Delta-Sigma Modulator-Based Compute-in-Memory Neural Network with Analog Feature Extraction and Classification for Edge Sensors

Tushar Gupta*‡, Vasundhara Damodaran*‡, Yuan Liao†, Jae-sun Seo†, and Arindam Sanyal*
*School of Electrical, Computer and Energy Engineering, Arizona State University, AZ, USA.
†School of Electrical and Computer Engineering Department, Cornell Tech, NY, USA.
‡equal co-authors.
Email: tgupta39@asu.edu

2026 IEEE International Symposium on Circuits and Systems (ISCAS) | 979-8-3315-7769-8/26/$31.00 ©2026 IEEE | DOI: 10.1109/ISCAS66217.2026.11562729

***Abstract*—This work presents an on-chip feature extractor and classifier that reduces sensor transmission by >1000× through on-chip data analysis. The classifier includes reservoir computing (RC) layer for unsupervised feature extraction followed by an artificial neural network (ANN) layer for supervised classification. The RC layer is implemented using switched-capacitor circuits, while the ANN uses a delta-sigma modulator-based 9T1C compute-in-memory (CIM) macro that supports variable-resolution precision. A test chip, fabricated in a 28nm process, is demonstrated for stress and seismic event detection. The test-chip achieves mean classification accuracies of 97.0% for stress detection and 95.8% for seismic event detection, with energy consumption of 1.3nJ/classification for stress and 4.25nJ/classification for seismic events.**

***Index Terms*—Classifier, Compute-in-Memory, Reservoir Computing, SRAM, Switched Capacitor circuits**

I. Introduction

Edge sensors are soon expected to produce exabytes of data per second that will consume Mega-watts power just for transmission. While techniques such as compressive sensing or adaptive sampling exploit data sparsity to compress data [1], [2] and reduce transmission power, they generally reduce sensor energy only by $< 10\times$. In contrast, this work proposes a machine-learning (ML) approach to perform data analysis inside the sensor itself and transmit only classification results which reduces data transmission volume by two-to-three orders-of-magnitude and sensor energy by $> 50\times$ compared to naive transmission of digitized sensor data and by $> 19\times$ compared to compressive-sensing analog-to-digital converter (ADC) before transmission as shown in Fig. 1. The energy numbers in Fig. 1 are calculated for state-of-the-art ADCs and transmitter reported in [1], [3], [4].

A fully analog approach is adopted for the proposed ML circuit design to reduce power consumption. The key contributions of this work are – a) unsupervised reservoir-computing (RC) for extracting features from the input data directly in analog domain; b) supervised artificial neural network (ANN) as the output layer that uses delta-sigma modulator (DSM)-based SRAM compute-in-memory (CIM) with switched-capacitors. The proposed technique is demonstrated for detecting stress from electro-cardiogram (ECG) signals and for identifying earthquake events from seismic data and reduces data transmission by $4200\times$ and $1400\times$ respectively.

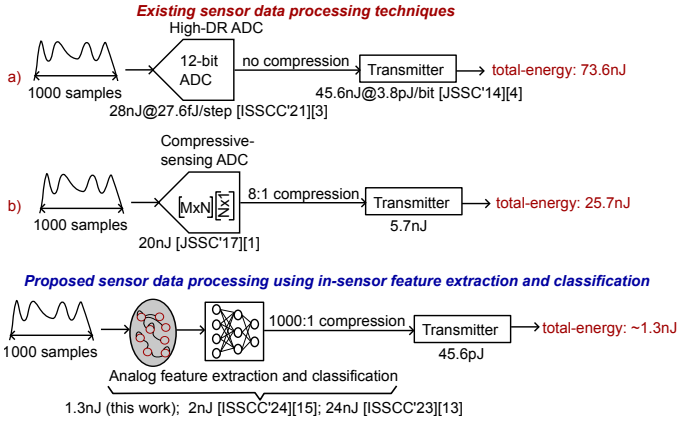

Fig. 1. Overview of the proposed in-sensor classification technique that addresses sensor data transmission problem and reduces sensor energy compared to existing techniques

