ReCIM-HD: A 40 nm On-Device Learning Accelerator with Hyperdimensional Computing via Importance-Based Hybrid Digital–ReRAM CIM

Keming Fan[†1], Chang Eun Song[†1], Soumil Jain[1], Weihong Xu[2], Gopabandhu Hota[1], Haichao Yang[1], Leo Liu[3], Meng-Fan Chang[3], Carlos H. Diaz[3], Gert Cauwenberghs[1], Tajana Rosing[1], and Mingu Kang[1] ([†]co-first)
[1]University of California, San Diego, USA [2]EPFL, Switzerland [3]Taiwan Semiconductor Manufacturing Company (TSMC)

***Abstract*—ReCIM-HD is an on-device learning (ODL) accelerator that integrates gradient-free hyperdimensional computing (HDC) with resistive RAM (ReRAM)-based analog compute-in-memory (CIM). ReCIM-HD introduces three key techniques: (1) ultra-low-power HDC-based edge ODL without costly back-propagation operations; (2) non-volatile ReRAM-based storage for class hypervectors and CIM similarity search, enabling energy-efficient inference and power-gating during idle periods without data loss; and (3) a hybrid digital–analog CIM inference flow that protects a small accuracy-critical subset digitally to preserve accuracy. Additionally, it employs a Kronecker HD encoder (KE) to reduce encoding overhead and a customized instruction set architecture (ISA) to support flexible programmability. Fabricated in 40 nm, ReCIM-HD achieves 44.17 TOPS/W (hybrid mode) and 75.8 TOPS/W (ReRAM CIM mode), delivering up to 20.1× higher energy efficiency than the state-of-the-art accelerator.**

***Index Terms*—On-Device Learning, Compute-in-memory, Hyperdimensional Computing**

I. Introduction

On-device learning (ODL) is crucial for continual adaptation to environmental changes while preserving data privacy and security. However, edge platforms operate under severe constraints in compute capability, memory capacity, and data bandwidth [1], [2]. Existing ODL accelerators [3]–[7] face three key challenges as described in Fig. 1. **(C1)** Gradient-based training incurs a high computational cost, and limited on-chip memory forces frequent data movement between processors and memory. **(C2)** Volatile memory (e.g., DRAM and SRAM) requires frequent model reloads when edge devices, which are largely idle due to intermittent inputs, are power-gated [4]. While non-volatile memory (NVM)-based compute-in-memory (CIM) (Fig. 2(b)) alleviates reload overhead [8], **(C3)** analog CIM and device non-idealities result in significant accuracy degradation.

To tackle these challenges, ReCIM-HD introduces several key techniques. **(S1)** We adopt gradient-free hyperdimensional computing (HDC) learning for efficient ODL (Fig. 2(a)), using simple and highly parallelizable bit-wise operations. However, conventional HDC relies on high-dimensional hypervectors (HVs) stored in large associative memories, incurring substantial storage and data movement overhead. We therefore exploit high-density resistive RAM (ReRAM) as an NVM substrate to store HVs in place. **(S2)** This enables near-zero standby power during idle periods under power-gating while preserving stored data, and low-power, highly parallel

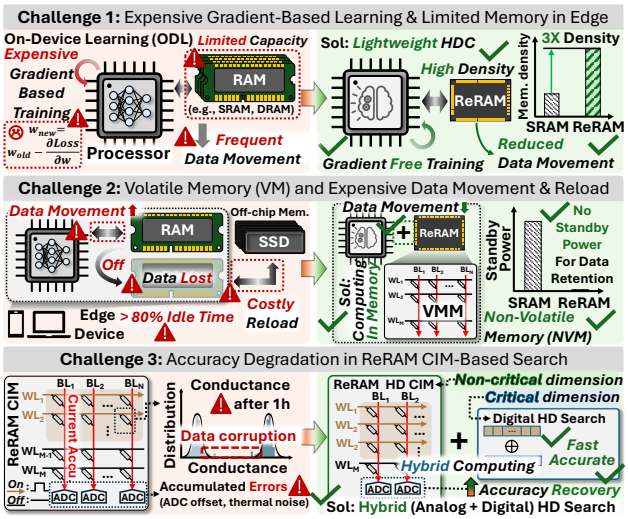


Fig. 1. Challenges of ODL at the edge and proposed solutions with hybrid HDC using ReRAM-based CIM.

