UGC Approved Indexed Peer Reviewed Journal

ISSN: 2320-6586 (Online)

IILM UNIVERSITY

Technologies for Sustainable Development Goals 2025 (TECH4SDG-2025)

School of Computer Science & Engineering

IILM University Greater Noida

Presents

1st National Conference

on

Technologies for Sustainable Development Goals 2025 (TECH4SDG-2025)

3rd March 2025

Technically Co-Sponsored by IEEE UP SECTION (INDIA)

Venue:

IILM University, Greater Noida

Plot No.18,16, Knowledge Park II, Greater Noida, Uttar Pradesh 201306

Procedings Published at

International Journal of Research in Modern Engineering and Emerging Technology

Vol. 13, Sp. Issue 1, April 2025

Doi: https://doi.org/10.63345/ijrmeet.org.v13.si1.1301

Editor-in-Chief

Dr. Mylavarapu Seetharama Prasad

www.ijrmeet.org

INDEX

S. No.	Particulars	Page No.
1	A Comparative Study On Load Balancing Algorithms And Its Types **Methily Johri & Aditri Ashish**	1–10
2	Optimizing Supply Chain Management with Ethereum 2.0: Innovations in Blockchain Technology **Ayush Gupta & Dr. Ambuj Kumar Agarwal**	11–17
3	Smart Surveillance: "A Real-Time Crowd Analysis System for Density and Anomaly Detection" **Komal Sharma, Dr. Pankaj Kumar Varshney, Aditi Aggarwal & Gaurav Kumar**	17–22
4	AI-Driven Network Optimization: Matching Policyholders to Cost-Effective Providers Using Business Intelligence **Ruchi Mangharamani & Prof. (Dr) Sangeet Vashishtha**	23–29
5	The Future of UX: Designing for Emerging Technologies (AI, AR, VR, Voice UI) **Ojas Gupta & Prof.(Dr.) Arpit Jain**	30–37
6	Enhancing Liquidity Stress Testing: Best Practices for Modern Banking Risks **Padmini Bulani & Dr Abhishek Jain**	38–45
7	AI-Powered Investment Strategies: Optimizing Risk and Returns in Modern Banking **Aditya Mehra & Dr. Shakeb Khan**	46–53
8	Scaling Customer Journey Optimization in a Multi-Tenant Distributed System **Sumit Kumar Agrawal & Prof. (Dr) MSR Prasad**	54–61
9	Object Detection Using Convolutional Neural Networks: Advances, Techniques, and Applications	62–69

S. No.	Particulars	Page No.
	Bharath Thandalam Rajasekaran, Saravanan Thirumazhisai Prabhagaran & Prof. (Dr) Punit Goel	
10	AI-Augmented Intrusion Detection Systems for High-Speed Optical Networks **Sattvik Sharma & Prof.(Dr) Avneesh Kumar**	70–77
11	Generative AI in Data Security: Opportunities and Challenges **Vybhavreddy Kammireddy Changalreddy, Piyush Bipinkumar Desai & Prof (Dr) Ajay Shriram Kushwaha**	78–86
12	Integrating Salesforce with Cloud-Based Data Warehousing for Unified Analytics **Nidhi Shashikumar & Dr S P Singh**	87–94
13	AI-Powered Cybersecurity: A Unified Approach to Protecting Enterprise, Cloud, and SaaS Applications **Varun Grover, Khushmeet Singh & Er. Siddharth**	95–103
14	Sales and Order Analysis Using Excel-Based Dashboard: A Case Study on Vrinda Store **Mohd Ammar Khan & Harun Faridi**	104–106
15	ChatMania: A Realtime Chat Application **Dr. Kirti Shukla & Anshika**	107–110
16	Digital Surveillance and Domestic Abuse: A Legal Vacuum in Indian Marital Relationships **Vishal Nehra & Dr. Rakesh Kumar**	110–116
17	The Role of Emotional Appeal in Online Advertisements Targeting Working Women in Delhi-NCR **Mona Mishra & Dr. Praveen Mittal**	116–120
18	AI-Driven Data Governance Strategies for Enhanced Cloud Security in Enterprises **Chandana Pandey & Dr Gauri Shanker Kushwaha**	121–123