II. Proposed Architecture

A. Comparison with prior silicon reservoir-computers

RC is a popular computing paradigm that uses static nonlinearity to project the input signal to a high-dimensional space, facilitating easier separation of input classes by a supervised output layer following the reservoir layer. Recurrent connections between the reservoir neurons impart properties of infinite-impulse response (IIR) filters to the reservoir neurons which makes their state-space infinite dimensional and allows RC to extract features from the input data with a small number of physical neurons. The weights in the input and reservoir layers are drawn from static, random distributions enabling unsupervised feature extraction, minimizing training requirements. Since RC uses nonlinearity for input projection, the otherwise undesired nonlinearities in analog circuits become useful attributes and reduces energy and area requirements of the circuits by trading-off power and area for linearity. This is a key motivation for adopting RC for analog feature extraction in

this work. The silicon RC realizations in [5], [6] require either large capacitors to realize biological time-constants which is energy-inefficient, or background calibration for analog delay elements or nonlinearity element. An on-chip analog RC in [7] used time-multiplexing to create multiple virtual neurons from a single physical neuron without requiring large capacitors or calibration. However, it relied on always-on continuous-time circuits which are not power efficient for sensor applications, and did not realize the output layer in silicon. The key differentiation of this work over [7] is the adoption of switched-capacitor architecture for reservoir neurons which reduced energy consumption by $20\times$ and silicon implementation of output layer. The work in [8] presented silicon RC and output layers but did not integrate them on a single die, used continuous-time circuits for RC and binary-weighted capacitors to encode model weights in the output layer which resulted in $8\times$ higher energy consumption than this work.

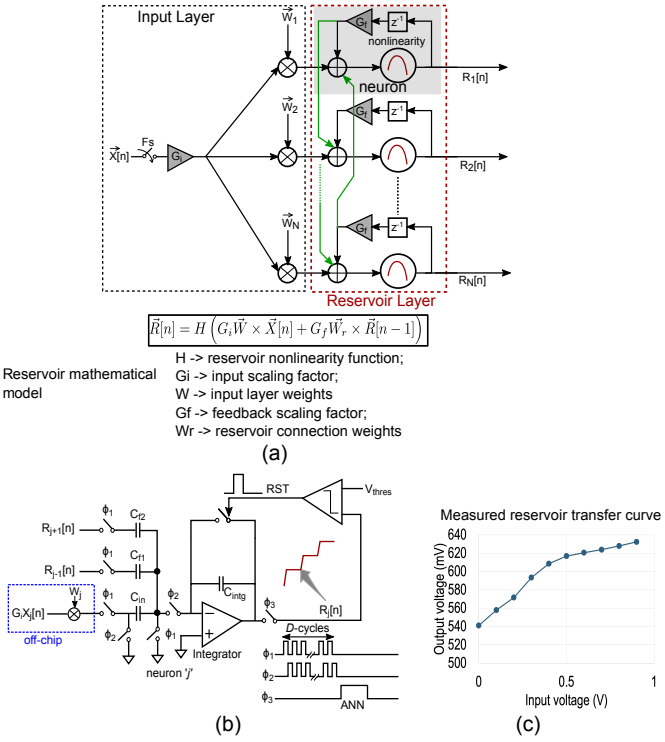

Fig. 2. a) Reservoir computing circuit schematic, b) switched-capacitor implementation of the reservoir neuron c) measured transfer curve of a reservoir neuron.

B. Proposed Reservoir Computer Layer

Fig. 2 shows circuit schematics of the proposed RC and the reservoir neuron implementation using switched-capacitor. The matrix multiplications are performed in the charge domain, and partial results are stored on feedback capacitors. The recurrent neural network behavior is realized in circuit using a leaky integrator with a low-gain amplifier (gain ≈ 25). The integrator resets if its output exceeds a threshold voltage. Each reservoir neuron is connected to two neighboring neurons. The reservoir has built-in lateral inhibition which resets all three neurons if one of the neurons is reset and detects transients in the input signal. The input weight matrix W is binarized and derived from random distribution. To test the chip with different input sample sizes D across different dataset, multiplication of the input samples with W is performed off-chip. Input and feedback scaling factors are set by ratios of sampling capacitors (C_{in}, C_{f1}, and C_{f2}) to feedback capacitor (C_{intg}), with sampling capacitors of 14.8fF. The circuit non-linearities are absorbed into the reservoir dynamics through its non-linear function $H(\cdot)$ (Fig. 2). Mismatches in the reservoir result in chip-to-chip variation which are addressed by re-training each chip. The key differentiation of this work over [7] are – a) adoption of switched-capacitor architecture for reservoir neurons which reduced energy consumption by $20\times$, and b) silicon implementation of output layer and integration with RC.

