A Delta-Sigma Based SRAM Compute-in-Memory Macro for Human Activity Recognition

Vasundhara Damodaran, Ziyu Liu, Jae-sun Seo, and Arindam Sanyal
School of Electrical, Computer and Energy Engineering, Arizona State University, AZ, USA.
Email: vdamoda2@asu.edu

2023 IEEE Biomedical Circuits and Systems Conference (BioCAS) | 979-8-3503-0026-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/BioCAS58349.2023.10388747

***Abstract*—This work presents an SRAM based compute-in-memory (CIM) macro that uses 1-bit $\Delta\Sigma$ modulators to convert input and output activations to binary pulse waveform. The SRAM macro uses switched-capacitors for vector matrix multiplications and together with binary input activation improves linearity compared to current-domain SRAM CIM macros and allows reconfigurable activation resolution. The proposed macro is fabricated in 65nm and used for human activity recognition classifier that achieves more than 96% accuracy while consuming 93-420.5pJ/classification**

***Index Terms*—compute-in-memory, human activity recognition, static random access memory, artificial neural network, delta-sigma**

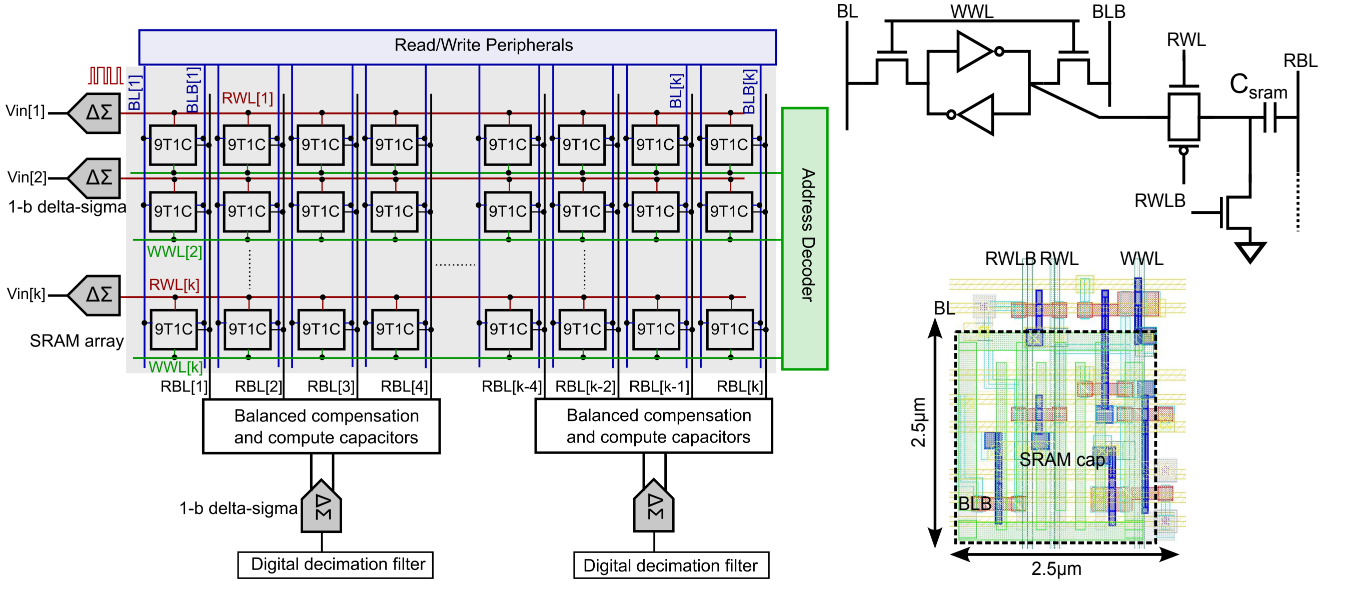
I. Introduction

Automatic recognition of daily human activity provides important contextual information that can be useful for health and wellness monitoring. Development of wearable sensing technology is an enabling factor for acquiring greater insights from long-term recording of human activity in daily life settings which might provide important clues to improving preventive healthcare. A key drawback of existing wearable devices for human activity recognition (HAR) task is that they cannot analyze activity patterns on-device and need to transmit sensor data to cloud server for processing. Radio-frequency (RF) transmission typically consumes 60-90% of the entire wireless sensor power consumption [1], [2] and limits battery life of wearable devices. The need to frequently recharge device disrupts monitoring and reduces user compliance as many users forget to put the device back on [3]. On-device artificial intelligence (AI) algorithms can significantly reduce transmission volume by analyzing sensor data locally and only transmitting inference results. A key challenge with embedding AI algorithms in resource constrained wearable devices is the energy requirement of complex AI models. Compute-in-memory (CIM) is an energy efficient technique to perform AI computations inside memory units storing the AI model weights. By reducing communication costs of bringing together many input activations, neuron weights and distributing output activations, CIM breaks the von-neumann bottleneck and improves energy-efficiency significantly compared to existing CPU/GPUs. Out of different CIM techniques, SRAM based CIM is widely used due to its high energy efficiency and easy integration with CMOS ICs [4], [5]. A fundamental limitation in SRAM-CIM is nonlinearity of the access transistors to which the input activation is applied. Large values of vector matrix multiplication (VMM) products computed by SRAM array make the currents through the access transistors change nonlinearly with the VMM products which in turn makes the VMM product nonlinear [6]. Recent works have tried to address this nonlinearity issue through - a) 1-bit activation [7]; b) converting analog input activations to binary pulse trains [6]; c) charge-domain computation using switched-capacitors [5]. 