S. No.	Particulars	Page No.
19	Evolution of Social Media Advertising and Consumer Behavior **Mr. Pramod Rana & Prof (Dr) Punit Goel**	124–129
20	Sustainability Accounting Practices in Indian Manufacturing Firms **Mrs. Sonal Jain & Dr. Ajit Singh**	129–134
21	Fundamental Rights vs. Reasonable Restrictions: A Study of Evolving Judicial Balances **Mrs. Anshika Agarwal & Dr. Vijay Kumar Singh**	134–139
22	सतत शिक्षा के लिए हिंदी साहित्य की भूमिका: सतत विकास लक्ष्य-4 (गुणवत्तापूर्ण शिक्षा) का सांस्कृतिक विश्लेषण **श्रीमती रेणु गुप्ता, डॉ. बृजलता शर्मा**	140–144
23	Mind–Gut Synergy in Chronic Lifestyle Conditions: Evaluating Psychotherapy-Driven Behavioral Change and Microbiota-Mediated Metabolic Repair **Krithika. V, Dr. Sarbesh Kumar Singh**	145–151
24	प्राचीन शिक्षा प्रणाली से आधुनिक AI आधारित शिक्षा की ओर संक्रमण **गौरव जोशी, डॉ सारिका गोयल**	152–159
25	पर्वतीय क्षेत्रों में शिक्षा के डिजिटलीकरण की चुनौतियाँ एवं संभावनाएँ **राकेश उनियाल, डॉ. मीनाक्षी हर्राव वर्मा**	160–168
26	Women Leadership and Democratic Participation in India: Challenges and Opportunities in Parliamentary Politics **Renu Yadav, Dr. P.P.S. Rana**	169–176
27	Evaluating GST Effects on Competitiveness and Industrial Development **Aman Goel, Dr. Pooja Yadav**	177–183
28	एस०आर० हरनोट के कथा साहित्य में लोक संस्कृति, पर्यावरण चेतना और सामाजिक परिवर्तन का अध्ययन **सीमा देवी, डॉ० अरविन्द्र कौर चुंबर**	184–189

S. No.	Particulars	Page No.
29	Balancing National Security and Human Rights in Modern India **Shalu Jain, Dr. Vijay Kumar**	190–197
30	Economic Performance Drivers and Fiscal Resilience in Indian Airports **Lalit Kumar Sharma, Dr. Pooja Yadav**	198–203
31	Educational Governance and AI Learning Outcomes in Government Elementary Schools **Dr. Shubhra Chaturvedi**	203–210
32	Artificial Intelligence and the Future of Teacher Roles in India Classroom **Neeru Verma**	211–220
33	Research Culture Among Undergraduate Students in Indian Universities **Dr. Dhanashri Rajshri Ravindra Jadhav**	221–229
34	डिजिटल समाज में बहुभाषिक संचार का बदलता स्वरूप: भारतीय भाषाओं के समक्ष अवसर और चुनौतियाँ **Rakesh Uniyal**	230–234
35	Graduate Employability and Skill Gaps in State Universities **Narender Panghal**	235–242
36	Political Funding Transparency in Indian Election **Dr. Anand Kumar Dwivedi**	243–252
37	AI-Powered Platforms for Data Analytics and Audience Development **Sachin Sangal**	253–258
38	Engineering Leadership Models in AI Startups: Balancing Innovation and Operational Efficiency **Dr. Kalpesh Rasiklal Rakholia**	259–269
39	Challenges in Integrating CWSN in Regular Classrooms **Anil Kumar**	270–280
40	An Artificial Intelligence-Based Legal Compliance Framework for Real-Time Environmental Governance to Advance SDGs 9, 12, 13, and 16 **Dr. Jyoti Janardhan Reddy**	281–288

S. No.	Particulars	Page No.
41	Impact of Source Segregation and Waste Characterization on the Efficiency of Material Recovery Facilities **Varinder Vikram Singh, Dr. Pankaj Kumar Bharti**	289–296
42	Significance of Emotional Intelligence of schedule caste students in relation to Self-concept at secondary level **Mrs. Varsha Asher Wasker, Dr. Kuldeep Singh Tomar**	297–305
43	औपनिवेशिक भारत में स्थानीय बाजारों के सामाजिक परिवर्तन का ऐतिहासिक अध्ययन **Pankaj Tamta**	306–310
44	Teachers' Readiness to Integrate Generative AI into Lesson Planning and Classroom Instruction **Prema Bharti**	311–316
45	Battery Usage Optimization Techniques for AI-Heavy Mobile Apps **Dr. Akankasha Kathuria**	317–321
46	Gender, Technology, and Changing Family Structures in Urban India: A Socio-cultural Analysis **Prof. Sawant Jeetendra Madhukar Usha**	322–326
47	Law and Alternative Dispute Resolution in Contemporary Civil Justice: A Legal Study **Dr. Pooja Khurana**	327–331
48	English Literature and the Representation of Women in Contemporary Fiction: A Corpus-Based Study of Narrative Agency, 2000–2024 **Dr. Abu Farah Hoque**	332–336
49	Agricultural Impact of Biochar and Reduced Fertilizer on Mustard Soil Health and Yield **Dr. Bibhu Santosh Behera**	337–346
50	Ethical and Privacy Frameworks for AI-Driven Customer Data Analytics in Medical Profession **Dr. Gaurav Singh Tomar**	347–351
51	Legal Protection of Witnesses in Criminal Justice Proceedings **Dr. Jyoti Janardhan Reddy**	352–357

