Multidisciplinary Journal of Academic Publications
ISSN (Online): 3107-538X
Vol. 01, Issue 02, November – December 2025

OPTIMIZED IMAGE PROCESSING TECHNIQUES FOR ENERGY-EFFICIENT AND SECURE EDGE AI SYSTEMS

Ankita Sonwani[1], Dr. Ghanshyam Sahu[2]

Research Scholar, Department of Computer Science & Engineering, Bharti Vishwavidyalaya, Durg[1]

Assistant Professor, Department of Computer Science & Engineering, Bharti Vishwavidyalaya, Durg[2]

ABSTRACT

The rapid proliferation of edge Artificial Intelligence (AI) systems has accelerated the deployment of image processing pipelines in resource-constrained environments such as Internet of Things (IoT) nodes, smart cameras, embedded processors, and autonomous systems. These deployments present two intertwined and critical challenges: ensuring energy efficiency under limited computational budgets and guaranteeing privacy preservation when sensitive visual data is processed at or near the point of capture. This review paper presents a comprehensive meta-analysis of prior research addressing these dual imperatives in edge AI image processing. Through a systematic survey of literature published between 2015 and 2024, this work synthesises findings from over one hundred studies spanning model compression, hardware-software co-design, federated learning, differential privacy, homomorphic encryption, and on-device inference optimisation. The review identifies dominant trends including the growing adoption of lightweight convolutional neural networks (CNNs), model quantisation, neural architecture search (NAS), secure multi-party computation, and privacy-by-design frameworks for vision tasks. Critical gaps are identified, including the absence of unified benchmarks that simultaneously evaluate energy cost and privacy loss, insufficient attention to adversarial robustness under energy constraints, and the limited applicability of current privacy mechanisms at sub-milliwatt inference regimes. The paper further discusses open research challenges and proposes a consolidated taxonomy to guide future work at the intersection of green edge AI and trustworthy visual computing.

Keywords: *Edge AI[1], Energy-Efficient Image Processing[2], Privacy-Preserving Computing[3], Federated Learning[4], Model Quantisation[5], Differential Privacy[6], IoT Vision Systems[7]*

1. INTRODUCTION

Artificial Intelligence systems capable of interpreting visual information are increasingly deployed outside centralised data centres, operating instead on battery-powered sensors, embedded microcontrollers, and field-programmable gate arrays located at the periphery of communication networks. This migration from cloud to edge has been driven by latency requirements that prohibit round-trip data transmission, bandwidth limitations

in wireless environments, and the sheer volume of image data generated by modern sensor networks. A surveillance camera network producing hundreds of terabytes of footage per day cannot practically relay all captured frames to the cloud for analysis; instead, inference must occur locally. Similarly, a wearable medical device performing retinal screening in a rural setting must function without persistent Internet connectivity. These scenarios define the operating context for edge AI image processing systems, in which computational resources measured in milliwatts and megabytes coexist with demands for real-time, accurate visual analysis.

Two fundamental tensions arise in this context. First, the computational cost of state-of-the-art deep learning models for image processing far exceeds what edge hardware can sustain in an energy-efficient manner. Models such as ResNet-152, EfficientNet-B7, or Vision Transformers deliver high accuracy on benchmark datasets but consume multiple gigaflops of arithmetic per inference, translating into energy footprints incompatible with devices drawing milliwatts from coin cells or solar harvesting. Second, image data is inherently sensitive. Visual information captured in domestic, medical, financial, and public environments is subject to privacy regulations including the General Data Protection Regulation (GDPR), the Health Insurance Portability and Accountability Act (HIPAA), and sector-specific legislation in India under the Digital Personal Data Protection Act, 2023. Processing image data on-device appears to offer a natural privacy advantage over cloud processing, but edge devices are themselves vulnerable to physical tampering, side-channel attacks, and model inversion attacks that can reconstruct private information from model parameters or intermediate activations.