1-bit activation requires boosting with multiple classifiers to achieve good performance which reduces energy efficiency. Pulsed input activation improves linearity but cannot fundamentally address nonlinearity due to current-domain accumulation in SRAM array and introduces quantization error due to conversion of analog input to pulsed input. Switched-capacitor based CIM improves linearity, but current designs are still limited by nonlinearity of access transistors that apply input activation to the capacitors.

This work presents a switched-capacitor based $\Delta\Sigma$ SRAM CIM technique that addresses the above challenges with a combination of 1-bit $\Delta\Sigma$ modulators that convert activations to binary pulse trains with lower in-band quantization error than in [6] and 9T1C SRAM bitcells that perform computations in charge-domain with high linearity. The use of $\Delta\Sigma$ modulator also allows variable resolution for input and output activations. Fabricated in 65nm CMOS, the proposed SRAM CIM macro consumes 93-420pJ/classification and identifies 5 human activities - sitting, standing, walking, running and dancing with more than 96% accuracy. The paper is organized as follows: Section II presents the proposed SRAM CIM architecture and circuit design techniques, Section III presents measurement results on the test-chip and Section IV brings up the conclusion.

II. Proposed Architecture

Fig. 1 shows architecture of the proposed $\Delta\Sigma$ SRAM array and the 9T1C bitcell schematic. 1-b $\Delta\Sigma$ modulators convert analog inputs into binary pulse train activation and apply to the SRAM bitcells through the RWL lines. The SRAM cells use switched-capacitor circuits to perform charge-domain multiplication of input activation and model weights stored in the SRAM cells. Accumulation is performed in charge-domain on the RBL bitlines and sent to output 1-b $\Delta\Sigma$ modulators and digitally decimated. Quantization noise-shaping in the $\Delta\Sigma$ modulators ensure lower in-band quantization error than the analog-to-binary pulse conversion technique using counters in [6]. Accumulation in charge-domain through charge redistribution in the proposed architecture results in higher



(a)

(b)

Fig. 1: a) Proposed ΔΣ based pulsed activation CIM macro b) schematic of 9T1C SRAM bitcell and layout

linearity than current-domain accumulation in conventional SRAM-CIM macros. Use of binary input activation and CMOS switch to sample input activation on the capacitor as shown in Fig. 1b) improves linearity compared to [5] by removing dependency of sampled voltage on input activation. The CMOS switch also allows the SRAM bit-cell to operate from low supply voltage and improves energy efficiency. The layout of a single SRAM bitcell is shown in Fig. 1b). A MOM capacitor is used to realize SRAM capacitor C_{sram} which is placed on top of the SRAM cell.