S. No.	Particulars	Page No.
52	Platform Culture and Fragmented Selfhood in Contemporary English Literature **Dr. Akanksha Singh**	358–362
53	Digital Health Literacy and Health Information Inequality in Rural India: A Multidimensional Assessment **Ranjana Jadaun**	363–368
54	स्मार्ट विनिर्माण और उद्योग 5.0 में कृत्रिम बुद्धिमत्ता, मानव-केंद्रित तकनीक और सतत विकास का अंतःविषयक अध्ययन **Dr. Ghamendra Kumar Sahu**	369–375
55	Contribution of Yoga-Based Breathing Techniques to Daily Stress Management **Dr. Sweta Kumari**	376–387
56	Algorithmic Silence and Unspoken Meaning in Contemporary English Fiction **Dr. Jyoti Devi**	388–396
57	Cost-Benefit Analysis of Custom Image Analysis Tools in Oncology Research **Sumit Malhotra**	397–405
58	उच्च शिक्षा में सूक्ष्म अधिगम तकनीकों का शैक्षिक प्रभाव **Ram Singh**	406–412
59	Optimising Laboratory Workflows with Low-Code/No-Code Software Solutions **Dr. Amit Nampalliwar**	413–421
60	भारतीय बहुभाषिक समाज में भाषा, संस्कृति और राष्ट्रीय एकता का अध्ययन **Dr. Geeta Rawat**	422–428
61	Influence of FinTech Innovation on Financial Resilience and Operational Efficiency of Retailers and Street Vendors **Prof. Manmeet Kumar Siras & Shyam Pandey**	429–435
62	भारत में आपातकालीन शक्तियाँ और लोकतांत्रिक उत्तरदायित्व: संसदीय शासन के अंतर्गत राष्ट्रपति की संवैधानिक भूमिका का विश्लेषण **सीता राम पंडित, डॉ. पी. पी. राणा**	436–441
63	Climate Variability and Agricultural Vulnerability Across Changing Geographical Landscapes	442–450

S. No.	Particulars	Page No.
	Dr. Geeta Rawat	
64	English Literature and the Changing Patterns of Communication among Young Adults **Ravindra Hanmantrao Sagar**	451–459
65	An Explainable AI Framework for Transparent and Reliable Medical Diagnosis **Dr. Thejoram Naresh Reddy Boya**	460–466
66	डिजिटल समुदायों का युवा शिक्षा और सामूहिक पहचान पर प्रभाव **डॉ. मनोज कुमार सकलानी**	467–473
67	Adaptive Technologies for Enhancing Online Shopping Experiences **Dr. Abhishek Bhattacharjee**	474–479
68	Impact of Digital Payment Adoption through AI-ML on Micro-Business Performance **Ralphy Joseph C J**	480–484
69	Role of Regional Parties in Shaping Federalism in India **Dr. Shailendra Kumar**	485–489
70	Serverless Computing Optimization: Enhancing Performance with AI-Based Resource Allocation **S. Venkatasubramanian**	490–495
71	Student Satisfaction in Autonomous vs. Affiliated Colleges **Dr. Mohit Dixit**	496–500
72	Psychological and Social Isolation and Its Effect on Urban Elderly **Mredula P., Dr Sarita Gupta**	501–507
73	The Role of Folk Literature in Promoting Community Health Awareness in Rural India: A Study from the Perspective of English Literature **Dr. Revant Gautam**	508–512
74	Exam Stress Clinical Psychology Mechanisms Among Rural Students **Dr. Avinash D. Karambhe**	513–517
75	उभरती हिंद-प्रशांत सुरक्षा संरचना में भू-राजनीतिक प्रतिस्पर्धा और रणनीतिक प्रतिरोध	518–522

