2024 IEEE Biomedical Circuits and Systems Conference (BioCAS) | 979-8-3503-5495-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/BioCAS61083.2024.10798153

MulPi: A *Mul*ti-Class and *P*atient-*I*ndependent Computing-in-SRAM Seizure Classifier

Bokyung Kim*, Qijia Huang†, Brady Taylor†, Qilin Zheng†, Jonathan Ku†, Nicky Ramos†, Eric Yeats†, Yiran Chen†, and Hai "Helen" Li†

* Department of Electrical and Computer Engineering, Rutgers University, NJ 08854, USA

† Department of Electrical and Computer Engineering, Duke University, NC 27708, USA

Email: bk.kim@rutgers.edu (corresponding author)

***Abstract*—Authentic detection and prediction of seizures require 1) multi-class (*Mul*) and 2) patient-independent (*Pi*) classification. Recent implementable chips for seizure classification rarely satisfy the two requirements due to restricted resources in small chips; therefore, high efficiency is imperative along with accuracy. This paper introduces an efficient *MulPi* chip, fabricated for the first time to simultaneously fulfill multi-class and patient independence, based on a co-design approach. We develop a 5-layer convolutional neural network (CNN), *MulPiCNN*, with advanced training techniques for lightness and accuracy. At the hardware level, our SRAM-based chip leverages computing-in-memory (CIM) for efficiency. The fabricated *MulPi* chip is distinguished from prior CIMs in two folds, namely *ISRW-CIM*: a) input-stationary (IS) CIM for resource-saving, and b) row-wise (RW) computing to address a challenge of SRAM CIM, empowered by our novel 2T-Hadamard product unit (HPU). *MulPi* outperforms state-of-the-art chips with 98.5% sensitivity and 99.2% specificity, classifying in 0.12s and 0.348mm^2.**

***Index Terms*—Seizure classification, computing-in-memory, diagnosis automation, input-stationary dataflow, SRAM CIM**

I. Introduction

Epilepsy patients over 50 million across the world suffer from unexpected seizures in their daily lives. Most patients up to 70% could live seizure-free with timely treatment, which necessitates instant and accurate diagnosis. Accordingly, implementable chips for automated diagnosis have been investigated to replace labor-expensive and time-consuming visual assessment or inaccurate and inconvenient wearable devices [12]. State-of-the-art chips attempt to employ machine learning (ML) techniques for high accuracy.

However, previous chips have limitations: binary classification and patient-specific designs as shown in Fig. 1. Firstly, predominant chips have adopted traditional ML binary algorithms for detection because of restricted resources [4], [9], [14], [16]. Unfortunately, patients still need to go through seizures until treatment takes effect and are exposed to danger unless they are in a secure place before the occurrence of unexpected seizures. Since seizure has at least three different states (inter-ical, pre-ical, and ictal) in practice according to medical definitions, a recent chip [5] proposes to predict seizure by recognizing the pre-ictal period. But for the sake of randomness of the differed states, multi-class assorting is necessary to discern each state in chip design, as indicated in algorithm-level exploration [1]. Second, prior arts have implemented ML models tailored for specific patients. These patient-specific models increase resources like area and power or even need additional fabrication of different chips to apply to other patients. Instead, patient independence (one model for all patients) is essential for economic chip costs [10].

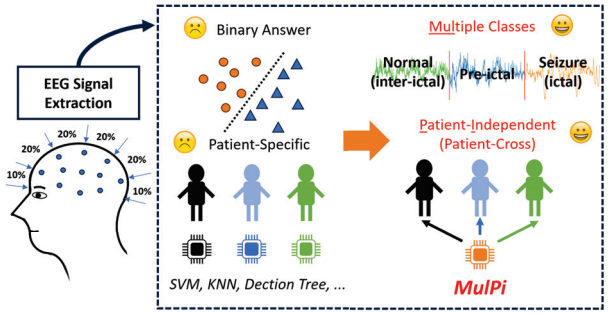


Fig. 1. Contribution of the proposed *MulPi* chip, differentiating from prior chips of binary and/or patient-specific seizure classification.

The two requirements for seizure classification demand a more complex algorithm leveraging deep learning techniques than traditional ML models. Convolutional neural networks (CNNs) have been experimented with this application, boasting high accuracy [6] and even patient independence [10]. However, the multi-class classification is still missing in previous chip designs because it needs further resources from a small chip. Hence, highly efficient design is imperative to enable the complex algorithm at the edge.