Fig. 2 shows the schematic of a single slice of the proposed macro using 4-bit signed weights. Compensation capacitors are used to ensure all RBL lines see the same capacitive load during computations. The capacitors and RBL lines are discharged during ϕ_1. During ϕ_2, the SRAM cells compute the VMM products on each RBL line which are charge-shared with the binary weighted compute capacitor bank. The compute capacitors are disconnected from the SRAM cells during ϕ_3 and charge-shared with additional balancing capacitors to ensure correct binary weighted MAC result with sign bit operation. At the end of ϕ_3, the sign-bit output V_{im}, and the remaining 3-bit output V_{ip} are given by

$$V_{im} = \frac{8}{15} \cdot \left[\left(\sum V_{in}[k] \cdot W_{i,k}\right)\right] \cdot \frac{C_{sram}}{NC_{sram} + 9C} \tag{1}$$

$$\begin{aligned} V_{ip} \; = \; & \left[\frac{4}{15} \cdot \left(\sum V_{in}[k] \cdot W_{i+1,k}\right)\right. \\ & + \frac{2}{15} \cdot \left(\sum V_{in}[k] \cdot W_{i+2,k}\right) \\ & + \left.\frac{1}{15} \cdot \left(\sum V_{in}[k] \cdot W_{i+3,k}\right)\right] \cdot \frac{C_{sram}}{NC_{sram} + 9C} \end{aligned} \tag{2}$$

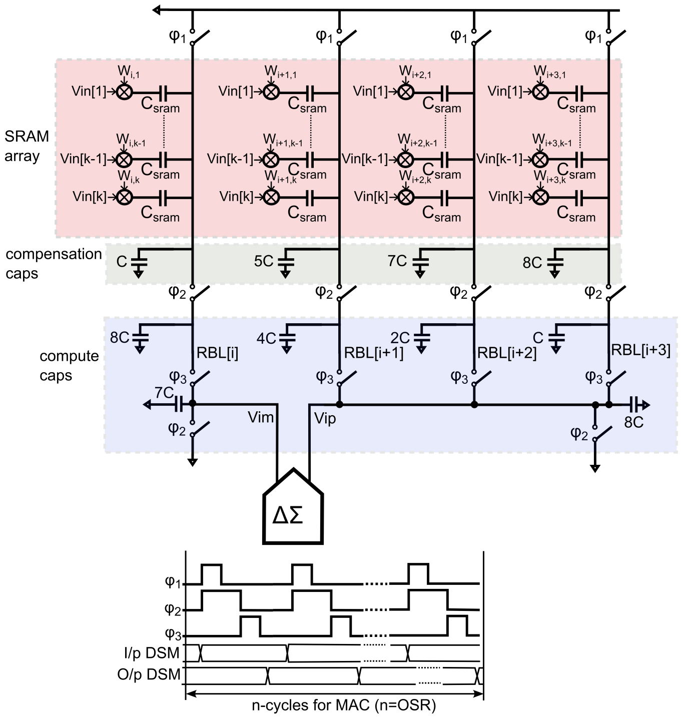


Fig. 2: Circuit schematic of a single slice of the CIM macro and associated timing diagram

where V_{in} is the input activation and W is the neural network weight array. V_{im} and V_{ip} are applied differentially to the output ΔΣ modulator as shown in Fig. 2 that provides a binary pulse train as output.