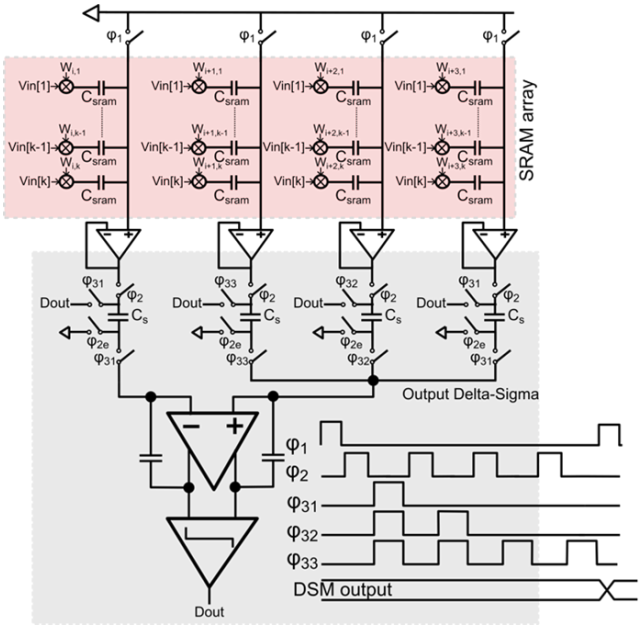

Fig. 3. Slice of SRAM macro with 1-bit DSM and 4-bit weights.

C. Proposed DSM-SRAM output layer

Fig. 3 shows schematic of a 4-bit slice of the proposed SRAM-CIM macro that uses 1-bit DSMs to convert input and output activations into binary pulses. DSM output from the last layer is digitally decimated using low-pass filters. Decimation is not needed in intermediate layers. Use of DSM improves linearity of matrix-multiplication since each SRAM cell sees either '0/1' input, reduces quantization error compared to [9] and allows re-configuring precision of input and output activations by simply changing oversampling ratio (OSR) without any change in hardware. Additionally, output DSM allows mapping higher-precision weights (e.g., 8-bits) onto the macro and easy combination of the partial results at the output through bit alignment. As in [10], 9T1C SRAM cells are used for charge-domain multiplications of input

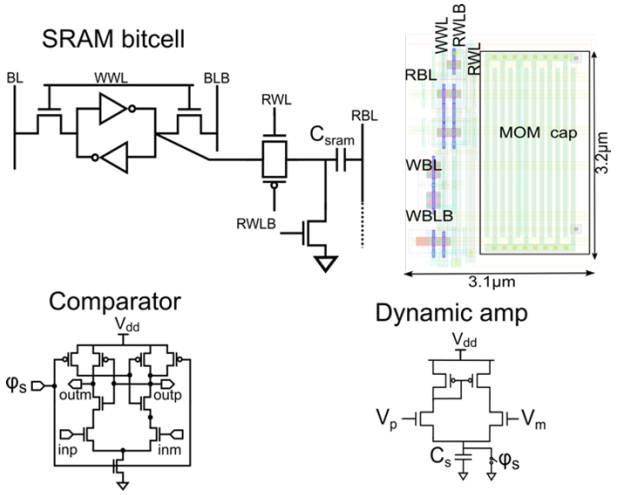

Fig. 4. 9T1C SRAM cell schematic, layout and circuit diagrams of the comparator and dynamic amplifier.

activations and model weights implemented on consecutive columns (Fig. 4). The key difference with [10] is that instead of passive combination of charge on the RBLs using additional compensation and compute capacitors, this work uses active charge transfer and combination which improves signal swing of the output activations. While [11] also uses DSM concept for CIM, a separate non-reconfigurable ADC is used for digitizing output activation. During Φ_1, capacitors are discharged, while in Φ_2, MAC outputs from RBLs are sampled into the output DSM. To combine the RBLs with appropriate binary weights, multi-cycle charge transfer (Φ_{31}, Φ_{32}, Φ_{33}) is performed, e.g., the RBL voltage corresponding to weight of 2 is sampled twice on the feedback cap (Φ_{32}). Active analog unity-gain buffers are used to sample RBL voltage on the corresponding DSM input capacitor. After the multi-cycle charge-transfer, a strong-arm latch performs 1-bit quantization.

