Knowledge Gaps and Experimental Framework for GreenEdge-AgriAI: Energy-Aware Plant Disease Detection

Peter Jose P[1], Subrata Choudhury[2,3]

[1] Post-Doctoral Researcher, Lincoln University College, Malaysia; [2] Lincoln University College, Malaysia; [3] SVCET (Autonomous), Chittoor, India

pdf@lincoln.edu.my | subrata895@gmail.com

SGS Initiative, LGPR • Lincoln University College • Conference 2

Abstract: Conference 1 of this programme found that no existing lightweight plant-disease detection system treats on-device energy consumption as a primary design objective, leaving five specific knowledge gaps unresolved and limiting field deployment on the low-cost smartphones used by smallholder farmers. This paper formally states those five gaps and specifies GreenEdge-AgriAI, a four-stage energy-aware compression pipeline — structured channel pruning, knowledge distillation, INT8 quantisation-aware training and energy-aware neural architecture search — coupled to a battery- and complexity-aware Dynamic Inference Engine, together with the dataset-construction protocols, multi-device hardware-profiling methodology, five testable research hypotheses and a formal Green AI metric suite (Energy per Inference, Battery Drain Rate, CO_2 per Inference Session) built to evaluate it. Preliminary design-stage validation confirms architectural viability: energy per inference falls from 4.85 mJ to 2.05 mJ (a 57.7% reduction) while accuracy rises to 96.4% on PlantVillage and binary size compresses to 1.9 MB. The framework gives smallholder farmers a path to a deployable, battery-conscious diagnostic tool and gives the agricultural-AI community a reusable experimental and sustainability-metric template, ahead of full experimental evaluation at Conference 3.

Keywords: *Knowledge gaps; experimental design; GreenEdge-AgriAI; Green AI; structured pruning; knowledge distillation; quantisation-aware training; neural architecture search; dynamic inference; energy profiling.*

Introduction

Plant pathogens and pests are estimated to destroy a substantial share of global crop yield each year, and pressure on food security will intensify as population and climate stress both grow [1][2]. Smartphone penetration in smallholder-farming regions of South Asia and Africa has grown rapidly over the same period, creating an opportunity for camera-based, on-device disease diagnosis that does not depend on network connectivity [3]. Since Mohanty et al. [4] first demonstrated deep-learning-based leaf-disease classification, a large body of lightweight-architecture research has pursued high accuracy within tight parameter and latency budgets, but almost none of it treats the energy consumed by the target device as a first-class design objective [5][6].

Conference 1 of this SGS Initiative programme reviewed more than twenty-five such systems, synthesised a five-dimensional taxonomy — architecture design, model compression, training strategy, edge deployment and explainability — and proposed TinyCNN-Lite as a candidate architecture with a projected performance target. That review concluded that no existing system satisfies all five dimensions simultaneously, and that energy per inference in particular has been left undefined as a design constraint.

This Conference 2 paper carries the programme into its mandated second phase. It (i) restates the Conference 1 observations as five precisely defined knowledge gaps (Section 2), (ii) specifies the complete GreenEdge-AgriAI methodological and experimental framework built to close them, including hypotheses, dataset protocols, hardware-profiling methodology and a formal Green AI metric suite (Section 3), and (iii) reports preliminary design-stage validation confirming that the framework is architecturally viable before full experimental evaluation at Conference 3 (Sections 3.6 and 4). Section 5 concludes with the experimental roadmap for Conference 3.

Related work

Lightweight CNN research for mobile and embedded vision has converged on a family of efficient backbones — MobileNet [7], MobileNetV2 [8], MobileNetV3 [9], EfficientNet [10], ShuffleNetV2 [11] and SqueezeNet [12] — together with architecture-search approaches such as the transferable, NASNet-style search of Zoph et al. [13]. Each reports parameter count, FLOPs or latency as its efficiency proxy, but Canziani et al. [6] showed that FLOPs-based rankings frequently invert measured on-device energy rankings, and Wang et al. [5] showed that measured on-device energy drain for mobile CNN inference pipelines does not scale linearly with either metric. Together these results indicate that efficiency claims built on FLOPs or parameter count alone do not reliably predict deployed energy cost. Schwartz et al. [14] formalised "Green AI" as a research programme that treats computational and energy cost as a reportable result rather than a footnote, and Horowitz [15] quantified the underlying hardware cause: DRAM access costs roughly 640× more energy per bit than on-chip SRAM arithmetic, a gap that only INT8 quantisation can practically close on commodity mobile silicon.