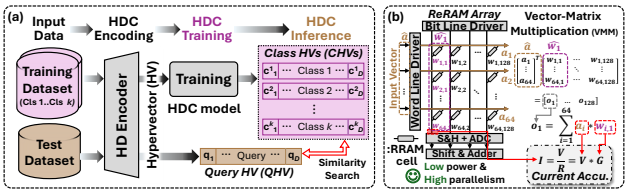


Fig. 2. Overview of (a) HDC and (b) ReRAM CIM.

CIM-based similarity search to reduce data movement. **(S3)** To mitigate non-idealities from analog CIM and device imperfection, ReCIM-HD employs an importance-based hybrid digital–analog execution scheme, where only a small accuracy-critical portion of dimensions is processed digitally while the majority is mapped to ReRAM CIM. Furthermore, ReCIM-HD incorporates a Kronecker HD encoder (KE) to reduce encoding overhead and a customized instruction set architecture (ISA) to flexibly program both HDC and CIM operations.

II. Proposed Design

Fig. 3 shows ReCIM-HD overall architecture. The HD module supports the full HDC pipeline (encoding, training, and similarity search) digitally, while the ReRAM CIM module stores bipolar (+1/-1) class HVs (CHVs) and performs inference via CIM-based analog similarity search, with a FIFO interface managing data exchange and synchronization.

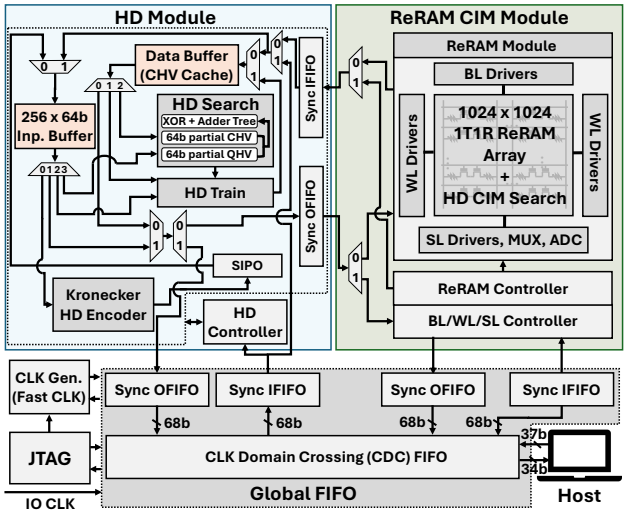


Fig. 3. Proposed ReCIM-HD hardware with key components including HD module, ReRAM CIM module, and global FIFO.

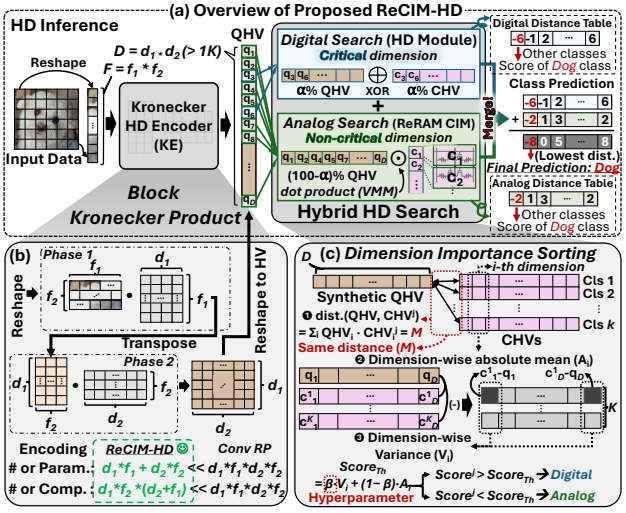


Fig. 4. (a) Proposed ReCIM-HD dataflow & HD inference via hybrid HD search, (b) overview of KE, and (c) identification of critical dimensions based on importance sorting.