III. Measurement Results

The test-chip is fabricated in 28nm CMOS (Fig. 5) and occupies an area of 0.1 mm^2. The reservoir consumes 7.2μW at 31.25kHz while the ANN macro with 30×32-bit SRAM consumes 12.3μW at 312.5kHz for an OSR of 10 (Fig. 6). Fig. 6 also shows the in-the-loop training of test-chips that suppresses chip-to-chip variations due to mismatches. 3 chips are trained and tested for stress detection from ECG samples obtained from wrist-worn sensor data [12] and seismic event detection obtained from proprietary dataset. Fig. 7 shows accuracy of chip 1 versus supply voltage and for 150 repeated evaluations at OSR=10 on stress dataset. Compared to [7] , the proposed analog RC achieves 2× better energy-efficiency as shown in Table III. The proposed analog neural network achieves 1:4200× compression for stress data and 1:1400× compression for seismic data (Table IV). Table V compares this work with state-of-the-art on-chip mixed-signal feature extractor and classifiers. This work consumes lower area and achieves competitive energy-efficiency as state-of-the-art for similar accuracies thanks to the RC that performs analog feature extraction with a small area and power consumption.

Table I summarizes these classification accuracies for stress dataset for 3 test-chips for different OSR. The maximum accuracy obtained for stress data is 97.7% and average accuracy across all chips is 97.0% with a standard deviation of 1.3% at an OSR of 10. We also conducted cross-chip testing by training on one chip and testing on the others using the trained model. The average accuracy drops by 3% when only cross-chip training data is considered, and training each chip individually reduces chip-to-chip accuracy variation by 3×. Similarly, Table II shows seismic dataset accuracies for 3 test-chips for different OSR. The average accuracy of seismic event detection is 95.8% at OSR=10 and training each chip individually reduces chip-to-chip variation by 4×.

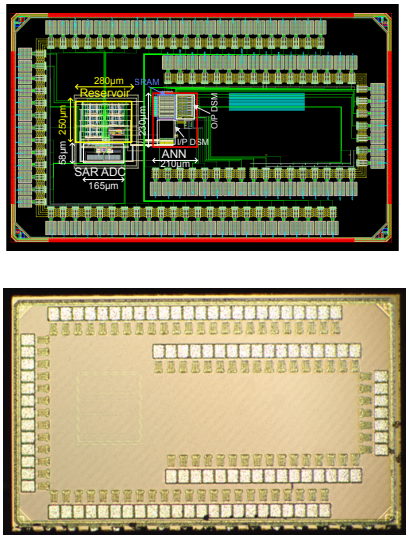

Fig. 5. Layout and die photo of 28nm test-chip

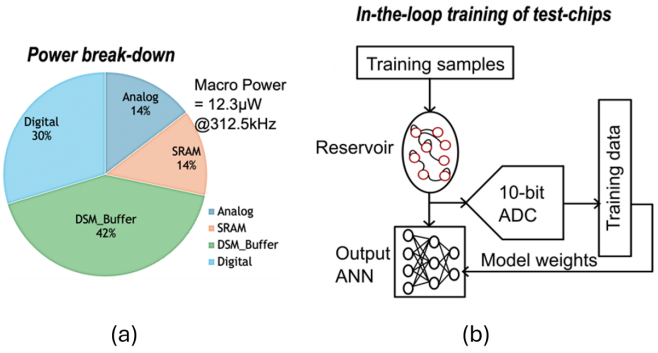

Fig. 6. a) Power break-down of the test-chip b) in-the-loop training of test-chips for suppressing chip-to-chip mismatches.

A. Conclusion

This paper presented an on-chip machine learning technique using analog circuits for in-sensor classification that

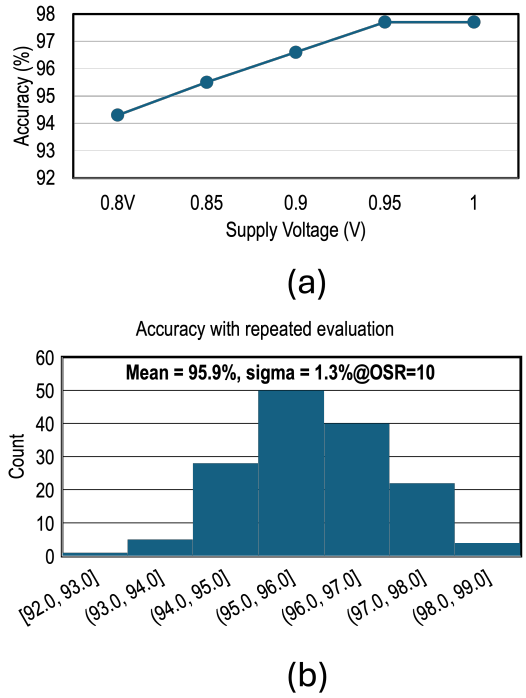

Fig. 7. a) Measured accuracy vs supply voltage and b) for 150 repeated evaluations at 0.9V for chip 1 on stress dataset at OSR=10.