The intersection of energy efficiency and privacy preservation in edge AI image processing has attracted growing scholarly attention, yet no unified review has synthesised progress across both dimensions simultaneously. Prior surveys have addressed either green AI or privacy-preserving machine learning in isolation, leaving practitioners without consolidated guidance for designing systems that satisfy both constraints. This paper addresses that gap through a systematic meta-analysis of research published over the past decade. The subsequent subsections contextualise the scope and significance of this review.

1.1 MOTIVATION AND SCOPE

The motivation for this review arises from the convergence of three concurrent technological trends. The first is the explosive growth in edge computing infrastructure, driven by 5G rollout, smart city deployments, industrial IoT, and consumer wearables. The second is the increasing stringency of global data protection law, which imposes legal liability on organisations that process personal image data without adequate safeguards. The third is the maturation of a research ecosystem encompassing model compression, federated learning, secure aggregation, and neuromorphic hardware, each offering partial solutions to the energy-privacy challenge. By situating these trends together, this review aims to provide a panoramic understanding of where the field stands and where it must progress. The scope is restricted to image and video frame processing tasks, encompassing classification, detection, segmentation, and feature extraction, as executed by neural network models deployed

on edge hardware including ARM Cortex-M microcontrollers, NVIDIA Jetson modules, Google Coral Edge TPUs, and custom VLSI accelerators.

1.2 CHALLENGES IN ENERGY-PRIVACY CO-OPTIMISATION

Energy efficiency and privacy preservation are not naturally aligned objectives in edge AI system design. Techniques that reduce energy consumption frequently introduce new privacy risks, while strong privacy mechanisms often impose substantial computational overhead. Model quantisation, for example, reduces bit-width representations of weights and activations, cutting memory and arithmetic energy by factors of four to sixteen, but quantised models may be more susceptible to membership inference attacks owing to reduced representational capacity. Conversely, differential privacy mechanisms inject calibrated statistical noise into gradients or outputs to bound information leakage, but noise addition requires additional computation and degrades model accuracy, indirectly increasing inference iterations or model size to compensate. Hardware-level techniques such as in-sensor computing and analogue inference offer extreme energy efficiency but currently lack the programmability needed to implement cryptographic privacy primitives. This tension implies that co-optimisation requires joint modelling of energy cost and privacy loss as coupled design variables rather than independent objectives.

1.3 OBJECTIVES AND ORGANISATION OF THE PAPER

The objectives of this review are fourfold. First, to systematically survey and categories existing literature on energy-efficient and privacy-preserving image processing for edge AI, establishing a coherent taxonomy of approaches. Second, to meta analyse quantitative results reported across studies, identifying performance envelopes and empirical trends in energy savings and privacy budgets. Third, to critically evaluate the methodological rig our of reviewed studies, identifying sources of bias, incompatible evaluation protocols, and reproducibility limitations. Fourth, to delineate open challenges and propose research directions that can advance the field toward practically deployable systems satisfying both constraints simultaneously. The remainder of the paper is organised as follows: Section 2 presents the systematic literature survey; Section 3 describes the meta-analysis methodology; Section 4 provides a critical analysis of reviewed work; Section 5 discusses implications and future directions; and Section 6 concludes the review.

2. LITERATURE SURVEY

The body of research addressing energy efficiency and privacy preservation in edge AI image processing spans hardware design, algorithm development, systems architecture, and cryptographic theory. This section organises the surveyed literature into thematic clusters corresponding to the principal technical approaches identified across reviewed publications.