Driven by the limitations in prior chips, we propose an efficient multi-class and patient-independent (*MulPi*) classifier chip, for the first time to satisfy both concurrently, with algorithm and hardware co-design. At the algorithm level, we introduce a new one-dimensional (1D) light CNN model, *MulPiCNN*. While pursuing lightness with shallow layers and quantized bits in domain conversion, we attain high accuracy through training techniques.

Then, we fabricated an SRAM-based chip with TSMC 65nm technology to implement the developed algorithm. We leverage the emerging computing-in-memory (CIM) paradigm, which is energy-efficient for data transfer and parallel multiplications [11]. While previous CIMs keep weights in macros and accu-

This work is supported by the National Science Foundation (NSF) under IIS-2332744, CNS-2233808, CNS-2112562 and Rutgers University Startup Package for Prof. Bokyung Kim. The authors thank the Shared Materials Instrumentation Facility (SMIF) at Duke University for the die photo.

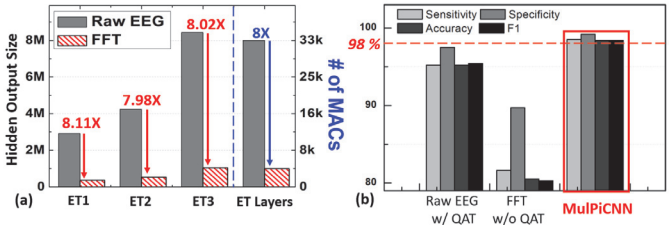


Fig. 2. (a) The reduced numbers in hidden output size from layers and MAC operations according to domain conversion. (b) Accuracy improvement by *MulPiCNN* keeping the reduced network size.

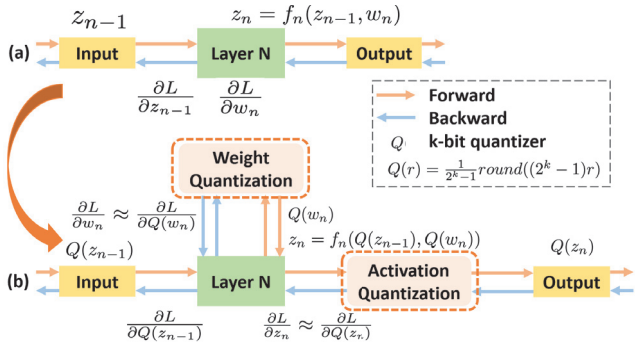


Fig. 3. (a) Typical training process and (b) the proposed training methodology to obtain parameters in *MulPiCNN* for high-performance inference.

mulate along bitlines, our chip proposes a cross-layer solution, *ISRW-CIM*, by a) forcing inputs to stay in macros (input-stationary, IS) at the architecture level for area efficiency and b) computing row-wise (RW), empowered by two-transistor-based Hadamard product unit (2T-HPU) at the circuit level, to avoid an SRAM-CIM challenge. *MulPi* outperforms state-of-the-art chips.

II. MulPiCNN Design

MulPiCNN leverages the power of CNN demonstrated in accurate classification, including EEG signals [6], [10]. For data preparation, we partition raw EEG signals into 4-second segments in a moving-window approach.

A. For Lightness

Table I summarizes the *MulPiCNN* architecture composed of three extracting (ET) and two flattened (FT) layers. With the shallow layers, we achieved *MulPi*, unlike prior CNNs of binary and patient-specific answers [6]. Considering the restricted environment, we applied a fast Fourier transform (FFT) to the 4-second EEG signals for domain conversion because data and multiply-accumulate (MAC) operations are reduced in the frequency domain, as shown in Fig. 2(a). The sampling rate for raw signals is 256Hz, and the frequency range for filtering noise is 0-128Hz. Also, data are quantized to 8 bits for compression, considering the hardware property.

B. For Accuracy

However, the transformed and quantized data aggravates the model accuracy, as Fig. 2(b) shows. Therefore, instead of typical learning in Fig. 3(a), we conceived a low-complex model with quantization-aware training (QAT) with straight-through estimator (STE), as illustrated in Fig. 3(b). QAT executes

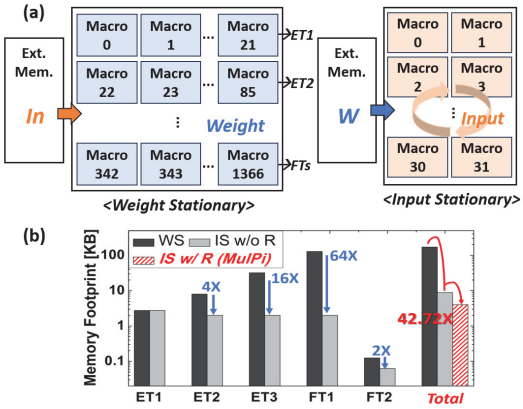


Fig. 4. (a) Comparison between weight- and input-stationary dataflow. (b) Layer-wise CIM memory array saving effect by input-stationary dataflow.