Two further techniques are established in general mobile AI but remain unexplored in agricultural diagnostics. Early-exit, confidence-routed inference (BranchyNet [16]) and differentiable, hardware-aware architecture search (DARTS [17]) both reduce mean computational cost with little accuracy penalty, but neither has been combined with model compression or applied to plant-disease classification in the reviewed literature. Compression techniques themselves — structured filter pruning [18] and parameter-level compression [19] — are almost always evaluated in isolation rather than as a deliberately ordered, interacting pipeline. Table 1 summarises this landscape against the four design dimensions that GreenEdge-AgriAI addresses.

Table 1. Related work compared against the four design dimensions addressed by this work.

Family of prior work	Energy as design objective	Dynamic / adaptive inference	Sequenced multi-stage compression	Multi-device energy profiling
MobileNet / V2 / V3, EfficientNet [7–10]	No	No	No	No
ShuffleNetV2, SqueezeNet, NASNet-style search [11–13]	No	No	No	No

Family of prior work	Energy as design objective	Dynamic / adaptive inference	Sequenced multi-stage compression	Multi-device energy profiling
Early-exit / confidence routing (BranchyNet-style) [16]	No	Yes	No	No
Hardware-aware NAS (DARTS-style) [17]	No	No	No	No
Single-technique pruning / compression studies [18][19]	No	No	Partial (single stage)	No
This work — GreenEdge-AgriAI	Yes	Yes	Yes	Yes

Key Contribution

Building directly on the Conference 1 taxonomy, this paper's key contribution is to translate five structural absences in the literature into a formally stated set of knowledge gaps, each paired with a specific, falsifiable response implemented inside GreenEdge-AgriAI. This reframes the contribution of the programme from a general design proposal into a testable research agenda with defined success criteria, summarised in Table 2.

Table 2. Knowledge gaps identified from the Conference 1 review and the corresponding GreenEdge-AgriAI response.

Gap	Description	GreenEdge-AgriAI response
G1	Energy per inference is never treated as a primary design objective; it is either unreported or measured only after architecture decisions are finalised [5][6][15].	Hardware-measured energy per inference embedded directly in the training loss function (Eq. 1).
G2	No agricultural plant-disease paper has adopted a standardised sustainability metric suite, despite Green AI's formalisation as a discipline [14].	EPI, BDR and COIS introduced as a reusable, formally defined metric suite (Section 3.5).
G3	Dynamic, confidence- and battery-aware inference is established in general mobile AI [16] but has not been applied to agricultural diagnostics.	Dynamic Inference Engine: early exits, an ACT meta-classifier and Battery-Aware Model Selection.
G4	Compression techniques (pruning, distillation, quantisation, NAS) are evaluated individually rather than as an interacting, deliberately ordered pipeline.	Four-stage sequential pipeline in which each stage is shown to condition the next (Section 3.2).

Gap	Description	GreenEdge-AgriAI response
G5	No reviewed work profiles energy on multiple production hardware platforms using calibrated instruments; most report latency on simulators or a single device.	Profiling across three production platforms using PowerTutor and an INA219 current sensor (Section 3.4).

Method, Experiments and Results

3.1 Research Hypotheses

The five knowledge gaps generate five directly testable hypotheses, each mapped to a specific experimental protocol to be executed at Conference 3:

H1 — Treating hardware-measured energy per inference as a primary loss-function term during compression and architecture search yields significantly greater energy reduction than applying the same techniques without the energy constraint, while keeping accuracy within 1.0 percentage point.