Model compression techniques constitute the most extensively studied class of energy reduction methods. Howard et al. [1] introduced MobileNets, demonstrating that depthwise separable convolutions could reduce computational cost by approximately an order of magnitude relative to standard convolutional networks with tolerable accuracy loss on ImageNet classification. Subsequent iterations, MobileNetV2 [2] and MobileNetV3 [3], refined this approach through inverted residual structures and hardware-aware neural architecture search, achieving inference times below ten milliseconds on Cortex-A55 processors at sub-watt power levels. Sandler et al. [2] demonstrated that the bottleneck design of inverted residuals significantly reduced activation memory, a critical constraint on devices with limited SRAM. Zhang et al. [4] proposed ShuffleNet, exploiting channel shuffle operations to equalise inter-group information flow in grouped convolutions, yielding competitive accuracy at forty megaflops per inference. These architectural innovations established that deep networks could be designed from the ground up for edge deployment rather than compressed post-hoc.

Post-training and quantisation-aware training methods have complemented architectural efficiency. Jacob et al. [5] formalised an integer arithmetic framework for quantising weights and activations to eight-bit representations, enabling deployment on processors without floating-point units. Subsequent work by Rastegari et al. [6] extended quantisation to one-bit representations in XNOR-Net, replacing multiply-accumulate operations with binary XNOR and popcount operations, reducing arithmetic energy by up to fifty-eight times at the cost of significant accuracy degradation on complex tasks. Liu et al. [7] demonstrated that mixed-precision quantisation, assigning different bit-widths to different layers based on sensitivity analysis, could recover most accuracy lost by aggressive quantisation while retaining substantial energy savings. Hubara et al. [8] showed that quantised neural networks could achieve within two percent of full-precision accuracy on CIFAR-10 image classification, validating the viability of sub-byte inference for edge image processing.

Knowledge distillation, originally proposed by Hinton et al. [9] as a model compression framework, has been widely applied to edge image processing. In this paradigm, a large teacher network trains a compact student network by supervising its soft output distributions, transferring representational knowledge without requiring the student to match the teacher's architecture. Romero et al. [10] extended this to intermediate feature matching through FitNets, showing that thin but deep student networks could surpass the performance of shallower compressed models. Mishra and Marr [11] combined distillation with quantisation, demonstrating that distilled quantised networks could achieve accuracy within one percent of floating-point baselines on ImageNet subsets while consuming under fifty millijoules per inference on an ARM processor.

Neural architecture search has emerged as a systematic approach to discovering energy-efficient models without manual design. Cai et al. [12] proposed ProxylessNAS, which directly searches on the target hardware platform, incorporating latency measurements into the search objective. Tan and Le [13] introduced EfficientNet through a compound scaling methodology discovered via NAS, achieving Pareto-optimal accuracy-efficiency trade-offs across a family of models. While EfficientNet was originally designed for server deployment, scaled-down variants have been adapted for edge use. Wu et al. [14] proposed FBNet, which uses differentiable architecture

search to optimise for inference latency on specific mobile chipsets, reducing latency by up to 1.5 times compared to MobileNetV2 at equivalent accuracy.

Hardware-software co-design approaches target the interaction between inference algorithms and silicon implementation. Chen et al. [15] described the Eyeriss accelerator, which minimises data movement energy by maximising data reuse through a row-stationary dataflow, demonstrating that memory access energy dominates arithmetic in convolutional workloads. Jouppi et al. [16] reported the Tensor Processing Unit architecture, whose systolic array organisation sustains high operational intensity during matrix multiplication, relevant to large-batch inference. For edge-specific constraints, Sze et al. [17] provided a comprehensive analysis of hardware efficiency in deep learning, establishing that dataflow architecture, memory hierarchy, and sparsity exploitation jointly determine system-level energy and latency. Sim et al. [18] proposed a reconfigurable accelerator supporting multiple precision modes, enabling dynamic energy scaling based on input difficulty.

In-sensor and near-sensor computing represents an emerging paradigm for extreme energy efficiency. Choi et al. [19] demonstrated image classification using a CMOS image sensor with integrated convolutional processing in the analogue domain, eliminating the energy cost of analogue-to-digital conversion and data transmission from the sensor. This approach inherently limits the data available for reconstruction by external parties, offering an intrinsic privacy benefit. Seo et al. [20] extended this concept to feature extraction at the pixel level using memristive crossbar arrays, reporting inference energy below one microjoule for simple classification tasks on thirty-two by thirty-two pixel images.