TABLE I
CLASSIFICATION ACCURACY SHOWING SAME-CHIP AND CROSS-CHIP TRAINING FOR STRESS DATASET

Training chip	Testing chip	OSR=10	OSR=15	OSR=20
Chip1	**Chip1**	96.6	96.6	96.6
	Chip2	95.5	95.5	95.5
	Chip3	93.2	94.3	94.3
Chip2	**Chip1**	93.6	93.6	93.6
	Chip2	96.8	96.8	96.8
	Chip3	96.8	96.8	93.6
Chip3	**Chip1**	96.6	96.6	96.6
	Chip2	93.2	93.2	93.2
	Chip3	97.7	97.7	97.7

TABLE II
CLASSIFICATION ACCURACY SHOWING SAME-CHIP AND CROSS-CHIP TRAINING FOR SEISMIC DATASET

Training chip	Testing chip	OSR=10	OSR=15	OSR=20
Chip1	**Chip1**	95.9	95.9	95.9
	Chip2	92.7	92.7	92.7
	Chip3	93.3	93.3	93.3
Chip2	**Chip1**	92.1	92.1	92.1
	Chip2	95.3	95.4	95.6
	Chip3	92.1	92.1	92.1
Chip3	**Chip1**	93.0	93.0	93.0
	Chip2	92.5	92.4	92.5
	Chip3	96.1	96.7	96.9

reduces sensor data transmission. The proposed circuits are designed using dynamic, switched-capacitor circuits that benefit from technology scaling and result in state-of-the-art energy-efficiency.

TABLE III
COMPARISON WITH STATE-OF-THE-ART RESERVOIR COMPUTERS

	[5]	[6]	[7]	**This Work**
Process(nm)	17	350	65	**28**
RC neuron type	spiking	Ring Oscillator	Continuous-time circuits	**Switched-cap**
# of neurons	1024	128	50	**30**
Speed (kHz)	-	0.05	40	**31.25**
Power (uW)	722	0.4	300	**7.2***
Energy/MAC (pJ)	17	5.2	50	**2.5****

*includes ADC power
**each reservoir neuron is considered to perform 3 MAC

TABLE IV
SUMMARY OF TRANSMISSION DATA COMPRESSION

Classification Task	Stress detection	Seismic events
RC neurons	20	30
ANN structure	30-1	20-1
Tx data compression ratio	1:4200	1:1400
Classification accuracy*	97.0%	95.8%

**average of 3 chips at OSR=10*

TABLE V
COMPARISON WITH STATE-OF-THE-ART ON-CHIP MIXED-SIGNAL CLASSIFIERS

	ISSCC'22 [13]	ISSCC'23 [14]	JSSC'19 [15]	ISSCC'24 [16]	**This work**	
Process (nm)	28	28	180	180	**28**	
Classifier	TD-CNN +BNN	Adaptive sample+ RNN	BNN	QTD-CNN +1-shot BNN	**Reservoir computer + ANN**	
Feature Extractor	Time-domain		Analog	Time-domain	**Analog**	
Classification task	Voice acitivity	Keyword spotiing	Voice acitvity	Arrythmia detection	**Seismic events**	**Stress detection**
# of classes	2	5	2	2	**2**	**2**
Classification speed	100Hz	60Hz	100Hz	2kHz	**312.5kHz**	
Accuracy (%)	90.1	92.8	84	98.9	**95.8***	**97.0***
Energy/class (nJ)	1.08	24	100	2	**4.25***	**1.3***
Chip area (mm^2)	0.8	0.8	2.1	2.9	**0.1**	

*at OSR of 10

ACKNOWLEDGMENT

This work is supported by Sony Research Award Program, Alphacore Inc.under DoE grant number DE-SC0022923 and SRC through task #3160.033.

REFERENCES

[1] W. Guo, Y. Kim, A. H. Tewfik, and N. Sun, "A Fully Passive Compressive Sensing SAR ADC for Low-Power Wireless Sensors," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 8, pp. 2154–2167, 2017.

[2] S. Mondal, C.-L. Hsu, R. Jafari, and D. A. Hall, "A Dynamically Reconfigurable ECG Analog Front-End With a 2.5× Data-Dependent Power Reduction," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 15, no. 5, pp. 1066–1078, 2021.

