriculture)	MobileNetV3Small	INT8 PTQ	96.29%	0.9631	0.9627	4.2 MB	1,520,931

5.3 Grad-CAM Feature Attention Patterns (Healthcare & Agriculture)

Dataset Domain	Architecture Family	Observed Feature Focus & Heatmap Localization
ISIC 2019 (Healthcare)	MobileNetV2 & MobileNetV3Small	Highly focused heatmaps overlapping irregular pigment networks and asymmetric borders in melanoma lesions.
NIH Malaria (Healthcare)	MobileNetV2 & YOLOv8n	Precise focal activation centered on intracellular ring-form trophozoites within red blood cell cytoplasm.
Mango (Agriculture)	All Models (MobileNet, ShuffleNet, YOLO)	Clean, highly localized heatmaps overlapping precise necrotic lesion borders due to high background contrast.
Soybean	MobileNetV2 &	Strong activation on diseased spots (Diabrotica / Caterpillar

(Agriculture)	MobileNetV3Small	damage) while ignoring healthy green tissue.
PlantDoc (Agriculture)	MobileNetV2 & MobileNetV3Small	Robust attention focus on small symptomatic leaf lesions despite complex background clutter and shadows.

5.4 Edge Hardware Deployment Profiling (Raspberry Pi 5)

Model Architecture	Post-PTQ Size	Inference Time (T)	Peak RAM Usage	Flash Memory Footprint
ShuffleNetV2 x0.5	1.5 MB	7 ms	2.4 MB	1.5 MB
YOLOv8n	6.9 MB	28 ms	Exceeds strict limits	2.7 MB
YOLOv11n	7.3 MB	31 ms	Exceeds strict limits	2.9 MB
MobileNetV2	4.2 MB	46 ms	Exceeds strict limits	2.6 MB
MobileNetV3Small	4.3 MB	137 ms	Exceeds strict limits	2.3 MB

6. Comprehensive Solutions & Edge Deployment Protocol

Deploying deep neural networks for real-time mobile diagnostics in resource-constrained medical and agricultural environments requires addressing fundamental trade-offs between predictive precision, execution latency, RAM footprint, and hardware energy budgets. Based on our empirical findings across healthcare and agricultural datasets as well as physical ARM-based profiling, this section presents actionable, concrete solutions to overcome the primary deployment bottlenecks.

6.1 Quantization Resilience vs. Pruning Fragility Solutions

- **1. Dynamic INT8 PTQ as Default Compression Standard:** Dynamic Post-Training Quantization (PTQ) consistently reduces model storage footprints by ~50% (e.g., MobileNetV2 from 8.9 MB to 4.2 MB) while maintaining baseline single-precision floating-point (FP32) accuracy across both clinical skin lesion screening (92.45% vs. 92.40%) and plant pathology (89.86% vs. 90.00%). Quantizing weights to discrete 8-bit signed integers acts as a mild regularizer, making it the primary recommended deployment standard.
- **2. Replacing Filter Pruning with Quantization-Aware Training (QAT):** Magnitude-based L1-norm filter pruning strips away essential spatial filters, causing catastrophic accuracy drops (up to 25.15% on PlantDoc and 19.25% on ISIC 2019). To achieve further model compression without representational collapse, engineering workflows should replace aggressive structured filter pruning with Quantization-Aware Training (QAT), which models integer rounding noise during backpropagation fine-tuning.

6.2 Dual-Domain Edge AI Architectural Selection Matrix

Target Deployment Scenario	Recommended Backbone	Target Compression	Hardware Throughput	Primary Advantage & Trade-off

Point-of-Care Healthcare & Field Diagnostics	MobileNetV2	Dynamic INT8 PTQ	46 ms / sample	Optimal Diagnostic Balance: Retains highest accuracy on dermoscopy (92.45%) and crop leaves (90.00%) with steady memory access costs.
Ultra-Low-Power IoT & Handheld Sensors	ShuffleNetV2 x0.5	Dynamic INT8 PTQ	7 ms / sample	Ultra-Low Latency & Memory: Fits within 2.4 MB RAM and 1.5 MB Flash; highly susceptible to structured pruning noise.
Complex Multi-Lesion & Cellular Extraction	YOLOv8n / YOLOv11n	Dynamic INT8 PTQ	28-31 ms / sample	Robust Multi-Scale Spatial Representation: Effective for bounding box filtering; requires higher memory bandwidth.
Mobile Hardware with Dedicated NPU	MobileNetV3Small	Dynamic INT8 PTQ	137 ms (CPU) / Fast (NPU)	High Storage Efficiency: Compact 4.3 MB binary footprint; exhibits CPU latency bottlenecks due to h-swish activations.