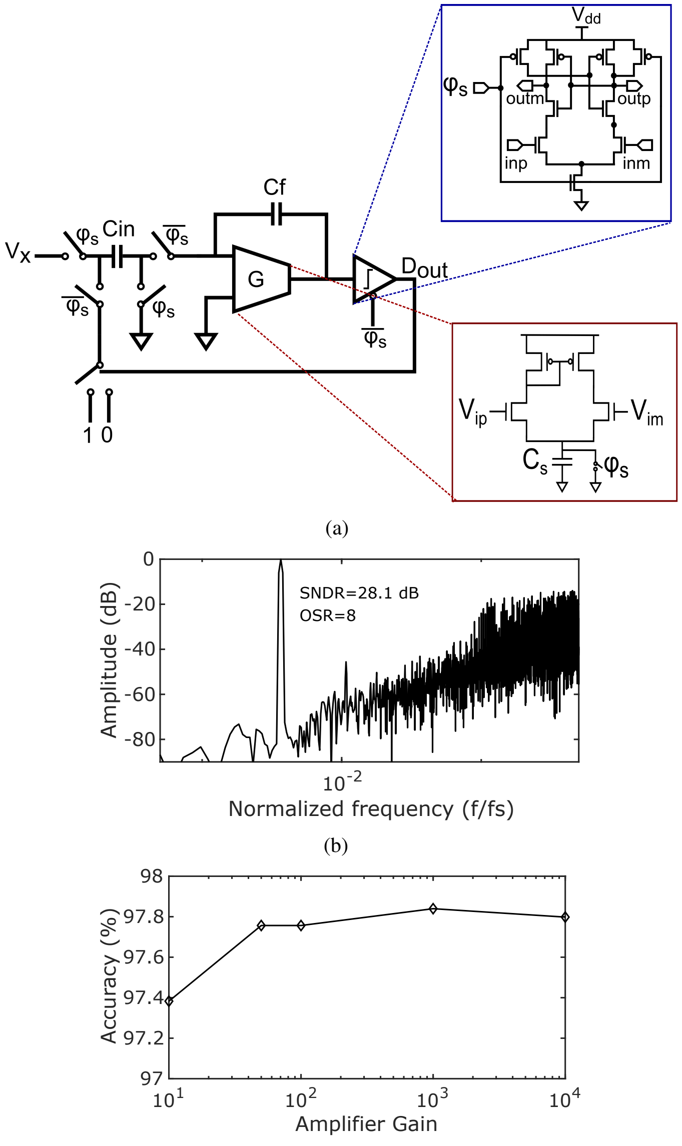
Fig. 3a) shows schematic of the input 1-bit ΔΣ modulator. The output ΔΣ modulator has the same architecture with the exception of differential inputs. The $\Delta\Sigma$ modulator is fully dynamic to reduce power consumption. A first-order loop filter is used in this work since the SRAM macro does not need a very high precision. The amplifier needs only a moderate gain since the 1-bit quantizer only uses the sign information of its inputs rather than amplitude to make decisions. The trade-off with moderate gain of the amplifier is weak in-band high-pass quantization noise shaping performance. This is reflected in the transfer function of the $\Delta\Sigma$ modulator given by

$$D_{out} = \frac{z^{-1}V_x + \left(1+2/G-(1+1/G)z^{-1}\right)Q}{1+2/G-z^{-1}/G} \tag{3}$$

where V_x is input to the $\Delta\Sigma$, Q is quantization error and G is gain of the amplifier. The in-band quantization error is attenuated by $1/(G+1)$ which corresponds to -34dB for $G = 50$. A single-stage dynamic amplifier shown in Fig. 3a) is used due to the moderate gain requirements. The amplifier uses a capacitor as tail current source [8]. The capacitor is discharged during the $\Delta\Sigma$ sampling phase (ϕ_s) and provides bias current to the amplifier during the amplification phase ($\overline{\phi_s}$). During the amplification phase, voltage across the capacitor rises which reduces gate-to-source voltage of the input transistors and increases the open-loop voltage gain until the transistors enter sub-threshold where the amplifier gain is maximum. Fig. 3b) shows the frequency response of the $\Delta\Sigma$ modulator. The $\Delta\Sigma$ modulator has an SNDR of 28.1dB at over-sampling ratio (OSR) of 8. Fig. 3c) shows the classification accuracy on the HAR dataset as a function of amplifier gain G at OSR of 8. The classification accuracy increases by only 0.4% as G is varied from 10 to 10000 thus demonstrating low dependency of accuracy on amplifier gain thanks to the 1-bit $\Delta\Sigma$ architecture. G is set to 50 in this work.