A. ReCIM-HD Inference and Training Dataflow

Fig. 4(a) shows the proposed ReCIM-HD data flow during HD inference. Input features are encoded into a bipolar (+1/-1) query HV (QHV) by the KE [9] (Fig. 4(b)), which consists of two phases, each performing a reshape followed by matrix multiplication to generate the encoded HV. The KE reduces storage by $204.8\times$ and computation by $21.3\times$ compared to random projection-based encoding [10] by factorizing a large projection matrix into two smaller matrices. Then, the encoded QHV is compared against all stored bipolar CHVs through dot-product–based similarity search to generate a distance table where the class with the smallest distance is taken as the final prediction. To balance efficiency and accuracy, ReCIM-HD proposes a *hybrid digital–analog HD search* scheme in which accuracy-critical dimensions, identified one-time offline, are processed digitally through the HD module while the remaining majority are mapped to ReRAM and processed via analog CIM (Section II-B). As shown in Fig. 4(c), we construct a synthetic vector that yields identical similarity scores across all CHVs, and analyze each dimension by measuring the mean (A_i) and variance (V_i) of the deviations of CHVs from this synthetic reference for the i-th dimension. Based on $\beta \cdot V_i + (1-\beta) \cdot A_i$, where β is a hyperparameter, all dimensions are ranked, and the top α fraction of dimensions is selected as the critical set, which is static during inference. For the digital portion, the critical dimensions of each CHV are stored in the on-chip cache. In each cycle, 64 MSBs of a CHV segment are fetched and compared against the corresponding QHV segment using the XOR and adder tree to compute partial similarity scores. For the analog portion, the remaining dimensions ($\approx 90\%$, i.e., $1-\alpha$) are stored in ReRAM and evaluated through analog CIM. The two similarity scores are summed to produce the final score in the distance table for classification.

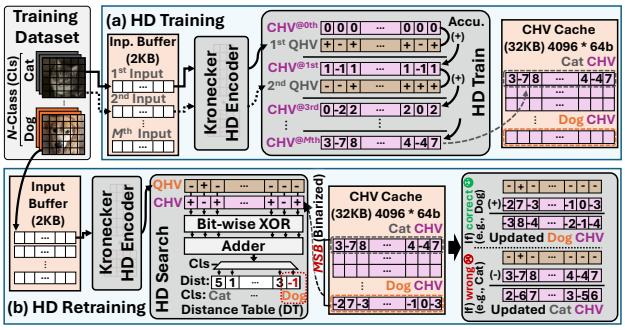


Fig. 5. HDC operations of (a) training and (b) retraining.

Fig. 5 shows the details of on-device HD training and retraining. (a) During HD training, the HD module encodes each sample into a bipolar QHV using the KE and accumulates same-class QHVs to form CHVs. (b) HD retraining further updates the initial CHVs by first performing HD inference. If the predicted class is correct, the corresponding CHV is reinforced by adding the QHV; otherwise, the QHV is subtracted from the misclassified CHV to refine class representations.

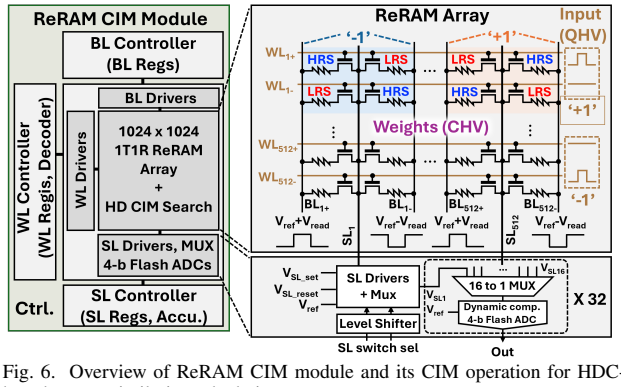


Fig. 6. Overview of ReRAM CIM module and its CIM operation for HDC-based vector similarity calculation.