Privacy-preserving image processing research has progressed along two principal axes: input data protection and model parameter protection. Federated learning, introduced by McMahan et al. [21], enables distributed model training without centralising raw data. In the context of edge image processing, federated learning allows each edge node to train on its locally captured images and contribute only gradient updates to a central aggregation server. Li et al. [22] surveyed federated learning challenges in heterogeneous edge environments, noting that non-independently and identically distributed data distributions, communication constraints, and device dropout introduce significant challenges not present in centralised training. Geyer et al. [23] proposed client-level differential privacy for federated learning, adding Gaussian noise to clipped gradient updates before transmission, providing formal privacy guarantees at the cost of increased communication rounds and reduced model accuracy.

Differential privacy as a formal privacy framework has been applied to image classification and feature extraction. Abadi et al. [24] developed DP-SGD, a differentially private stochastic gradient descent algorithm that clips per-sample gradients and injects calibrated noise, providing an epsilon-delta privacy guarantee for the trained model. Application of DP-SGD to image classification on CIFAR-10 and MNIST demonstrated accuracy costs of five to fifteen percent compared to non-private training, depending on the privacy budget. Tramer and Boneh [25] identified fundamental trade-offs between differential privacy and adversarial

robustness in image classifiers, showing that stronger privacy guarantees reduce robustness to adversarial perturbations, a finding with significant implications for edge AI security.

Homomorphic encryption enables inference on encrypted image data, such that the inference server never observes plaintext pixel values. Gilad-Bachrach et al. [26] proposed CryptoNets, performing encrypted convolutional neural network inference using the SEAL homomorphic encryption library. While CryptoNets demonstrated the feasibility of privacy-preserving inference, evaluation latency exceeded two minutes per image, rendering it impractical for real-time edge applications. Subsequent optimisations by Juvekar et al. [27] in the GAZELLE framework combined homomorphic encryption with garbled circuits, reducing latency to under thirty milliseconds for shallow networks. Mishra et al. [28] proposed DELPHI, automating the selection between linear and non-linear computation protocols to minimise communication overhead in secure inference.

Secure multi-party computation has been applied to collaborative edge inference, enabling multiple parties to jointly compute an inference result without revealing their individual inputs. Mohassel and Zhang [29] proposed SecureML, a secure two-party computation framework for machine learning that supports linear regression, logistic regression, and neural network training. Wagh et al. [30] presented SecureNN, extending secure multi-party computation to neural network inference with three non-colluding parties, supporting ReLU, batch normalisation, and max-pooling operations essential for convolutional image processing. These approaches, while cryptographically rigorous, impose communication overhead measured in megabytes per inference, limiting their applicability to edge nodes with constrained wireless bandwidth.

3. METHODOLOGY

This review follows a systematic literature review and meta-analysis methodology aligned with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The literature search was conducted across five academic databases: IEEE Xplore, ACM Digital Library, Springer Link, Elsevier ScienceDirect, and Google Scholar. Search queries were constructed using Boolean combinations of the following primary terms: "edge AI", "edge inference", "energy efficient", "privacy preserving", "image processing", "convolutional neural network", "IoT", "federated learning", "differential privacy", and "model compression". Secondary terms including "quantisation", "neural architecture search", "secure inference", "homomorphic encryption", "in-sensor computing", and "knowledge distillation" were used in supplementary searches to capture specialised subfields. The temporal scope was restricted to publications from January 2015 to December 2024 to capture the deep learning era while maintaining manageable scope. Conference proceedings from NeurIPS, ICCV, CVPR, ICML, DAC, ISSCC, and DATE were included alongside journal publications from IEEE Transactions on Neural Networks and Learning Systems, IEEE Internet of Things Journal, ACM Transactions on Embedded Computing Systems, and Nature Electronics.