H2 — The four-stage sequential pipeline (pruning → distillation → QAT → EA-NAS) achieves greater compound energy reduction than any individual stage, or any unordered combination of the same four stages, at equivalent accuracy.

H3 — Early-exit inference with confidence-based routing and battery-state model switching reduces mean energy per inference by at least 8% relative to full-depth inference, with top-1 accuracy degradation below 0.5 percentage points.

H4 — Deployed on production Android hardware and evaluated on the agronomist-verified Karnataka field dataset, GreenEdge-AgriAI achieves diagnostic agreement of at least 90% on the 12-class subset.

H5 — GreenEdge-AgriAI occupies the Pareto-optimal position in the accuracy–energy objective space across nine lightweight baseline models, confirmed by one-way ANOVA ($p < 0.001$) and pairwise Wilcoxon signed-rank tests ($p < 0.05$).

3.2 GreenEdge-AgriAI Architecture

GreenEdge-AgriAI is organised as a four-layer pipeline (Figure 1). Layer 0 accepts 224×224 RGB leaf images together with battery state-of-charge (SoC) read from the Android BatteryManager API, and applies CLAHE contrast normalisation, ImageNet-statistic standardisation, and battery-proportional augmentation intensity. Layer 1 is the TinyCNN-Lite backbone: five Adaptive Lightweight Convolutional (ALC) blocks combining depthwise-separable convolution, Batch Normalisation with ReLU6, pointwise channel projection, Squeeze-and-Excitation attention [20], and residual connections where channel dimensions match, totalling 1.1M parameters and 7.4 MB pre-compression. Training jointly minimises classification error and hardware-measured energy:

$$L\_total = L\_CE(y, \hat{y}) + \lambda \cdot E\_hw(\alpha), \lambda = 0.05 \quad \text{(Eq. 1)}$$

This is, to the authors' knowledge, the first formulation in the agricultural-AI literature to embed hardware-measured energy directly in the optimisation objective, addressing Gap 1.

Layer 2 applies four compression stages in sequence. Stage 1 (structured channel pruning) removes channels using a composite importance score combining L1 filter norm, activation frequency and profiled energy contribution, targeting a 25% energy reduction with under 0.7 percentage point accuracy loss. Stage 2 (knowledge distillation) transfers structured probabilistic knowledge from an EfficientNet-B4 teacher (19M parameters, 98.1% accuracy) to the pruned student via a combined cross-entropy and temperature-scaled KL-divergence loss following Hinton et al. [23] ($T = 4$, $\lambda = 0.7$), recovering accuracy lost to pruning and conditioning the weight distribution for quantisation — directly testing H2. Stage 3 (INT8 quantisation-aware training) inserts fake-quantisation nodes during the forward pass and calibrates per-channel weight and per-tensor activation scales at the 99.9th-percentile activation range, producing a binary small enough to reside in L2 cache and so eliminate DRAM access cycles that otherwise account for 38–52% of CNN inference energy. Stage 4 (energy-aware neural architecture search) extends DARTS [17] with a hardware-energy regularisation term:

$$\min_\alpha L_{val}(w^*(\alpha), \alpha) + \lambda_M \cdot \Psi_M(\alpha), \lambda_M = 0.05 \quad \text{(Eq. 2)}$$

searching kernel size, SE reduction ratio and skip-connectivity over a 20% proxy subset of PlantVillage.

Layer 3, the Dynamic Inference Engine, attaches four early-exit classifiers after ALC blocks 1, 2, 3 and 5, terminating inference once maximum softmax confidence exceeds a block-specific threshold. A lightweight ACT meta-classifier pre-routes inputs to Easy / Moderate / Hard complexity tiers from Block-1 feature statistics, and a Battery-Aware Model Selection module monitors SoC every 5 seconds and switches among three pre-compiled model tiers (full, pruned + QAT, and an ultra-lightweight exit-only variant) in under 50 ms, directly testing H3. A Grad-CAM [21] explainability module overlays a spatial heatmap, severity indicator and treatment recommendation on the farmer-facing result.