the forward pass using round-nearest quantized weights and activations [7]. The challenge arises from the computational difficulty of calculating gradients for backpropagation when weights and activations are discrete values. Our solution is STE to update the weights using gradients calculated from full-precision weight values, which provides a sufficiently accurate approximation [2]. As a result, *MulPiCNN* beats the model performance with raw EEG (Fig. 2(b)) and prior works exploiting deep learning, as shown in Section IV.

III. MulPi Chip with ISRW-CIM

The *MulPi* chip proposes a cross-layer solution, for high performance and efficiency across circuit and architecture levels. While we utilize CIM for efficiency [11], here are two design keys distinguished from prior CIMs: 1) IS for saving (architecture-level design) and 2) RW computing to tackle read disturbance with our novel low-cost 2T-HPU (circuit-level).

A. "IS", Architectural Design with Dataflow and Mapping

To our knowledge, previously fabricated CIM chips have been based on the weight-stationary (WS) dataflow by maintaining weights in CIM macros and supplying layer's inputs from external memories, as displayed in Fig. 4(a). Yet WS fundamentally requires a large area for weight parameters, over 1300 macros in our case despite the tiny size of *MulPiCNN*, because of the neural network feature, i.e., #weights > #inputs [8]. Therefore, we implement *MulPI* as an IS CIM, where weights are supplied from the outside, to reduce memory and

TABLE I
THE *MulPiCNN* ARCHITECTURE. IN: INPUT, W: WEIGHT KERNEL.

	In_size	#In_channels	W_size	Conv #	Stride	Padding
ET1	128	22	4	32	1	2
ReLU	128	32	-	-	-	-
MaxPool	128	32	2	-	2	0
ET2	64	32	4	64	1	2
ReLU	64	64	-	-	-	-
MaxPool	64	64	2	-	2	0
ET3	32	64	4	128	1	2
ReLU	32	128	-	-	-	-
MaxPool	32	128	2	-	2	0
Flatten	16	128	-	-	-	-
FT1	1	2048	2048	64	-	0
FT2	1	64	64	2	-	0