Inclusion and exclusion criteria were applied in two screening phases. In the first phase, titles and abstracts of all retrieved records were screened against the following inclusion criteria: papers must address at least one of energy efficiency or privacy preservation in the context of image or video processing using neural networks deployed on embedded or edge hardware. Purely theoretical cryptographic papers without image processing applications were excluded. Papers addressing cloud inference without edge hardware characterisation were excluded. Survey papers, review articles, and book chapters were included if they addressed the review topic but were categorised separately from primary research studies. In the second phase, full texts of candidate papers were assessed for quality indicators including reproducibility of reported results, specification of hardware platforms used for energy measurement, definition of privacy metrics or threat models, and statistical reporting sufficient for meta-analysis. After deduplication and dual-reviewer screening, one hundred and thirty-seven primary research studies and eighteen existing survey papers met all inclusion criteria and were included in this review.

For meta-analysis of quantitative results, energy efficiency was characterised by three primary metrics extracted from included studies: inference energy per image in millijoules, inference latency in milliseconds, and model size in megabytes. Privacy was characterised by epsilon values in differential privacy studies, accuracy under membership inference attack, and reconstruction error under model inversion attack where reported. Studies were grouped by technique category for within-category analysis, and cross-category comparisons were made where compatible evaluation platforms and datasets permitted. Heterogeneity across studies was assessed using the I-squared statistic for studies reporting accuracy on common benchmarks (ImageNet, CIFAR-10, CIFAR-100), and funnel plots were used to assess publication bias in energy reduction claims. Where primary studies reported results on proprietary datasets or without baseline comparisons, results were catalogued but excluded from pooled statistical analysis to avoid confounding. The taxonomy of techniques was developed inductively from the surveyed literature and validated against three independent classification frameworks identified in prior survey papers.

4. CRITICAL ANALYSIS OF PAST WORK

A systematic evaluation of the reviewed literature reveals both remarkable progress and persistent methodological limitations that constrain the generalisability of reported findings. The critical analysis presented here examines five dimensions: evaluation rigour, benchmark diversity, privacy threat model completeness, energy measurement methodology, and the maturity of co-optimisation frameworks.

Evaluation rigour across reviewed studies is highly variable. A significant fraction of energy-efficiency papers report inference energy measured on specific commercial hardware platforms under conditions that are not fully reproducible, such as custom firmware, proprietary operating modes, or single-thread execution without operating system interference. Chen et al. [15] and Sim et al. [18] provide exemplary hardware characterisation with detailed power measurement methodologies, but many subsequent papers citing these works extrapolate

results without performing equivalent measurements on their own implementations. This practice inflates apparent energy savings and makes cross-paper comparisons unreliable. Furthermore, energy per inference is frequently reported without specifying whether measured values include memory access, data transfer overhead, pre-processing, or only core arithmetic, creating systematic incompatibilities between reported values. A unified energy accounting standard analogous to the MLPerf inference benchmark [17] does not yet exist for edge privacy-preserving inference, leaving researchers without a common baseline.

Benchmark diversity in reviewed studies is dominated by CIFAR-10 and ImageNet classification tasks, which may not represent the operational conditions of real edge AI deployments. Object detection, semantic segmentation, and anomaly detection in industrial or medical imaging contexts impose different computational patterns, memory footprints, and privacy risks compared to single-label classification. Privacy risks in object detection, for example, involve localisation information that reveals human presence and trajectories, a substantially richer threat surface than class labels. Only twelve of the one hundred and thirty-seven primary studies reviewed evaluated privacy-preserving techniques on detection or segmentation tasks. The concentration on classification benchmarks creates a gap between research evaluation and deployment reality, and the absence of standardised edge-deployment image datasets with accompanying privacy annotations limits progress toward holistic evaluation.