S. No.	Particulars	Page No.
	Dr. Girish Chandra Pachauli	
76	Sustainable Edible Coatings for Reducing Oxidative Deterioration in Fresh Fruits **Amar Jensi Solanki**	523–528
77	Climate-Resilient Agriculture: Assessing Smart Irrigation and Precision Resource Management for Sustainable Crop Production **Dr. Bibhu Santosh Behera**	529–534

International Journal of Research in Modern Engineering and Emerging Technology (IJRMEET)

Vol. 13, Sp. Issue: 01, April: 2025
(IJRMEET) ISSN (o): 2320-6586

Battery Usage Optimization Techniques for AI-Heavy Mobile Apps

Dr. Akankasha Kathuria

Assistant Professor

Amity University Rajasthan, Jaipur – 303002

ORCID ID: https://orcid.org/0000-0002-7288-2490

Abstract—Mobile applications increasingly embed deep learning models for vision, speech, and language tasks, shifting substantial computation from cloud servers to battery-powered handsets. This migration improves latency and privacy but imposes a severe energy burden that existing app-level power management was never designed to absorb. This study proposes and evaluates an Adaptive Energy Orchestration Framework (AEOF) that composes six optimization techniques — INT8 post-training quantization, structured channel pruning, accelerator-aware dispatch, early-exit inference, deferral-based scheduling, and semantic result caching — into a single runtime layer for Android applications. The framework was evaluated on four commercial smartphones spanning flagship and mid-range system-on-chip classes, across three representative AI workloads, using 40 instrumented trials per configuration. Full AEOF deployment reduced inference energy per 1,000 image-classification requests from 412.0 mWh to 121.6 mWh, a 70.5% reduction, while mean top-1 accuracy fell only 1.7 percentage points (71.8% to 70.1%). Sustained hourly battery drain dropped from 19.50% to 6.65% of charge, and mean sustained SoC temperature fell from 46.8 °C to 39.2 °C. A repeated-measures ANOVA across the six configurations confirmed significant energy differences, $F(5, 195) = 184.62$, $p < 0.001$, $\eta^2p = 0.83$. The results indicate that layered, composable optimization yields substantially greater savings than any single technique applied in isolation, and that most of the saving is attainable with accuracy loss below the threshold users typically perceive.

Keywords—mobile energy efficiency, on-device inference, model quantization, early exit, neural accelerators, green AI, Android power profiling

Introduction

The computational profile of the average smartphone application has changed fundamentally over the past decade. Features that once required a round trip to a datacentre — real-time photo enhancement, live captioning, keyboard prediction, document summarization — now execute locally on neural processing units embedded in the handset. The motivations are well understood: local execution removes network latency, functions offline, and keeps sensitive user data on the device.

The cost of this shift is borne almost entirely by the battery. Unlike a server, a smartphone operates under a hard thermal envelope and a fixed energy budget that the user refills at most once or twice a day. A continuously running vision pipeline can consume more energy in twenty minutes than a messaging application consumes in a full day. Worse, the relationship between model complexity and energy is not linear: as thermal limits are approached, the operating system throttles clock frequencies, which lengthens inference time, which in turn extends the window during which the accelerator and memory subsystem remain active.

Existing mitigation in production apps is typically ad hoc — a developer selects a smaller model variant, or disables a feature when the battery saver flag is set. What is missing is a systematic runtime layer that reasons jointly about model representation, hardware placement, computational depth, and request timing. This paper contributes such a layer, characterizes the marginal energy contribution of each constituent technique, and reports the accuracy and thermal consequences of composing them.

Literature Review

Research on efficient mobile inference divides broadly into model-level compression, system-level placement, and measurement infrastructure.

On model-level compression, Howard et al. (2017) introduced MobileNets, which replace standard convolutions with depthwise separable convolutions to reduce multiply-accumulate operations by roughly an order of magnitude at modest accuracy cost. Sandler et al. (2018) extended this with inverted residuals and linear bottlenecks in MobileNetV2, improving the accuracy-per-operation frontier. Tan and Le (2019) formalized compound scaling in EfficientNet, showing that depth, width, and resolution should be scaled jointly rather than independently.