III. Measurement Results

The test-chip is fabricated in 65nm process and the die photograph is shown in Fig. 4. The core circuits occupy an area of 0.1mm^2 with the 64x64 SRAM array occupying an area of 0.03mm^2. The test-chip operates from a supply voltage of 0.5V-1.2V for the SRAM array and 1.2V for the other circuits. The operating speed of the entire macro is 325kHz which is limited by buffers driving the sampling capacitor in the output $\Delta\Sigma$. The $\Delta\Sigma$ modulators and clock generator consume 1.8μW from 1.2V supply while the SRAM array consumes 0.6μW-3.6μW from 0.5V-1.2V supply. Offset in the input and output $\Delta\Sigma$ modulators are calibrated once in the foreground before characterization of the complete macro. Performance of the macro is summarized in the table in Fig. 4 and benchmarked on the HAR dataset using a 2-layer artificial neural network (ANN) in which the input layer has 60 neurons, the hidden layer has 50 neurons and the output layer has 5 neurons. The hidden layer uses tanh activation and the output layer uses softmax activation. The proposed macro achieves 96.6% accuracy at 0.5V SRAM supply voltage and OSR of 4 and 97.8% accuracy at 1.2V SRAM supply voltage and OSR of 8. The macro consumes 16.2fJ at 1.2V supply. The energy consumption for each classification is 420.5pJ at 1.2V supply that does not include energy consumption of tanh/sigmoid activation functions and decimation filter.



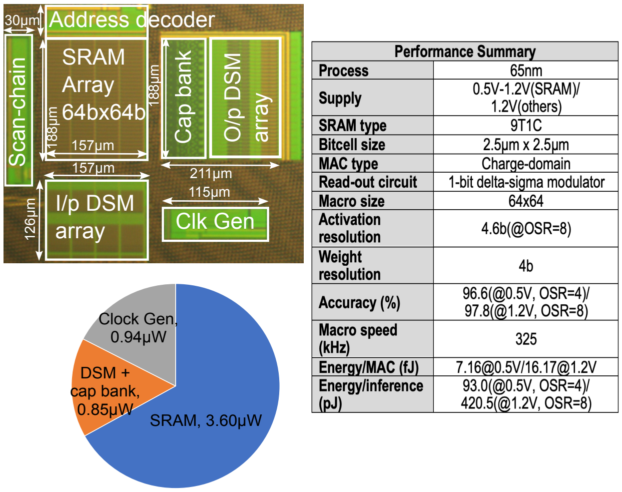
(a)

(b)

(c)

Fig. 3: a) Schematic of 1-bit $\Delta\Sigma$ modulator b) FFT plot of $\Delta\Sigma$ modulator with sinusoidal input c) accuracy vs amplifier gain at OSR of 8.

Fig. 5a) plots the measured root-mean-squared-error (RMSE) of dot-product that varies between 1.16mV at 0.5V SRAM power supply to 1.49mV at 1.2V SRAM power supply. Linearity of the dot product is measured by the ratio of maximum RBL swing to worst-case RMSE (Δ RBL/RMSE) which varies from 40.5 at 0.5V to 68.8 at 1.2V SRAM power supply. The RBL swing is obtained by spatial averaging of bitcells in a column and and the output $\Delta\Sigma$ bitstreams 1M times. Nonlinearity in the RBL transfer curve is due to static random mismatch between individual bitcells and



Performance Summary	
Process	65nm
Supply	0.5V-1.2V(SRAM)/ 1.2V(others)
SRAM type	9T1C
Bitcell size	2.5µm x 2.5µm
MAC type	Charge-domain
Read-out circuit	1-bit delta-sigma modulator
Macro size	64x64
Activation resolution	4.6b(@OSR=8)
Weight resolution	4b
Accuracy (%)	96.6(@0.5V, OSR=4)/ 97.8(@1.2V, OSR=8)
Macro speed (kHz)	325
Energy/MAC (fJ)	7.16@0.5V/16.17@1.2V
Energy/inference (pJ)	93.0(@0.5V, OSR=4)/ 420.5(@1.2V, OSR=8)

Fig. 4: Die micro-photograph, power breakdown and performance summary

capacitors, as well as charge-injection error in the switched-capacitor circuits. In comparison to this work, Δ RBL/RMSE varies from 12.6 [9] to 44 [5] for state-of-the-art SRAM CIM macros which demonstrate better linearity of the proposed architecture.