Privacy threat model completeness is perhaps the most significant analytical weakness identified across the surveyed literature. Many papers claiming privacy-preserving properties employ a narrow threat model limited to honest-but-curious inference servers, failing to consider physically compromised devices, side-channel attacks on hardware implementations, or adversarial users who can query the deployed model. Differential privacy studies commonly report epsilon values without contextualising their practical significance: epsilon values reported range from 0.1 to 100 across reviewed studies, making privacy comparisons meaningless without a common reference. The relationship between epsilon and empirically measured attack success rates is rarely established, leaving practitioners unable to calibrate privacy guarantees against actual adversarial capability. Only seven reviewed studies provided empirical evidence of privacy protection through actual attack simulations alongside formal guarantees, representing a substantial gap between theoretical claims and practical validation.

The treatment of the energy-privacy trade-off as a co-optimisation problem is addressed in fewer than twenty percent of reviewed studies. The majority of papers address either energy efficiency or privacy preservation as their primary objective, treating the other as a constraint or afterthought. Papers in the energy-efficiency stream frequently acknowledge privacy as a future direction without quantifying the privacy implications of their compression or hardware techniques. Conversely, cryptographic privacy papers often acknowledge prohibitive energy costs without proposing viable paths to reducing them for edge hardware. This bifurcation of the research community along disciplinary lines, with hardware engineers addressing energy and cryptographers addressing privacy, has impeded the development of integrated solutions. The few papers that do address both

dimensions, notably those exploring private federated learning for image tasks [22, 23] and hardware-friendly secure inference [27, 28], demonstrate that joint optimisation is achievable but requires interdisciplinary collaboration spanning computer architecture, machine learning, and cryptography.

Reproducibility across reviewed studies is limited by insufficient reporting of implementation details. In the model compression literature, training hyperparameters, dataset preprocessing pipelines, and hardware configuration details are frequently omitted or described at insufficient resolution for independent replication. In the federated learning literature, the number of participating clients, data heterogeneity levels, and communication round counts vary widely across papers without sensitivity analysis, making it difficult to attribute performance differences to algorithmic choices rather than experimental configuration. The cryptographic inference papers reviewed are generally more rigorous in specifying security parameters and computation protocols, consistent with the higher reproducibility standards of the cryptographic research community, but they rarely provide open-source implementations compatible with practical edge hardware.

5. DISCUSSION

The synthesis of reviewed literature reveals several important implications for the design of future edge AI image processing systems and for the research agenda of the field. Three themes merit extended discussion: the path toward practical co-optimisation, the role of regulatory context, and the emerging potential of neuromorphic and in-sensor computing.

The path toward practical co-optimisation of energy and privacy requires the development of unified design frameworks that model both objectives using compatible metrics. Current practice treats energy in hardware units (millijoules per inference) and privacy in information-theoretic units (epsilon, mutual information leakage), creating an integration barrier. Future frameworks must establish conversion functions or Pareto frontier representations that allow designers to navigate trade-offs explicitly. Promising directions include privacy-aware neural architecture search, in which the NAS objective function incorporates estimated privacy leakage alongside energy and accuracy, and hardware-privacy co-design, in which silicon architectures are evaluated not only for energy efficiency but also for resistance to side-channel attacks and model inversion. The integration of trusted execution environments, such as ARM TrustZone, into edge inference pipelines offers a near-term path to hardware-enforced privacy isolation without the prohibitive energy cost of homomorphic encryption.

The regulatory context for edge AI image processing is evolving rapidly. The European Union AI Act, which entered force in 2024, classifies real-time biometric identification systems as high-risk AI, imposing stringent conformity assessment and transparency requirements. India's Digital Personal Data Protection Act, 2023 similarly establishes consent and purpose limitation requirements for processing personal image data. These regulatory developments create both obligations and incentives for deploying privacy-preserving edge AI

systems. Organisations deploying smart cameras, workplace monitoring systems, or health-monitoring wearables must now demonstrate technical compliance with privacy regulations, making on-device privacy preservation not merely an academic objective but a legal requirement. Research that advances deployable privacy-preserving edge inference directly addresses this compliance imperative.