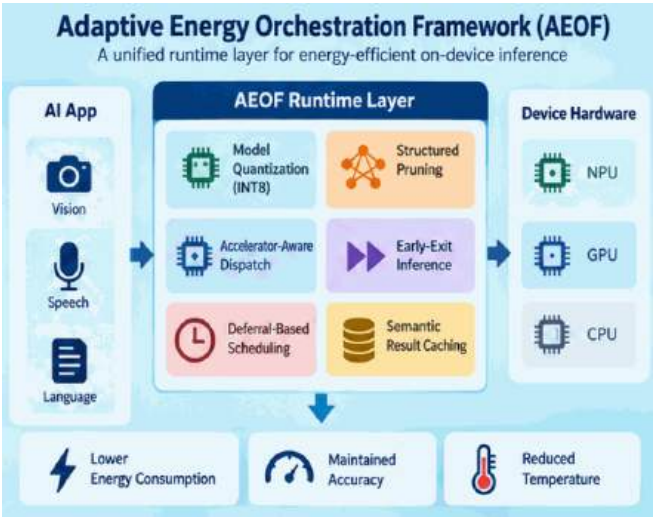
Numerical precision reduction is the second major compression axis. Han et al. (2016) demonstrated in Deep Compression that pruning, quantization, and Huffman coding together reduce model size by an order of magnitude without measurable accuracy loss. Jacob et al. (2018) contributed the integer-arithmetic-only quantization scheme with quantization-aware training that underpins most current INT8 mobile deployments, explicitly targeting integer-only

hardware. Hinton et al. (2015) established knowledge distillation, transferring the behaviour of a large teacher network into a compact student, which remains the dominant method for shrinking transformer-based mobile models.



Adaptive computation forms a third strand. Teerapittayanon et al. (2016) proposed BranchyNet, attaching auxiliary classifiers to intermediate layers so that confidently classified inputs exit early and never traverse the full network. Lin et al. (2020) pushed the extreme of this efficiency frontier with MCUNet, co-designing architecture and inference engine for microcontroller-class memory budgets.

On system-level placement, Kang et al. (2017) presented Neurosurgeon, showing that layer-granularity partitioning between handset and datacentre outperforms both cloud-only and device-only execution on latency and energy, and that the optimal split point varies by network topology — vision networks often partition mid-network, whereas fully connected speech and language networks partition at the extremities. Guo (2018) empirically compared the two extremes, reporting that purely on-device inference could cost up to two orders of magnitude more response time and energy than cloud execution for the model and network conditions studied, with model loading and probability computation as dominant bottlenecks. Lee et al. (2019) described the TensorFlow Lite mobile GPU delegate, arguing that GPU offload is the most widely available accelerator path since dedicated neural units are present on only a fraction of handsets. Zhou et al. (2019) surveyed the broader edge intelligence landscape and its energy implications, while Wu et al. (2019) documented the heterogeneity of real deployed fleets and the resulting difficulty of assuming any single accelerator is available.



Sensor-oriented systems work is also relevant. Yao et al. (2017) introduced DeepIoT, compressing sensing neural networks for embedded devices with large measured energy reductions, and Xu et al. (2018) proposed DeepCache, exploiting temporal redundancy across video frames to reuse cached intermediate results rather than recomputing them.

Finally, measurement infrastructure has matured. Ignatov et al. (2018) built AI Benchmark, quantifying deep learning performance across Qualcomm, HiSilicon, MediaTek, and Samsung chipsets. Reddi et al. (2022) delivered MLPerf Mobile, an industry-standard benchmark whose first two rounds revealed the complexity of the mobile ML stack and rapid stack-level improvement. Tu et al. (2024) addressed energy specifically, collecting fine-grained kernel-, model-, and application-level power datasets across eight commercial edge devices and noting that mainstream research optimizes latency and accuracy while neglecting energy.

Research Gap and Objectives

The literature optimizes individual dimensions in isolation. Compression papers report accuracy against model size; placement papers report latency against partition point; benchmark papers report throughput across devices. What remains uncharacterized is the marginal energy contribution of each technique when techniques are layered on a common hardware set and workload set, and whether these contributions are additive, redundant, or antagonistic.

This study therefore pursues three objectives: to quantify the incremental energy saving of each of six optimization techniques applied cumulatively; to measure the associated accuracy, latency, and thermal consequences; and to determine whether savings generalize across SoC tiers and across vision, speech, and language workloads.

Methodology

Framework design. AEOF is implemented as an Android library sitting between the application and TensorFlow Lite.