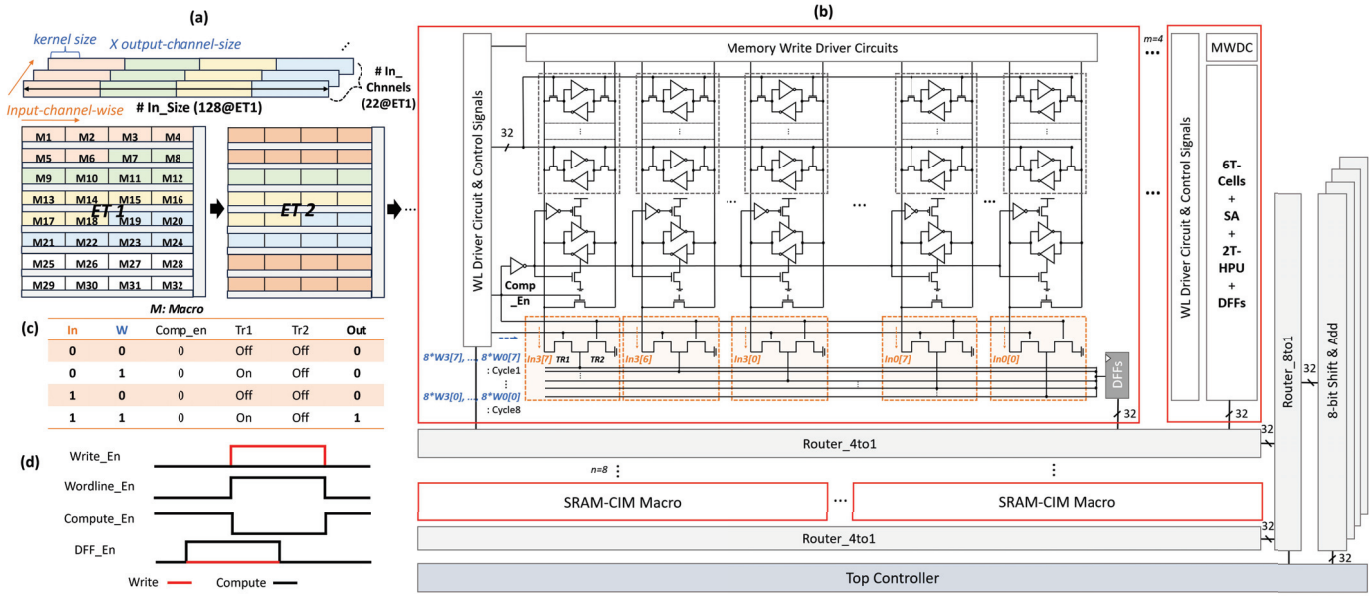


Fig. 5. (a) Channel-wise mapping for one-shot accumulation in each macro. (b) The proposed *MulPi* architecture of macros, mainly including 6T-cells, sense-amp, 2T-HPU in the orange box, and DFFs, and peripheries of routers and shift&adds, (c) operation results following the Hadamard product, and (d) timing diagram of control signals.

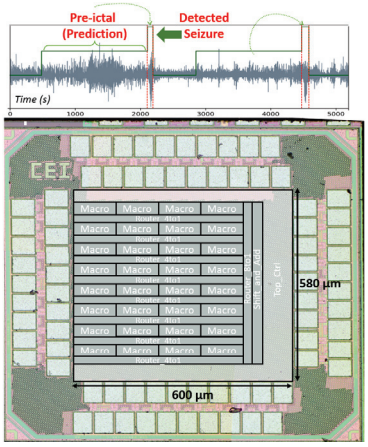
area consumption. As generated outputs including intermediate results are forwarded to CIM in IS, IS needs even fewer macros with the feasibility of recycling cells. IS CIM with recycling decreased the memory footprint by 42.7× and 2.2× than WS and IS without recycling, as compared in Fig. 4(b).

Fig. 5(a) shows the proposed mapping scheme to maximize the utilization rate of SRAM arrays in IS. Since inputs of 1D CNN are 2D and the total input size is larger than the total macro size, we should split and map 2D data into multiple macros. However, one macro in RW accumulates all partial results from one row in one shot, so naive mapping could cause inefficiency due to array under-utilization or redundant repentance of the same operations. Based on our observation that input-channel-wise accumulation always follows kernel-wise and output-channel-wise sum, we prioritize input-channel-wise data in our mapping, which enables the one-shot accumulation of each macro. We perform addition across kernel-sized subsequent rows by hopping the rows under the kernel window, repeating it as many times as the output channel size.

B. "RW", Macro and Circuit Design

Unlike emerging-memory-based CIM [11], SRAM-CIM has a representative issue with the prevalent column-wise (CS) computing, called read disturbance, which interferes with adjacent memory cell data. Rather than employing CS with stopgaps (e.g., under-driven wordline with slow operations [3] or adding transistors [15]), we enforce RW with a novel 2T-HPU circuit.

Fig. 5(b) provides the overview of *MulPi*, including macro details. The macro operates cells in rows corresponding to the kernel window sequentially to multiply the memory data (inputs) with supplied weights through 2T-HPU colored in orange. The 2T-HPU follows the outcomes of the truth table



Technology		TSMC 65nm 1p9m
Area	Chip	1.02 × 1.00mm²
	Core	0.60 × 0.58mm²
Core V_{DD}		0.6V – 1V
I/O		2.5V – 3.3V
SRAM		32Kb
Power		0.28mW @ 0.6V
Latency		0.12s
Frequency		100MHz
Precision		FXP8
Sensitivity		98.5%
Specificity		99.2%
Accuracy		98.4%
PPV		98.4%

Fig. 6. Classification results, which successfully predicts and detects seizures, a die photo, and chip performance summary.

in Fig. 5(c). Specifically, operands are given to TR1 in 2T-HPU, and COMP_EN (active_low) controls TR2. When TR2 is OFF, TR1 delivers V_d to the output according to V_g. TR2 is added to restrain reverse currents in TR1.

The final accumulation is performed based on the bit-slicing technique, where each bit of the corresponding weight is supplied per cycle and the partial outputs are summed through shift-and-add circuits. To reduce the number of peripheral circuitry sets, routers are used to deliver outputs and control signals between macro arrays and other circuits. Fig. 5(d) gives a simplified timing diagram for important control signals.

IV. Measurement and Results

The unique feature of this work is its ability to efficiently perform multi-class classification with highly accurate results across patients for authentic detection and prediction. We verified *MulPi* with the CHB-MIT dataset by the cross-patient

TABLE II

COMPARISON BETWEEN THE PROPOSED CHIP, *MulPi*, AND STATE-OF-THE-ART WORKS

(*LS: LEAST-SQUARES, *GTCA: GUIDED TIME-CHANNEL AVERAGING, -: NOT REPORTED, ∧ : ESTIMATED BASED ON THE REPORTED DATA).