[3] C. Lo, J. Lee, Y. Lim, Y. Yoon, H. Hwang, J. Lee, M. Choi, M. Lee, S. Oh, and J. Lee, "A 116μ W 104.4 dB-DR 100.6 dB-SNDR CT Δ audio ADC using tri-level current-steering DAC with gate-leakage

compensated off-transistor-based bias noise filter," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 64. IEEE, 2021, pp. 164–166.

[4] P. P. Mercier, S. Bandyopadhyay, A. C. Lysaght, K. M. Stankovic, and A. P. Chandrakasan, "A sub-nW 2.4 GHz transmitter for low data-rate sensing applications," *IEEE journal of solid-state circuits*, vol. 49, no. 7, pp. 1463–1474, 2014.

[5] F. C. Bauer, D. R. Muir, and G. Indiveri, "Real-Time Ultra-Low Power ECG Anomaly Detection Using an Event-Driven Neuromorphic Processor," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, no. 6, pp. 1575–1582, 2019.

[6] Y. Chen, E. Yao, and A. Basu, "A 128-Channel Extreme Learning Machine-Based Neural Decoder for Brain Machine Interfaces," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 3, pp. 679–692, 2016.

[7] S. T. Chandrasekaran, I. Banerjee, and A. Sanyal, "7.5nJ/inference CMOS Echo State Network for Coronary Heart Disease prediction," in *ESSDERC 2021 - IEEE 51st European Solid-State Device Research Conference (ESSDERC)*, 2021, pp. 103–106.

[8] S. Sadasivuni, S. P. Bhanushali, I. Banerjee, and A. Sanyal, "In-sensor neural network for high energy efficiency analog-to-information conversion," *Scientific reports*, vol. 12, no. 1, p. 18253, 2022.

[9] Q. Dong, M. E. Sinangil, B. Erbagci, D. Sun, W.-S. Khwa, H.-J. Liao, Y. Wang, and J. Chang, "A 351TOPS/W and 372.4GOPS Compute-in-Memory SRAM Macro in 7nm FinFET CMOS for Machine-Learning Applications," in *2020 IEEE International Solid-State Circuits Conference - (ISSCC)*, 2020, pp. 242–244.

[10] V. Damodaran, Z. Liu, J.-S. Seo, and A. Sanyal, "A 138-TOPS/W Delta-Sigma Modulator-Based Variable- Resolution Activation in-Memory Computing Macro," in *2023 IEEE Custom Integrated Circuits Conference (CICC)*, 2023, pp. 1–2.

[11] P. Chen, M. Wu, W. Zhao, J. Cui, Z. Wang, Y. Zhang, Q. Wang, J. Ru, L. Shen, T. Jia *et al.*, "A 22nm Delta-Sigma Computing-In-Memory (Δ CIM) SRAM Macro with Near-Zero-Mean Outputs and LSB-First ADCs Achieving 21.38 TOPS/W for 8b-MAC Edge AI Processing." in *ISSCC*, 2023, pp. 140–141.

[12] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ser. ICMI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 400–408.

[13] F. Chen, K.-F. Un, W.-H. Yu, P.-I. Mak, and R. P. Martins, "A 108nW 0.8mm2 Analog Voice Activity Detector (VAD) Featuring a Time-Domain CNN as a Programmable Feature Extractor and a Sparsity-Aware Computational Scheme in 28nm CMOS," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 65, 2022, pp. 1–3.

[14] J.-H. SeoI, H. Yang, R. Rothe, Z. Fan, Q. Zhang, H.-S. Kim, D. Blaauw, and D. Sylvester, "A 1.5μW End-to-End Keyword Spotting SoC with Content-Adaptive Frame Sub-Sampling and Fast-Settling Analog Frontend," in *2023 IEEE International Solid-State Circuits Conference (ISSCC)*, 2023, pp. 1–3.

[15] M. Yang, C.-H. Yeh, Y. Zhou, J. P. Cerqueira, A. A. Lazar, and M. Seok, "Design of an Always-On Deep Neural Network-Based 1- μ W Voice Activity Detector Aided With a Customized Software Model for Analog Feature Extraction," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 6, pp. 1764–1777, 2019.

[16] J. Cho, Y. J. Pyeon, J. Yeom, H. Kim, S. Cho, Y. Kim, T. Kim, J.-H. Kwak, G. Choi, Y. Lee, H. Shin, H. E. Jeong, and J. J. Kim, "33.7 An Adhesive Interposer-Based Reconfigurable Multi-Sensor Patch Interface with On-Chip Application Tunable Time-Domain Feature Extraction," in *2024 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 67, 2024, pp. 554–556.