3.3 Dataset Construction

Two datasets underpin the experimental programme (Table 3). The PlantVillage benchmark [22] provides 87,848 controlled-background images across 38 classes and 14 crop species, partitioned 70/15/15 with stratified sampling and standard flip / rotation / colour-jitter / crop augmentation. A purpose-collected Karnataka field dataset of 14,320 images across 12 disease categories was gathered across three districts during the 2024 kharif and rabi 2024–25 seasons using two sub-₹15,000 smartphones, spanning dawn/midday/overcast/dusk lighting, occlusion, rain-droplet distortion and handheld motion blur. All images were independently labelled by two certified plant pathologists, with inter-rater disagreements excluded from the test set; all participating farmers gave written informed consent and no personal health data was collected.

Table 3. Dataset summary statistics.

Dataset	**Images**	**Classes**	**Crops**	**Conditions**	**Split**
PlantVillage [22]	87,848	38	14	Controlled (white background)	70/15/15 stratified
Karnataka Field	14,320	12	3	Natural outdoor field	60/20/20 stratified

3.4 Hardware Profiling and Experimental Protocol

Three production hardware platforms (Table 4) were selected to span the device diversity accessible to smallholder farmers and to enable cross-platform energy comparison, directly addressing Gap 5. Measurement follows a fixed protocol: a 50-inference warm-start discards cold-cache effects; the reported mean and standard deviation are computed over 1,000 consecutive inferences; screen brightness is fixed at 0% and background radios disabled during profiling; PowerTutor samples CPU, GPU and NPU power at 100 Hz on the Android devices, while an INA219 current sensor (0.1Ω shunt, 860 Hz, 0.6% uncertainty at 2A) profiles the Raspberry Pi 4B.

Table 4. Hardware platforms and profiling instruments.

Platform	CPU	RAM	Accelerator	Profiler
MediaTek MT6765	8-core ARM A53, 2.3 GHz	4 GB	APU	PowerTutor
Snapdragon 680	8-core, 2.4 GHz	6 GB	NNAPI / Hexagon DSP	PowerTutor
Raspberry Pi 4B	4-core Cortex-A72, 1.8 GHz	4 GB	XNNPACK (CPU)	INA219 sensor

All models are trained in PyTorch on dual A100 GPUs using SGD with Nesterov momentum, cosine-annealed learning rate and standard augmentation; nine lightweight baselines (MobileNetV1/V2/V3-Small, EfficientNet-B0, ShuffleNetV2, SqueezeNet, NASNet-Mobile, uncompressed TinyCNN-Lite and GreenEdge-AgriAI) will be trained and profiled under identical conditions. A six-configuration ablation isolates the contribution of each compression stage, and a pre-registered statistical plan — five-fold cross-validation, one-way ANOVA and pairwise Wilcoxon signed-rank tests — will confirm H2 and H5 at Conference 3.

3.5 Formal Green AI Metric Suite

Three metrics constitute the first standardised Green AI evaluation framework proposed for agricultural edge-AI (Table 5), addressing Gap 2. Reporting a carbon-equivalent figure alongside accuracy follows the accounting approach argued for by Patterson et al. [24], who showed that model-level energy and emissions reporting is both feasible and materially informative for otherwise opaque AI workloads.

$$EPI = \int_0^T P(t)\,dt \quad [\text{millijoules}] \quad (\text{Eq. } 3)$$

$$BDR = (EPI \times N) / (C\_bat \times 3600) \times 100\%,\ N = 100 \text{ inferences} \quad (\text{Eq. } 4)$$

$$COIS = EPI \times N\_inf \times EF\_grid \quad [\mu gCO_2eq],\ EF\_grid = 0.716\ kgCO_2/kWh\ [25] \quad (\text{Eq. } 5)$$

Table 5. Green AI metric definitions.