The HAR dataset has 24075 observations of five activities and 60 features extracted from acceleration data measured by smartphone accelerometer sensors. The dataset is randomly partitioned into 90% split for training and 10% split for testing. The measured accuracy changes from 96.6% at 0.5V and OSR of 4 to 97.8% at 1.2V and OSR of 8 compared to software baseline accuracy of 98.2% as shown in Fig. 5b). Fig. 5 c) shows the measured confusion matrices on the test chip at SRAM supply voltages of 0.5V and 1.2V and OSR of 4 and 8 with classes 1 through 5 corresponding to sitting, standing, walking, running and dancing respectively. Table I compares this work with state-of-the-art software HAR classifiers. The proposed 2-layer ANN classifier achieves competitive accuracy as state-of-the-art.

TABLE I: Comparison with state-of-the-art HAR works

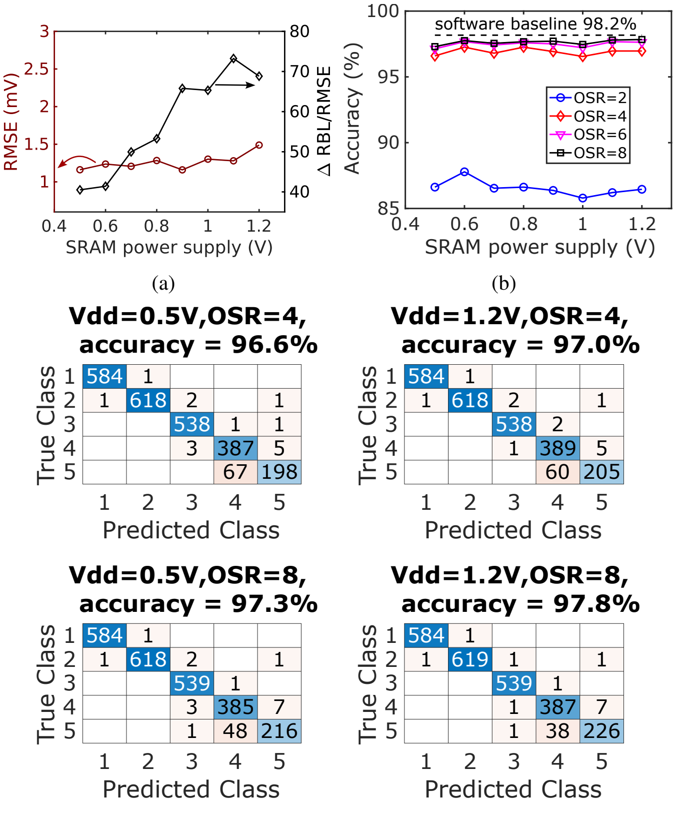
	This work	[10]	[11]	[12]	[13]	[14]
Model	ANN	Ada-boost		Bayesian		CNN
Accuracy	96.6-97.8%	98%	89.2%	91.4%	89.2%	95.7%
Class #	5	5	5	9	17	6

IV. Conclusion

This work has presented a ΔΣ based SRAM CIM macro that uses switched-capacitor circuits for charge-domain matrix multiplications with high linearity. The proposed macro is used for HAR task and achieves state-of-the-art accuracy while consuming only 420.5pJ/classification.

Acknowledgment

This work is partially supported by NSF grant CCF-1948331.



(a)

(b)

Vdd=0.5V,OSR=4, accuracy = 96.6%

True Class \ Predicted Class	1	2	3	4	5
1	584	1			
2	1	618	2		1
3			538	1	1
4			3	387	5
5				67	198

Vdd=1.2V,OSR=4, accuracy = 97.0%

True Class \ Predicted Class	1	2	3	4	5
1	584	1			
2	1	618	2		1
3			538	2	
4			1	389	5
5				60	205

Vdd=0.5V,OSR=8, accuracy = 97.3%

True Class \ Predicted Class	1	2	3	4	5
1	584	1			
2	1	618	2		1
3			539	1	
4			3	385	7
5			1	48	216

Vdd=1.2V,OSR=8, accuracy = 97.8%

True Class \ Predicted Class	1	2	3	4	5
1	584	1			
2	1	619	1		1
3			539	1	
4			1	387	7
5			1	38	226

(c)

Fig. 5: a) Measured RMSE and linearity of dot-product b) accuracy of software baseline and measured accuracy as a function of OSR and SRAM supply voltage c) confusion matrices for different supply voltages and OSR

References

[1] Y. Luo, K.-H. Teng, Y. Li, W. Mao, Y. Lian, and C.-H. Heng, “A 74-μW 11-Mb/s wireless vital signs monitoring SoC for three-lead ECG, respiration rate, and body temperature,” *IEEE transactions on biomedical circuits and systems*, vol. 13, no. 5, pp. 907–917, 2019.