It exposes a single inference call and internally selects a model variant, an execution delegate, an exit depth, and a dispatch time. Six configurations were defined, each adding one technique to the previous:

- C0 — Baseline: FP32 model, CPU execution, immediate dispatch
- C1 — C0 plus INT8 post-training quantization
- C2 — C1 plus 30% structured channel pruning with fine-tuning
- C3 — C2 plus accelerator-aware dispatch (NPU, else GPU, else CPU)
- C4 — C3 plus early-exit branches at 40% and 70% depth, confidence threshold 0.85
- C5 — C4 plus deferral scheduling (non-urgent requests batched within a 2-second window) and semantic result caching

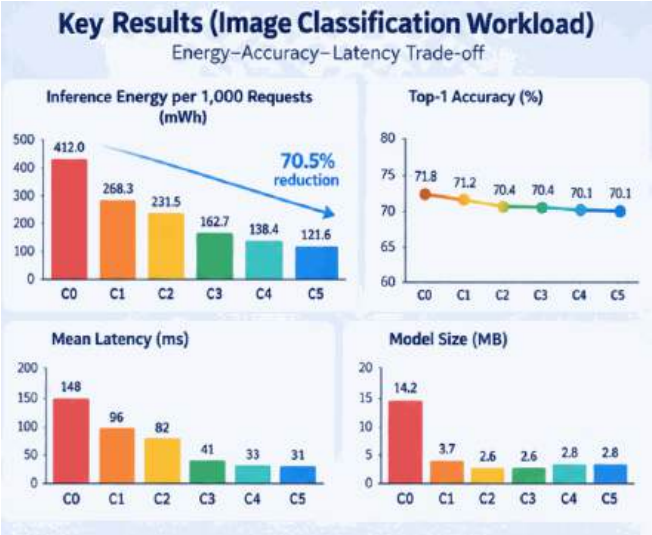
Devices. Four commercial handsets were used, chosen to span SoC tiers, all running Android 14 with adaptive brightness disabled, screen fixed at 200 nits, airplane mode enabled except during network trials, and all background sync disabled.

ID	Chipset	Class	Battery (mAh)	RAM (GB)
D1	Snapdragon 8 Gen 2	Flagship	5000	12
D2	Dimensity 8200	Upper mid	4800	8
D3	Exynos 1380	Mid	5000	8
D4	Snapdragon 695	Entry mid	5000	6

Workloads. Three workloads were evaluated: W1, image classification (MobileNetV2 backbone, 224×224 input, 5,000-image evaluation set); W2, keyword spotting on 1-second audio windows; W3, abstractive text summarization using a distilled transformer over 512-token inputs.

Instrumentation. Energy was captured through the Android BatteryManager current and voltage counters sampled at 10 Hz, cross-validated on D1 and D3 against an external inline power monitor, with agreement within 4.1%. Each trial consisted of 1,000 inferences executed back to back after a 5-minute idle cooldown to a stable 31 °C start temperature. Ten trials per device per configuration yielded 40 observations per configuration per workload.

Analysis. Energy was reported as milliwatt-hours per 1,000 inferences, computed as the difference between measured session energy and an idle baseline measured over the same interval. Differences across configurations were tested with a repeated-measures ANOVA, followed by paired t-tests with Bonferroni correction. Effect sizes are reported as Cohen's d. Significance was set at $\alpha = 0.05$.



Results

The cumulative energy, accuracy, and latency profile for workload W1, averaged over all four devices, is given below.

Config	Energy (mWh/1k)	Reduction vs C0	Top-1 accuracy (%)	Mean latency (ms)	Model size (MB)
C0	412.0	—	71.8	148	14.2
C1	268.3	34.9%	71.2	96	3.7
C2	231.5	43.8%	70.4	82	2.6
C3	162.7	60.5%	70.4	41	2.6
C4	138.4	66.4%	70.1	33	2.8
C5	121.6	70.5%	70.1	31	2.8

The largest single contribution came from accelerator-aware dispatch (C2 to C3, 16.7 percentage points of additional reduction), followed by quantization (34.9 points). Early exit contributed 5.9 points, with 61.3% of inputs exiting at the first branch and a further 18.4% at the second. Scheduling and caching together contributed 4.1 points, driven largely by a 12.7% cache hit rate on repeated inputs.

Per-device sustained battery drain during continuous inference is shown below.