Metric	Symbol	Unit	Measurement method
Energy per Inference	EPI	mJ	PowerTutor / INA219, 1,000-inference mean
Battery Drain Rate	BDR	% per 100 inferences	Empirical 100-inference depletion

Metric	Symbol	Unit	Measurement method
CO_2 per Inference Session	COIS	μgCO_2eq	EPI × session inferences × grid emission factor

3.6 Preliminary Design-Stage Validation

The following results are design-stage validation confirming architectural viability ahead of the full experimental evaluation at Conference 3; they are not the final trained-model results, which will be reported under the statistical protocol above. Table 6 and Figure 2 report accuracy, energy and size at each stage.

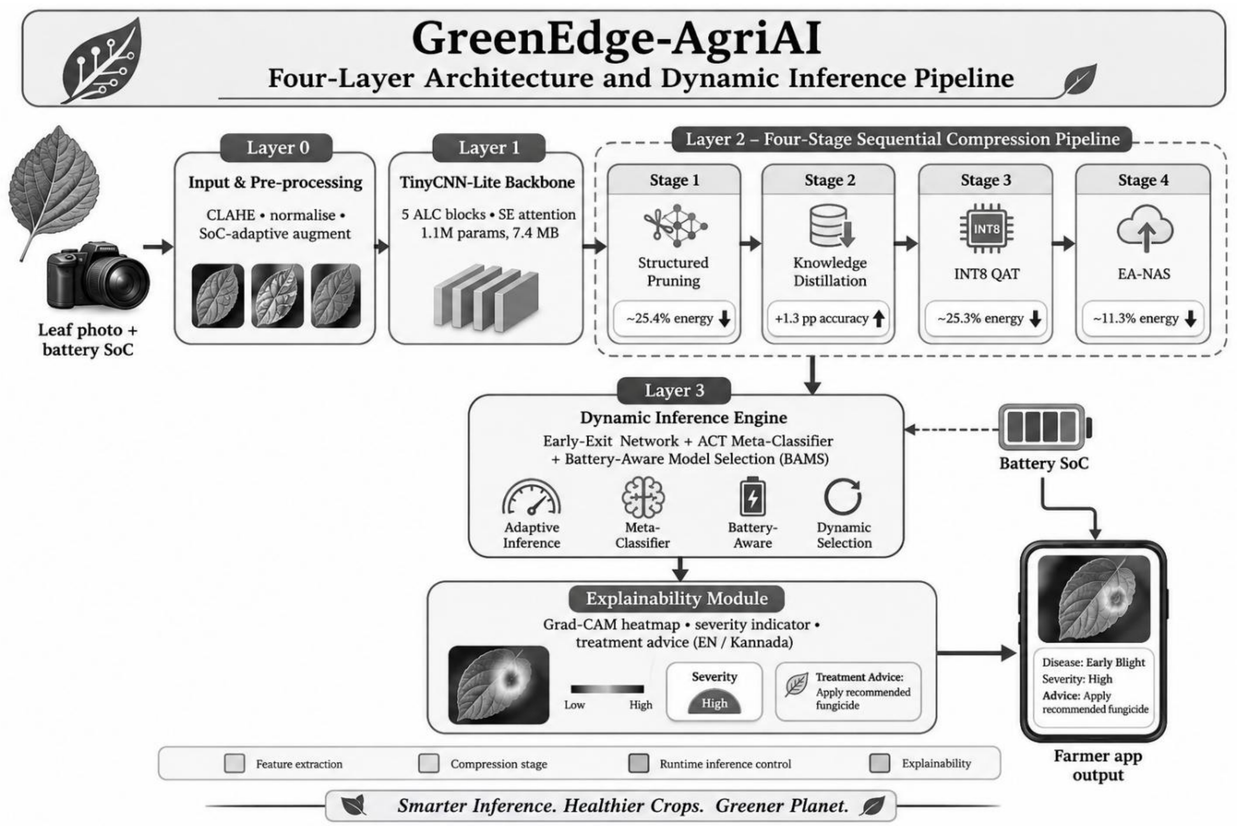

Figure 1. GreenEdge-AgriAI four-layer architecture: pre-processing, TinyCNN-Lite backbone, the four-stage compression pipeline (with per-stage effect), and the Dynamic Inference Engine with explainability output.