[2] X. Zhang, Z. Zhang, Y. Li, C. Liu, Y. X. Guo, and Y. Lian, “A 2.89-μW clockless wireless dry-electrode ECG SoC for wearable sensors,” in *IEEE Asian Solid-State Circuits Conference (A-SSCC)*, 2015, pp. 1–4.

[3] B. P. Lo, H. Ip, and G.-Z. Yang, “Transforming health care: body sensor networks, wearables, and the internet of things,” 2016.

[4] C. Yu, T. Yoo, T. T.-H. Kim, K. C. T. Chuan, and B. Kim, “A 16K current-based 8T SRAM compute-in-memory macro with decoupled read/write and 1-5bit column ADC,” in *IEEE Custom Integrated Circuits Conference (CICC)*, 2020, pp. 1–4.

[5] Z. Jiang, S. Yin, J.-S. Seo, and M. Seok, “C3SRAM: An in-memory-computing SRAM macro based on robust capacitive coupling computing mechanism,” *IEEE Journal of Solid-State Circuits*, vol. 55, no. 7, pp. 1888–1897, 2020.

[6] Q. Dong, M. E. Sinangil, B. Erbagci, D. Sun, W.-S. Khwa, H.-J. Liao, Y. Wang, and J. Chang, “A 351TOPS/W and 372.4 GOPS compute-in-memory SRAM macro in 7nm FinFET CMOS for machine-learning applications,” in *IEEE International Solid-State Circuits Conference-(ISSCC)*, 2020, pp. 242–244.

[7] A. Biswas and A. P. Chandrakasan, “Conv-RAM: An energy-efficient SRAM with embedded convolution computation for low-power CNN-based machine learning applications,” in *IEEE International Solid-State Circuits Conference-(ISSCC)*, 2018, pp. 488–490.

[8] B. J. Hosticka, “Dynamic CMOS amplifiers,” *IEEE Journal of Solid-State Circuits*, vol. 15, no. 5, pp. 881–886, 1980.

[9] X. Si, J.-J. Chen, Y.-N. Tu, W.-H. Huang, J.-H. Wang, Y.-C. Chiu, W.-C. Wei, S.-Y. Wu, X. Sun, R. Liu *et al.*, “A twin-8T SRAM computation-in-memory unit-macro for multibit CNN-based AI edge processors,” *IEEE Journal of Solid-State Circuits*, vol. 55, no. 1, pp. 189–202, 2019.

[10] P. Li, Y. Wang, Y. Tian, T.-S. Zhou, and J.-s. Li, “An automatic user-adapted physical activity classification method using smartphones,” *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 3, pp. 706–714, 2016.

[11] P. Mayer, M. Magno, and L. Benini, “Energy-Positive Activity Recognition-From Kinetic Energy Harvesting to Smart Self-Sustainable Wearable Devices,” *IEEE Transactions on Biomedical Circuits and Systems*, vol. 15, no. 5, pp. 926–937, 2021.

[12] M. Forouzanfar, M. Mabrouk, S. Rajan, M. Bolic, H. R. Dajani, and V. Z. Groza, “Event recognition for contactless activity monitoring using phase-modulated continuous wave radar,” *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 2, pp. 479–491, 2016.

[13] M. S. Totty and E. Wade, “Muscle activation and inertial motion data for noninvasive classification of activities of daily living,” *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 5, pp. 1069–1076, 2017.

[14] J. Huang, S. Lin, N. Wang, G. Dai, Y. Xie, and J. Zhou, “TSE-CNN: A two-stage end-to-end CNN for human activity recognition,” *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 292–299, 2019.