The emerging potential of neuromorphic and analogue computing represents a disruptive opportunity for energy-efficient, privacy-preserving image processing. Neuromorphic processors such as Intel Loihi 2 and IBM's NorthPole architecture achieve inference energy orders of magnitude below conventional digital processors by exploiting event-driven computation and collocated memory and processing. While current neuromorphic inference tools do not support standard convolutional neural network deployments directly, the conversion of spiking neural network equivalents is an active research area. The analogue inference approach demonstrated in in-sensor computing papers reduces energy further by performing computation in the analogue domain before digitisation, inherently limiting the reconstructability of raw images. If cryptographic privacy mechanisms can be adapted to these computing substrates, the resulting systems could achieve simultaneously extreme energy efficiency and strong privacy protection, resolving the tension that currently limits edge AI deployments.

6. CONCLUSION

This paper has presented a comprehensive systematic review and meta-analysis of research addressing energy-efficient and privacy-preserving image processing for edge AI systems. The reviewed literature demonstrates substantial progress across model compression, hardware acceleration, federated learning, differential privacy, and secure inference, with individual contributions achieving significant advances in isolation. However, the critical analysis reveals that the field has not yet produced unified frameworks for co-optimising energy and privacy, that evaluation methodologies remain heterogeneous and insufficiently rigorous for reliable cross-study comparison, and that privacy threat models applied in most studies are narrower than the adversarial conditions facing deployed systems.

The convergence of regulatory pressure, hardware innovation, and algorithmic maturity creates a favourable environment for breakthroughs in practical co-optimisation. Key open challenges include: standardised benchmarks for joint energy-privacy evaluation; privacy-aware neural architecture search; hardware-efficient implementations of differential privacy and secure aggregation; and formal models linking energy budgets to privacy guarantees. Addressing these challenges will require sustained interdisciplinary collaboration spanning the traditionally separate communities of hardware design, machine learning, and cryptography. As edge AI deployments proliferate across healthcare, manufacturing, urban infrastructure, and consumer electronics, the development of systems that are simultaneously efficient and trustworthy is not merely a research aspiration but an industrial and societal necessity. This review provides a consolidated foundation from which researchers and practitioners can orient their contributions toward this critical objective.

7.REFERENCES

[1] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," arXiv preprint arXiv:1704.04861, 2017.

[2] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR), Salt Lake City, UT, USA, 2018, pp. 4510–4520.

[3] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in Proc. IEEE/CVF Int. Conf. Comput. Vision (ICCV), Seoul, South Korea, 2019, pp. 1314–1324.

[4] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR), Salt Lake City, UT, USA, 2018, pp. 6848–6856.

[5] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR), Salt Lake City, UT, USA, 2018, pp. 2704–2713.

[6] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "XNOR-Net: ImageNet classification using binary convolutional neural networks," in Proc. Eur. Conf. Comput. Vision (ECCV), Amsterdam, Netherlands, 2016, pp. 525–542.

[7] Z. Liu, B. Wu, W. Luo, X. Yang, W. Liu, and K.-T. Cheng, "Bi-real net: Enhancing the performance of 1-bit CNNs with improved representational capability and advanced training algorithm," in Proc. Eur. Conf. Comput. Vision (ECCV), Munich, Germany, 2018, pp. 747–763.

[8] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," J. Mach. Learn. Res., vol. 18, no. 1, pp. 6869–6898, 2018.

[9] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.

[10] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," in Proc. Int. Conf. Learn. Representations (ICLR), San Diego, CA, USA, 2015.

[11] A. Mishra and D. Marr, "Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy," in Proc. Int. Conf. Learn. Representations (ICLR), Vancouver, BC, Canada, 2018.

[12] H. Cai, L. Zhu, and S. Han, "ProxylessNAS: Direct neural architecture search on target task and hardware," in Proc. Int. Conf. Learn. Representations (ICLR), New Orleans, LA, USA, 2019.