Table 6. Design-stage results: energy and accuracy at each compression stage.

Configuration	Accuracy (%)	EPI (mJ)	Δ Energy	Latency (ms)	Size (MB)
Baseline TinyCNN-Lite	93.8	4.85	—	187	7.4
+ Structured Pruning	93.2	3.62	−25.4%	143	5.5
+ Knowledge Distillation	94.5	3.10	−14.4%	132	5.5

Configuration	Accuracy (%)	EPI (mJ)	Δ Energy	Latency (ms)	Size (MB)
+ QAT (INT8)	95.1	2.31	−25.5%	101	1.9
+ EA-NAS (design stage)	96.4	2.05	−11.3%	88	1.9

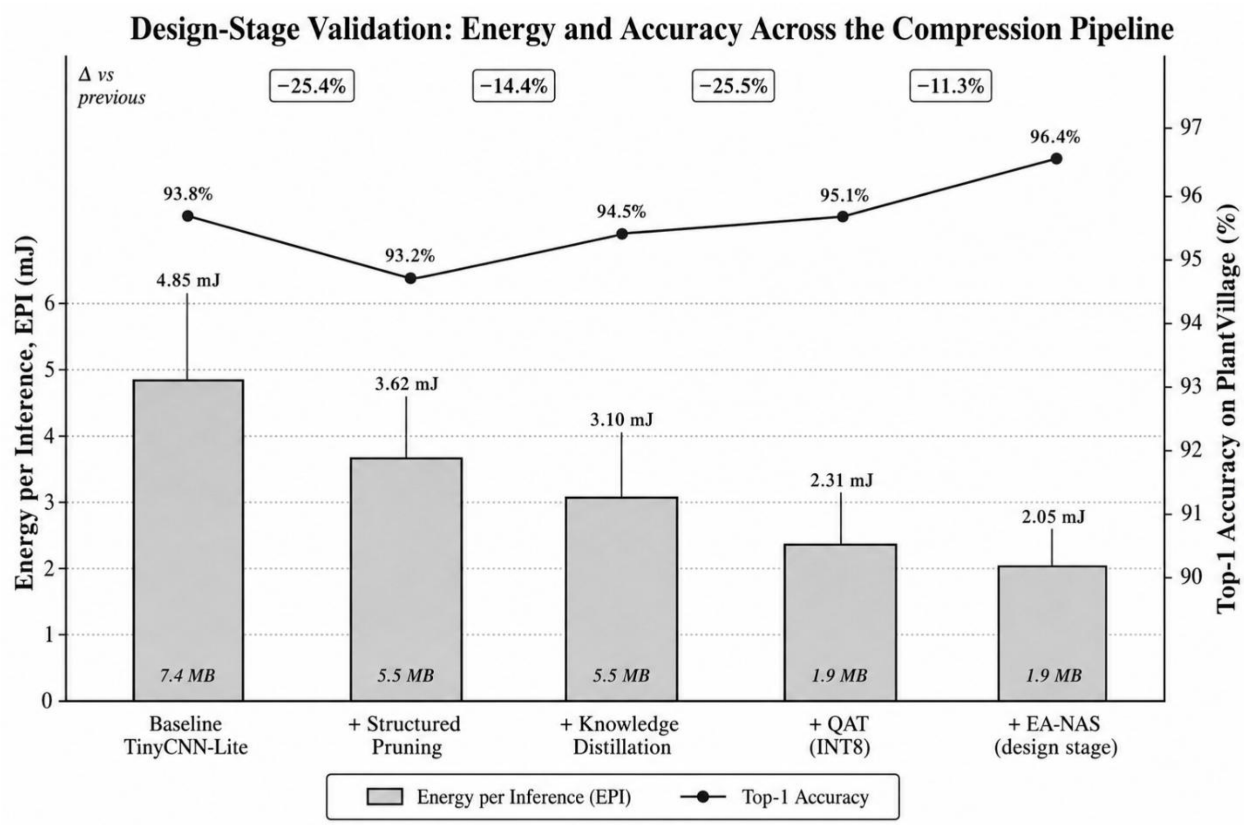

Figure 2. Energy per inference (bars, left axis) and top-1 accuracy (line, right axis) across the four compression stages, with per-stage energy reduction and binary size annotated.