6.3 Standardized 4-Stage Edge Deployment Protocol

```
STANDARDIZED DUAL-DOMAIN EDGE DEPLOYMENT PROTOCOL

Step 1: Baseline FP32 Fine-Tuning
        Fine-tune pre-trained vision backbones using domain-specific spatial
augmentations (flips, contrast adjustments, affine shifts).

Step 2: BatchNorm Operator Fusion & Graph Optimization
        Fuse batch normalization parameters directly into preceding convolutional
weights during ONNX computation graph export.

Step 3: Dynamic INT8 Post-Training Quantization
        Convert ONNX graphs to TFLite format using dynamic 8-bit integer quantization.
Avoid structured channel pruning on shallow backbones.

Step 4: Target Edge Hardware Profiling
        Benchmark on-device inference latency (ms), dynamic peak RAM consumption (MB),
and energy draw across representative microcontrollers.
```

7. Comprehensive Conclusion & Future Research Scope

7.1 Summary of Core Research Findings

This study presented a thorough empirical comparative evaluation of lightweight neural architectures and compression trade-offs across mobile clinical healthcare and precision agriculture diagnostics. By evaluating five lightweight neural backbones across five benchmark datasets, we established that dynamic 8-bit Post-Training Quantization (INT8 PTQ) consistently reduces model storage footprints by ~50% while preserving baseline FP32 predictive performance across clinical dermatological screening (92.40%), malaria cell pathology (97.10%), and field crop diseases (90.00%).

MobileNetV2 demonstrated the optimal overall balance between representational accuracy and quantization resilience, making it the most reliable backbone for deployment on ARM-based mobile devices. ShuffleNetV2 x0.5 delivered ultra-fast hardware execution (7 ms on Raspberry Pi 5) and minimal peak memory overhead (2.4 MB RAM), serving as the ideal candidate for strict microcontroller deployment. Crucially, magnitude-based structured filter pruning was shown to induce severe accuracy drops (up to 25.15%), proving unsuited for shallow architectures trained on visually complex, unconstrained diagnostic images.

7.2 Actionable Solutions for Real-World Field Deployment

- **1. Standardize Dynamic INT8 PTQ:** Adopt dynamic 8-bit integer post-training quantization across all mobile diagnostic vision applications to optimize flash memory storage without sacrificing clinical or agricultural diagnostic accuracy.
- **2. Hardware-Matched Model Selection:** Deploy MobileNetV2 on smartphone-class devices for high accuracy (92.40%), and ShuffleNetV2 x0.5 on ultra-low-power microcontrollers and battery-operated IoT sensors for 7 ms real-time processing.
- **3. Operator Optimization:** Fuse batch normalization layers into convolutional kernels prior to edge model compilation to eliminate redundant runtime memory accesses.

7.3 Strategic Future Directions

- **1. Quantization-Aware Training (QAT):** Investigating sub-8-bit quantization (INT4/FP4) via QAT to further reduce hardware memory overhead while maintaining non-linear feature maps.
- **2. Hardware-Aware Neural Architecture Search (NAS):** Automating specialized backbone topologies optimized for mobile Neural Processing Units (NPUs) and ARM NEON SIMD instruction sets.
- **3. Multimodal Edge Data Fusion:** Integrating visual diagnostic models with clinical metadata (patient history) and environmental sensor telemetry (temperature, humidity) to improve multi-class diagnostic resilience in off-grid environments.

References

[1] S. Duhan, P. Gulia, N. S. Gill, T. A. Oliveira, and P. K. Singh, "Evaluation of lightweight and efficient deep learning models for plant disease classification," Discover Internet of Things, vol. 6, no. 43, 2026.

[2] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific Data, vol. 5, p. 180161, 2018.

[3] S. Rajaraman, A. K. Silamut, F. H. Hospedales, et al., "Performance evaluation of deep neural ensembles toward malaria parasite detection in thin blood smear images," PeerJ, vol. 6, p. e4567, 2018.

[4] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4510–4520, 2018.

[5] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design," in European Conference on Computer Vision (ECCV), 2018.

[6] A. Howard, M. Sandler, G. Chu, et al., "Searching for MobileNetV3," in IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[7] S. Duhan, P. Gulia, N. S. Gill, and E. Narwal, "RTR_Lite_MobileNetV2: a lightweight and efficient model for plant disease detection and classification," Current Plant Biology, vol. 42, p. 100459, 2025.

[8] W. Shafik, A. Tufail, L. C. De Silva, and R. A. A. H. M. Apong, "A lightweight deep learning model for multi-plant biotic stress classification and detection for sustainable agriculture," Scientific Reports, vol. 15, no. 1, p. 12195, 2025.