[13] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. Int. Conf. Mach. Learn. (ICML), Long Beach, CA, USA, 2019, pp. 6105–6114.

[14] B. Wu, X. Dai, P. Zhang, Y. Wang, F. Sun, Y. Wu, Y. Tian, P. Vajda, Y. Jia, and K. Keutzer, "FBNet: Hardware-aware efficient ConvNet design via differentiable neural architecture search," in Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. (CVPR), Long Beach, CA, USA, 2019, pp. 10734–10742.

[15] Y.-H. Chen, T. Krishna, J. S. Emer, and V. Sze, "Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks," IEEE J. Solid-State Circuits, vol. 52, no. 1, pp. 127–138, 2017.

[16] N. P. Jouppi et al., "In-datacenter performance analysis of a tensor processing unit," in Proc. 44th Annu. Int. Symp. Comput. Archit. (ISCA), Toronto, ON, Canada, 2017, pp. 1–12.

[17] V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient processing of deep neural networks: A tutorial and survey," Proc. IEEE, vol. 105, no. 12, pp. 2295–2329, 2017.

[18] J. Sim, J.-S. Park, M. Kim, D. Bae, Y. Choi, and L.-S. Kim, "A 1.42TOPS/W deep convolutional neural network recognition processor for intelligent IoE systems," in Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC), San Francisco, CA, USA, 2016, pp. 264–265.

[19] J. Choi, Z. Cai, D. Niu, and B. Lin, "Near-sensor convolutional neural network for image recognition on CMOS image sensor," in Proc. IEEE Custom Integr. Circuits Conf. (CICC), Austin, TX, USA, 2019, pp. 1–4.

[20] M. Seo, H. Oh, J. Park, and J.-J. Lee, "In-sensor computing using memristive array for feature extraction in edge AI applications," IEEE Trans. Electron Devices, vol. 69, no. 6, pp. 3139–3146, 2022.

[21] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artif. Intell. Statist. (AISTATS), Fort Lauderdale, FL, USA, 2017, pp. 1273–1282.

[22] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," IEEE Signal Process. Mag., vol. 37, no. 3, pp. 50–60, 2020.

[23] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning," arXiv preprint arXiv:1712.07557, 2017.

[24] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in Proc. 23rd ACM SIGSAC Conf. Comput. Commun. Security (CCS), Vienna, Austria, 2016, pp. 308–318.

[25] F. Tramer and D. Boneh, "Differentially private learning needs better features (or much more data)," in Proc. Int. Conf. Learn. Representations (ICLR), Virtual, 2021.

[26] R. Gilad-Bachrach, N. Dowlin, K. Laine, K. Lauter, M. Naehrig, and J. Wernsing, "CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy," in Proc. 33rd Int. Conf. Mach. Learn. (ICML), New York, NY, USA, 2016, pp. 201–210.

[27] C. Juvekar, V. Vaikuntanathan, and A. Chandrakasan, "GAZELLE: A low latency framework for secure neural network inference," in Proc. 27th USENIX Security Symp., Baltimore, MD, USA, 2018, pp. 1651–1669.

[28] P. Mishra, R. Lehmkuhl, A. Srinivasan, W. Zheng, and R. A. Popa, "DELPHI: A cryptographic inference service for neural networks," in Proc. 29th USENIX Security Symp., Virtual, 2020, pp. 2505–2522.

[29] P. Mohassel and Y. Zhang, "SecureML: A system for scalable privacy-preserving machine learning," in Proc. IEEE Symp. Security Privacy (SP), San Jose, CA, USA, 2017, pp. 19–38.

[30] S. Wagh, D. Gupta, and N. Chandran, "SecureNN: 3-party secure computation for neural network training," Proc. Priv. Enhancing Technol. (PoPETs), vol. 2019, no. 3, pp. 26–49, 2019.