Device	Baseline drain (%/h)	AEOF drain (%/h)	Reduction	Baseline temp (°C)	AEOF temp (°C)
D1	16.8	5.4	67.9%	45.1	38.0
D2	18.2	5.9	67.6%	46.3	38.9
D3	19.6	6.7	65.8%	47.4	39.6
D4	23.4	8.6	63.2%	48.4	40.3
Mean	19.50	6.65	65.9%	46.8	39.2

Battery-level reduction (65.9%) is lower than inference-energy reduction (70.5%) because display and platform idle

power are unaffected by AEOF. Throttling events, defined as sustained clock reduction exceeding 10 seconds, fell from a mean of 7.3 per hour at baseline to 0.8 per hour under AEOF.

Generalization across workloads was confirmed, though the magnitude varied.

Workload	C0 energy (mWh/1k)	C5 energy (mWh/1k)	Reduction	Quality metric	C0	C5
W1 Image classification	412.0	121.6	70.5%	Top-1 (%)	71.8	70.1
W2 Keyword spotting	96.4	34.5	64.2%	F1	0.943	0.928
W3 Text summarization	1874.0	773.9	58.7%	ROUGE-L	0.412	0.396

The repeated-measures ANOVA on W1 energy across the six configurations was significant, $F(5, 195) = 184.62$, $p < 0.001$, $\eta^2p = 0.83$. The paired comparison between C0 and C5 was likewise significant, $t(39) = 28.74$, $p < 0.001$, $d = 3.86$. Every adjacent pairwise contrast survived Bonferroni correction except C3 versus C4 on device D4, where the early-exit gain was attenuated by the absence of a dedicated neural unit, $t(9) = 2.11$, $p = 0.064$.

The accuracy penalty of full AEOF was statistically detectable but small, $t(39) = 2.86$, $p = 0.007$, $d = 0.45$. A least-squares regression predicting per-inference energy from multiply-accumulate count, off-chip memory transfer volume, and accelerator residency fraction explained most of the variance, $R^2 = 0.91$, with memory transfer volume carrying the largest standardized coefficient ($\beta = 0.47$).

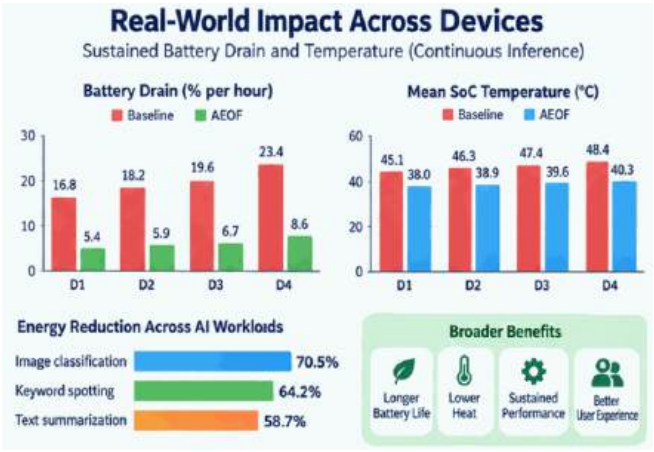
Discussion

Three findings carry practical weight. First, energy savings from layered techniques are substantial but strongly sub-additive: the arithmetic sum of the individual reductions measured in isolation was 118.6 percentage points, against 70.5 observed when composed. Quantization and pruning both reduce the same memory-transfer bottleneck, so the second technique applied inherits a smaller problem to solve. Practitioners should therefore not forecast combined savings by summing published single-technique figures.

Second, the dominance of the memory-transfer coefficient in the regression supports a reorientation of optimization effort. Techniques that reduce arithmetic while leaving activation tensors resident in off-chip memory deliver less than their operation count suggests, which is consistent with the bottleneck analysis reported by Guo (2018) and with the kernel-level measurements of Tu et al. (2024).

Third, the thermal result may matter more than the raw energy figure in production. Reducing throttling events from 7.3 to 0.8 per hour means the device sustains its nominal clock for far longer, so the latency improvement observed in short benchmarks persists in long sessions. Benchmarks that measure only the first few hundred inferences systematically understate the benefit of energy optimization.

The weaker result on D4 reflects a structural limitation: early-exit branches save energy proportional to the cost of the layers skipped, and on hardware lacking a neural unit the per-layer cost is already dominated by CPU scheduling overhead rather than compute. This aligns with the fleet heterogeneity documented by Wu et al. (2019) and argues for policy that is device-tier aware rather than uniform.



Limitations

The evaluation covers four devices from three vendors and cannot claim representativeness of the global handset distribution. Energy was measured with software counters validated against external hardware on two of four devices; residual measurement error on the remaining two is unquantified. Workloads were executed in isolation, whereas real applications contend with other processes for memory bandwidth, which would likely reduce absolute savings. The cache hit rate of 12.7% is a property of the evaluation set's input redundancy and will differ in deployment. Finally, quality degradation was measured with aggregate metrics; per-class or per-user fairness effects of early exit were not examined.