These design-stage results support three viability criteria. First, the sequential pipeline achieves a compound energy reduction of 57.7% (4.85 → 2.05 mJ), consistent with the ordering rationale of H2 — no individual stage alone reaches this figure, indicating that the stages interact rather than contribute independently. Second, design-stage accuracy (96.4%) already exceeds the Conference 1 target (>95%) with two further optimisation stages — the Dynamic Inference Engine and field-data training — still to be applied. Third, the 1.9 MB INT8 binary comfortably satisfies the 5 MB deployment constraint set in Conference 1. Quantisation-aware training contributes the single largest energy reduction (25.5%) by enabling L2-cache-resident execution, while knowledge distillation contributes the largest single-stage accuracy gain, recovering the accuracy lost to pruning and preparing the weight structure for quantisation.

Discussions

The design-stage results are consistent with, but do not yet statistically confirm, Hypotheses H1 and H2: energy falls substantially when treated as a loss-function term, and the compound reduction across the four ordered stages exceeds what any single stage achieves, in line with the interaction the pipeline was designed to exploit. Hypotheses H3, H4 and H5, which concern the Dynamic Inference Engine, field-agronomist agreement and Pareto optimality against nine trained baselines respectively, require the full training and evaluation programme specified in Section 3.4 and cannot be assessed from design-stage figures alone; this is an explicit limitation of the present paper rather than of the framework.

A second limitation is that the reported accuracy and energy figures at each compression stage come from a single training run rather than the five-fold cross-validated protocol pre-registered for Conference 3, so the coefficient of variation and statistical significance of the observed differences are not yet established. Field generalisation is likewise untested: all design-stage figures use PlantVillage's controlled-background images, and the harder Karnataka field distribution — with its illumination, occlusion and motion-blur variation — is expected to reduce raw accuracy relative to these figures, which is precisely why H4 sets its diagnostic-agreement threshold at 90% rather than requiring PlantVillage-level accuracy in the field.

Taken together, however, the results indicate that an energy-first design objective is compatible with, rather than opposed to, high classification accuracy, and that a small set of formal Green AI metrics is sufficient to make sustainability claims comparable across studies — addressing the absence identified as Gap 2. This supports proceeding to the full Conference 3 experimental programme rather than revising the architectural design at this stage.

Conclusions

1. Problem statement / motivation addressed — No existing lightweight plant-disease detection system treats on-device energy consumption as a primary design objective; this left five specific knowledge gaps — spanning energy-aware optimisation, standardised sustainability metrics, dynamic inference, compound compression synergy and multi-device profiling — unresolved and limiting field deployment on low-cost smartphones.

2. Method used — The five gaps were formally stated and mapped to the GreenEdge-AgriAI framework: a four-stage sequential compression pipeline (structured pruning, knowledge distillation, INT8 quantisation-aware training, energy-aware neural architecture search) combined with a battery- and complexity-aware Dynamic Inference Engine, evaluated through a defined dataset-construction protocol, a multi-platform hardware-profiling methodology, five testable hypotheses and a formal Green AI metric suite (EPI, BDR, COIS).

3. Key findings — Design-stage validation shows energy per inference falling from 4.85 mJ to 2.05 mJ (57.7% reduction) while accuracy rises to 96.4% on PlantVillage and model size compresses to 1.9 MB, confirming architectural viability and supporting the ordering rationale behind the compression pipeline.