B. ReRAM Macro

Fig. 6 outlines the ReRAM module, which consists of a digital ReRAM controller, a 1k×1k 1T1R ReRAM array, and

its peripherals, including drivers and ADCs. The ReRAM array supports two operating modes. For normal read/write operations, ReCIM-HD uses a single-differential mapping with two cells per bipolar bit (+1/-1) to maximize storage density. For CIM-based similarity search, it adopts a double-differential weight mapping scheme using four cells per bipolar bit (i.e., representing +1/-1 as shown in Fig. 6), improving accuracy and noise tolerance. During CIM, each CHV is stored per column, and if the vector dimension exceeds the number of rows, the CHV is split into segments and distributed over multiple columns. The bipolar input vector QHV is encoded through wordlines (WLs), while BL+ and BL− are biased at $V_{\text{ref}} + V_{\text{read}}$ and $V_{\text{ref}} - V_{\text{read}}$, respectively, independent of the input. Depending on the stored CHV polarity and the applied QHV polarity, each activated cell contributes a positive or negative current to the shared sourceline (SL). At steady state, the accumulated voltage on the SL capacitor (V_{SL}) represents the dot-product-based similarity score, as shown in Fig. 7. The resulting SL voltage is then multiplexed and converted by a 4-bit flash analog-to-digital converter (ADC). Each ADC is shared across 16 SLs to save area while matching throughput. These partial sums are further accumulated in the ReRAM controller to cover full dimensions, and are then forwarded to HD module for end-to-end inference.

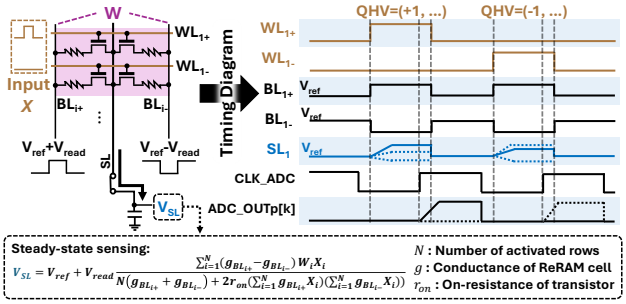
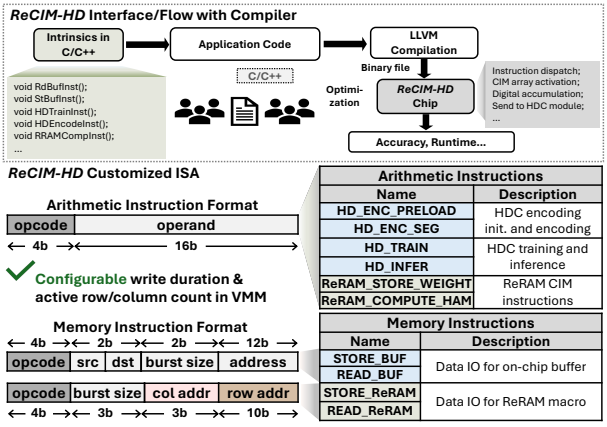


Fig. 7. ReRAM analog CIM operation, and timing diagram.