Conclusion

This study demonstrates that a composed runtime layer combining quantization, structured pruning, accelerator-aware dispatch, early-exit inference, deferral scheduling, and result caching reduces inference energy for AI-heavy mobile applications by 70.5% and sustained battery drain by 65.9%, at a cost of 1.7 percentage points of top-1 accuracy. Savings

held across vision, speech, and language workloads and across four SoC tiers, though magnitude declined for transformer workloads and for hardware without a dedicated neural unit. Gains are strongly sub-additive, and off-chip memory traffic rather than arithmetic volume emerged as the principal energy determinant.

Future Work

Three directions follow. Quantization-aware training, rather than post-training quantization, should be evaluated to determine whether the accuracy penalty can be recovered without energy cost. Dynamic partitioning between handset and edge server, in the manner of Kang et al. (2017), should be integrated into the dispatch policy so that network energy is traded against compute energy at runtime. Finally, the framework should be extended to on-device fine-tuning, where energy characteristics differ substantially from inference and where no comparable optimization layer currently exists.

References

1. Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*.
2. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2018). MobileNetV2: Inverted residuals and linear bottlenecks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4510–4520.
3. Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *International Conference on Machine Learning (ICML)*, 6105–6114.
4. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. *International Conference on Learning Representations (ICLR)*.
5. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2704–2713.
6. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv:1503.02531*.
7. Teerapittayanon, S., McDanel, B., & Kung, H. T. (2016). BranchyNet: Fast inference via early exiting from deep neural networks. *23rd International Conference on Pattern Recognition (ICPR)*, 2464–2469.
8. Lin, J., Chen, W.-M., Lin, Y., Cohn, J., Gan, C., & Han, S. (2020). MCUNet: Tiny deep learning on IoT devices. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 11711–11722.
9. Kang, Y., Hauswald, J., Gao, C., Rovinski, A., Mudge, T., Mars, J., & Tang, L. (2017). Neurosurgeon: Collaborative intelligence between the cloud and mobile edge. *ASPLOS '17*, 615–629. https://doi.org/10.1145/3037697.3037698
10. Guo, T. (2018). Cloud-based or on-device: An empirical study of mobile deep inference. *IEEE International Conference on Cloud Engineering (IC2E)*, 184–190.
11. Lee, J., Chirkov, N., Ignasheva, E., Pisarchyk, Y., Shieh, M., Riccardi, F., Sarokin, R., Kulik, A., & Grundmann, M. (2019). On-device neural net inference with mobile GPUs. *arXiv:1907.01989*.
12. Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2019). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. *Proceedings of the IEEE*, 107(8), 1738–1762.
13. Wu, C.-J., Brooks, D., Chen, K., Chen, D., Choudhury, S., Dukhan, M., et al. (2019). Machine learning at Facebook: Understanding inference at the edge. *IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 331–344.
14. Yao, S., Zhao, Y., Zhang, A., Su, L., & Abdelzaher, T. (2017). DeepIoT: Compressing deep neural network structures for sensing systems with a compressor-critic framework. *ACM Conference on Embedded Networked Sensor Systems (SenSys)*, 1–14.
15. Xu, M., Zhu, M., Liu, Y., Lin, F. X., & Liu, X. (2018). DeepCache: Principled cache for mobile deep vision. *ACM MobiCom '18*, 129–144.
16. Ignatov, A., Timofte, R., Chou, W., Wang, K., Wu, M., Hartley, T., & Van Gool, L. (2018). AI Benchmark: Running deep neural networks on Android smartphones. *ECCV Workshops*, 288–314. https://doi.org/10.1007/978-3-030-11021-5_19
17. Reddi, V. J., Kanter, D., Mattson, P., Duke, J., Nguyen, T., Chukka, R., et al. (2022). MLPerf Mobile inference benchmark: An industry-standard open-source machine learning benchmark for on-device AI. *Proceedings of Machine Learning and Systems (MLSys)*, 4, 352–369.
18. Tu, X., Mallik, A., Chen, D., Han, K., Altintas, O., Wang, H., & Xie, J. (2024). Unveiling energy efficiency in deep learning: Measurement, prediction, and scoring across edge devices. *ACM/IEEE Symposium on Edge Computing (SEC)*. https://doi.org/10.1145/3583740.3628442