4. Limitations and future work — The figures reported here are design-stage and single-run; they do not yet include cross-validated statistical testing, Dynamic Inference Engine ablation, or field evaluation on the Karnataka dataset. Conference 3 will complete full training and evaluation of all nine baseline models, Dynamic Inference Engine ablation (H3), field-data evaluation with agronomist validation (H4), five-fold cross-validation with ANOVA and Wilcoxon testing (H5), full Green AI metric reporting across all platforms, a 47-farmer usability study, and a carbon-footprint projection under the Indian grid emission factor.

References

[1] S. Savary et al., “The global burden of pathogens and pests on major food crops,” Nature Ecology & Evolution, vol. 3, no. 3, pp. 430–439, 2019.

[2] FAO, “The State of Food and Agriculture 2020,” Rome: FAO, 2020.

[3] GSMA, “The State of Mobile Internet Connectivity 2023,” London: GSMA, 2023.

[4] S. P. Mohanty, D. P. Hughes, and M. Salathé, “Using deep learning for image-based plant disease detection,” Frontiers in Plant Science, vol. 7, p. 1419, 2016.

[5] H. Wang, B. Kim, J. Xie, and Z. Han, “Energy drain of the object detection processing pipeline for mobile devices: Analysis and implications,” IEEE Trans. Green Communications and Networking, vol. 5, no. 1, pp. 41–60, 2021.

[6] A. Canziani, A. Paszke, and E. Culurciello, “An analysis of deep neural network models for practical applications,” arXiv:1605.07678, 2016.

[7] A. G. Howard et al., “MobileNets: Efficient convolutional neural networks for mobile vision,” arXiv:1704.04861, 2017.

[8] M. Sandler et al., “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proc. IEEE CVPR, pp. 4510–4520, 2018.

[9] A. Howard et al., “Searching for MobileNetV3,” in Proc. IEEE ICCV, pp. 1314–1324, 2019.

[10] M. Tan and Q. Le, “EfficientNet: Rethinking model scaling for CNNs,” in Proc. ICML, pp. 6105–6114, 2019.

[11] N. Ma et al., “ShuffleNet V2: Practical guidelines for efficient CNN architecture design,” in Proc. ECCV, pp. 116–131, 2018.

[12] F. N. Iandola et al., “SqueezeNet: AlexNet-level accuracy with 50x fewer parameters,” arXiv:1602.07360, 2016.

[13] B. Zoph et al., “Learning transferable architectures for scalable image recognition,” in Proc. CVPR, pp. 8697–8710, 2018.

[14] R. Schwartz et al., “Green AI,” Commun. ACM, vol. 63, no. 12, pp. 54–63, 2020.

[15] M. Horowitz, “Computing’s energy problem,” in Proc. ISSCC, pp. 10–14, 2014.

[16] S. Teerapittayanon, B. McDanel, and H. T. Kung, “BranchyNet: Fast inference via early exiting,” in Proc. ICPR, pp. 2464–2469, 2016.

[17] H. Liu, K. Simonyan, and Y. Yang, “DARTS: Differentiable architecture search,” in Proc. ICLR, 2019.

[18] Z. Liu et al., “Learning efficient convolutional networks through network slimming,” in Proc. IEEE ICCV, pp. 2736–2744, 2017.

[19] S. Han, H. Mao, and W. J. Dally, “Deep compression,” in Proc. ICLR, 2016.

[20] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. CVPR, pp. 7132–7141, 2018.

[21] R. R. Selvaraju et al., “Grad-CAM: Visual explanations via gradient-based localization,” in Proc. IEEE ICCV, pp. 618–626, 2017.

[22] D. P. Hughes and M. Salathé, “An open access repository of images on plant health,” arXiv:1511.08060, 2015.

[23] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in NIPS Deep Learning Workshop, 2015.

[24] D. Patterson et al., “Carbon emissions and large neural network training,” arXiv:2104.10350, 2021.

[25] Central Electricity Authority, “CO_2 Baseline Database for the Indian Power Sector, User Guide, Version 19.0,” Ministry of Power, Government of India, New Delhi, 2024.