C. Customized ISA

Fig. 8 presents the customized ISA of ReCIM-HD and its software-to-hardware execution flow. The ISA adopts a 20-bit instruction format consisting of a 4-bit opcode and a 16-bit operand, enabling flexible programmability across both digital HDC operations and ReRAM CIM execution. ReCIM-HD supports basic HDC arithmetic instructions, such as encoding and training, as well as memory-related instructions, such as store and read. For CIM execution, the ReRAM controller exposes several programmable knobs through the ISA. It supports configurable write durations to balance write energy and latency against programming accuracy. Configurable row-level parallelism is available during vector–matrix multiplication (VMM), with 16/32/64-row activation modes, along with configurable column parallelism through burst-size–controlled column activation. We implement C/C++ intrinsics for each instruction so that inline assembly in high-level code compiles directly to the corresponding bytecode. This lightweight ISA/compiler stack enables flexible programmability while preserving low-overhead hardware control.

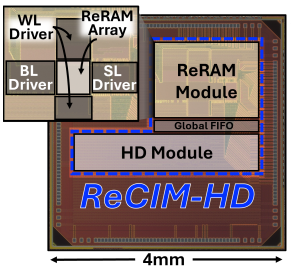


Arithmetic Instructions	
Name	Description
HD_ENC_PRELOAD	HDC encoding init. and encoding
HD_ENC_SEG	
HD_TRAIN	HDC training and inference
HD_INFER	
ReRAM_STORE_WEIGHT	ReRAM CIM instructions
ReRAM_COMPUTE_HAM	

Memory Instructions	
Name	Description
STORE_BUF	Data IO for on-chip buffer
READ_BUF	
STORE_ReRAM	Data IO for ReRAM macro
READ_ReRAM	

Fig. 8. Customized ISA and compiler interface for fast and flexible programming for ReCIM-HD.

III. Silicon Measurement and Results



Chip Summary Table	
Technology	40nm CMOS
Core Size	5.1 mm² (10.6 mm² w/ IO PAD)
ReRAM	Foundry 1T1R ReRAM
Capacity (KB)	ReRAM: 32–128,SRAM: 34 (HDC)
Supply Voltage	0.7V–1.1V
Frequency	50MHz – 200MHz
Model	HDC
Precision	HDC inference: INT1-8 HDC training: INT8 for CHV
Feature Dim. (*F*)	8-1024
Max HDC Dim. (*D*)	1024 (Training), 8192 (Inference)
Max # of Class	32
Peak Energy Efficiency (TOPS/W)	HDC: 1.53-3.78, ReRAM CIM: 75.8 HDC + ReRAM CIM:10.03-44.17

Fig. 9. ReCIM-HD die photo with chip specifications.

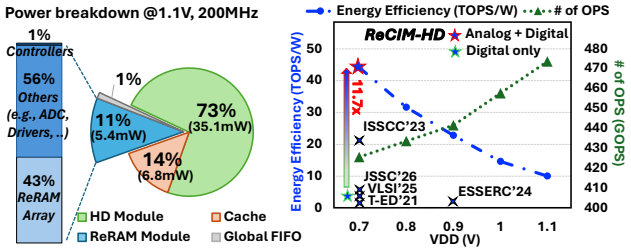


Fig. 10. ReCIM-HD power breakdown and energy efficiency of ReCIM-HD across different supply voltages of the HD module.

Fig. 9 shows the chip micrograph and specifications of ReCIM-HD, fabricated in 40 nm CMOS with a die area of 5.1 mm^2, integrated with a foundry 128 KB 1T1R ReRAM macro. The HD module and ReRAM CIM module can operate over a frequency range of 50–200 MHz. The power breakdown in Fig. 10 shows that the ReRAM module consumes only 5.4 mW, accounting for 11% of the total chip power while achieving 75.8 TOPS/W, highlighting the high energy efficiency of analog CIM. By using hybrid search, ReCIM-HD achieves up to 44.17 TOPS/W during inference at 0.7–1.1 V when the critical dimension ratio is $\alpha = 10\%$. We measure both ReRAM read operations for storage and CIM operations for in-memory vector-matrix multiplication (VMM). Each ReRAM cell's resistance is characterized by collecting the cumulative distribution function (CDF) of readout values from 1024 cells using the on-chip 4-bit flash ADC configured as a

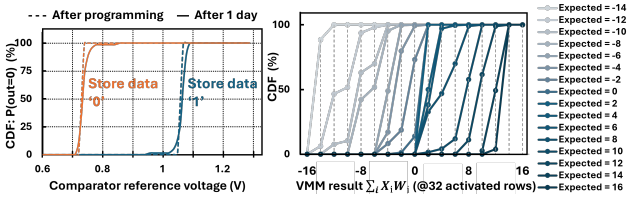


Fig. 11. Measured results of ReRAM read operation and VMM CIM.

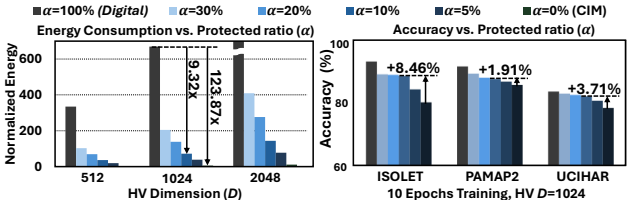


Fig. 12. Energy consumption and accuracy of ReCIM-HD across different protected ratios (α).

1-bit comparator while sweeping its reference voltage. Fig. 11 shows the binary readout distributions measured immediately after programming and again one day later. It also shows the CIM results as CDF of expected and measured VMM outputs with 32 activated rows per cycle during inference.

End-to-end system accuracy is evaluated on three benchmarks, including ISOLET for speech recognition [11], PAMAP2 [12] for physical activity monitoring, and UCI-HAR [13] for human activity recognition. Fig. 12 summarizes the accuracy and energy efficiency across different critical dimension ratios α. While a fully analog CIM implementation achieves 123.87× higher energy efficiency than fully digital processing, this gain comes at the cost of reduced accuracy. By contrast, the hybrid HD search in ReCIM-HD effectively balances accuracy and efficiency: with only $\alpha = 10\%$, it improves accuracy by 1.91–8.46% over the CIM-only design while still reducing energy consumption by 9.32× relative to the digital-only implementation. By increasing α, the accuracy can be further recovered, at the cost of higher energy consumption.

TABLE I
COMPARISON WITH STATE-OF-THE-ART MODELS.

	Our work (ReCIM-HD)	JSSC'26[3]	VLSI'25[4]	ESSERC'24[5]	ISSCC'23[6]	TED'21[7]
Node	40 nm	40 nm	40 nm	40 nm	22 nm	40 nm
Model	HDC	CNN	HDC	HDC	CNN	MANN
Training & Inference	Yes	Yes	Yes	Yes	Inference only	Yes
Design	Digital + CIM (Hybrid)	Digital	Digital	Digital	ReRAM CIM	ReRAM CIM
Precision	INT1-8	INT4	BF16/ INT1-8	BF16/ INT16	INT1-8	FP32
On-chip Mem. (kB)	34 (SRAM)+ 128 (ReRAM)	71	200	424	512 (SRAM)+ 4096 (ReRAM)	8
Area (mm^2)	10.6	0.83	14.4	11.3	30.6	0.2
Freq. (MHz)	50-200	150	50-250	100-250	50-200	200
Supply voltage (V)	0.7-1.1	0.6-1.1	0.7-1.2	0.9-1.2	0.7-0.8	-
Area Effi. (Peak TOPS/mm^2)	0.092 (HDC + ReRAM) 0.080 (ReRAM CIM)	0.104	0.027	0.014	0.086 (Scaled to 40nm [14])	0.002
Energy Effi. (Peak TOPS/W)	44.17 (HDC + ReRAM) 75.8 (ReRAM CIM)	6	3.78	0.78	20.84 (Scaled to 40nm [14])	0.12

Table I compares ReCIM-HD with prior accelerators. Unlike prior ReRAM CIM designs [6] focusing on inference only, ReCIM-HD supports both learning and inference. CNN-based accelerators enable training [3] but rely on costly gradient operations, while HDC accelerators [4], [5] allow lightweight learning but are fully digital and do not utilize the benefits of analog CIM. ReCIM-HD is the first silicon prototype to combine HDC-based ODL with CIM-based inference, achieving 20.1–56.6× and 3.4–6.6× higher energy and area efficiency, respectively, than prior works by leveraging the high efficiency of analog CIM and the high density of ReRAM.

IV. Conclusion

This paper presents ReCIM-HD, the first silicon prototype to combine HDC-based on-device learning with ReRAM CIM-based inference. ReCIM-HD enables low-power gradient-free edge learning, non-volatile model retention under power-gating, and energy-efficient in-memory inference. To mitigate analog-CIM non-idealities, ReCIM-HD further adopts an importance-based hybrid digital–analog search that protects only a small accuracy-critical subset of dimensions digitally. Finally, it employs an optimized HD encoder and a customized ISA. Fabricated in 40 nm CMOS with a foundry 128 KB 1T1R ReRAM macro, ReCIM-HD achieves up to 20.1× higher energy efficiency than SOTA accelerators [4].

Acknowledgment

This work was supported by TSMC and in part by PRISM and CoCoSys, centers in JUMP 2.0, an SRC program sponsored by DARPA.

References

[1] G. Karunaratne *et al.*, "In-memory realization of in-situ few-shot continual learning with a dynamically evolving explicit memory," *ESSCIRC*, pp. 105–108, 2022.
[2] X. Ma *et al.*, "Cost-effective on-device continual learning over memory hierarchy with miro," *International Conference on Mobile Computing and Networking*, pp. 1–15, 2023.
[3] D. Blanken *et al.*, "Chameleon: A multiplier-free temporal convolutional network accelerator for end-to-end few-shot and continual learning from sequential data," *JSSC*, 2026.
[4] C. E. Song *et al.*, "Clo-hdnn: A 4.66 tflops/w and 3.78 tops/w continual on-device learning accelerator with energy-efficient hyperdimensional computing via progressive search," *VLSI*, pp. 1–3, 2025.
[5] H. Yang *et al.*, "Fsl-hdnn: A 5.7 tops/w end-to-end few-shot learning classifier accelerator with feature extraction and hyperdimensional computing," *ESSERC*, pp. 33–36, 2024.
[6] W.-H. Huang *et al.*, "A nonvolatile al-edge processor with 4mb slc-mlc hybrid-mode reram compute-in-memory macro and 51.4-251tops/w," *ISSCC*, pp. 15–17, 2023.
[7] H. Li *et al.*, "Sapiens: A 64-kb rram-based non-volatile associative memory for one-shot learning and inference at the edge," *IEEE Transactions on Electron Devices*, vol. 68, no. 12, pp. 6637–6643, 2021.
[8] I. Kazi *et al.*, "Energy/reliability trade-offs in low-voltage reram-based non-volatile flip-flop design," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 11, pp. 3155–3164, 2014.
[9] W. Xu *et al.*, "Hypermetric: Efficient hyperdimensional computing with metric learning for robust edge intelligence," *TCAD*, 2025.
[10] A. Hernández-Cano *et al.*, "Reghd: Robust and efficient regression in hyper-dimensional learning system," in *DAC*, 2021, pp. 7–12.
[11] R. Cole and M. Fanty, "ISOLET," UCI Machine Learning Repository, 1991, DOI: https://doi.org/10.24432/C51G69.
[12] A. Reiss, "PAMAP2 Physical Activity Monitoring," UCI Machine Learning Repository, 2012, DOI: https://doi.org/10.24432/C5NW2H.
[13] J. Reyes-Ortiz *et al.*, "Human Activity Recognition Using Smartphones," UCI Machine Learning Repository, 2013, DOI: https://doi.org/10.24432/C54S4K.
[14] A. Stillmaker *et al.*, "Scaling equations for the accurate prediction of cmos device performance from 180 nm to 7 nm," *Integration*, vol. 58, pp. 74–81, 2017.