	TBioCAS'21 [14]	JSSC'22 [4]		JSSC'22 [16]	ISSCC'22 [9]		JSSC'23 [5]	ISSCC'23 [10]	This Work
Dataset	CHB-MIT	UoM	CHB-MIT	CHB-MIT	IEEE.org	CHB-MIT	CHB-MIT	CHB-MIT	**CHB-MIT**
#Chnnels	8	8		16	256		-	22	**22**
Prediction	No	No		No	No		Yes	No	**Yes**
Classification	Binary	Binary		Binary	Binary		Binary	Binary	**Multi-Class**
Classifier	LS*-SVM	Logistic Regression		GTCA* -SVM	NeuralTree		SVM	SciCNN	**MulPiCNN**
Sensitivity [%]	97.8	97.9	97.5	97.8	94	95.6	92	90.3	**98.5**
Specificity [%]	99.7	98.2	98.2	99.5	96.9	96.8	99.1	93.6	**99.2**
Target Patient	Specific	Specific		Specific	Specific		Specific	Independent	**Independent**
Classification Latency [s]	12.2^	2.6	1.6	<1	<1		-	8.3	**0.12**
Technology [nm]	180	28		40	65		40	40	**65**
Supply Voltage [V]	1.5	0.5		1.1 (A) 0.7 (D)	1.2		0.49	1.1 (A) 0.9 (D)	**0.6**
Energy Efficiency [μJ/classification]	14.2	3.9	2.4	123.73^	0.23		-	28.33	**34.0**
Average Power for Classification [μW]	1.16	1.5		123.73^	-		2310.0	3.41^	**283.8**
Core Area [mm^2]	5.83	0.1		2.08	0.336		1.96	2.508	**0.348**

train-test data split, different from the patient-specific training and testing in other works. While more seizure classes can exist according to various definitions, we aim to detect and predict seizures with the uncontroversial three classes. *MulPi* predicts seizure by identifying the pre-ictal phase, which expects a potential seizure to occur within the next 30 minutes. The chip was tested with an FPGA for inputs and outputs.

Fig. 6 summarizes the specs and performance of the chip fabricated in TSMC 65nm technology and integrated into a core area of 0.348mm^2. While the inference from the chip successfully predicts and detects the seizure based on the multi-class classification, inference takes 0.12s with a 100MHz clock and consumes 0.28mW at 0.6V. *MulPi* shows high accuracy, sensitivity, specificity, F1, and PPV of over 98%. FAR for each alarm is 1.6%, and activations and weights are represented as 8-bit fixed-point numbers. In Fig. 7, it is worth noting that *MulPi* achieved similar and even higher performance than state-of-the-art works of patient-specific binary prediction [4], [13] and patient-independent binary detection [10].

The area-saving effect of *Pi* and IS stands out in the inter-comparison to a patient-specific chip. As Fig. 8(a) describes, a patient-specific chip [14] necessitates a considerable area for the overall implementation. Furthermore, the area should increase according to the number of patients because separate classification units are required for added patients. On the contrary, *MulPi* keeps a small area regardless of the number of patients due to *Pi* with IS reducing the total area noticeably. High accuracy across patients is confirmed as Fig. 8(b) displays. Fig. 9 provides breakdowns in power and area.

TABLE II compares *MulPi* to recently fabricated chips for seizure classification based on results per inference with CHB-MIT for a fair comparison. Since the papers reported their latency and efficiency with different standards, we defined the classification latency as the total time consumed for one patient. Our chip shows lower power than the prediction chip but higher than the detection chips, where prediction is unavailable. Our fast classification compensates for this higher power, consuming comparable energy. In short, *MulPi* is competitive with the latest arts for binary prediction, *Pi*, and even the simpler binary detectors, providing further functionality.

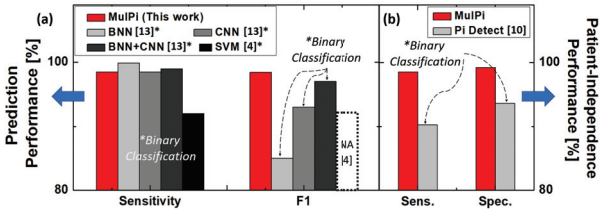


Fig. 7. Model performance comparison in (a) prediction and (b) patient independence. Unlike *MulPi*, compared works attained either one with binary results.

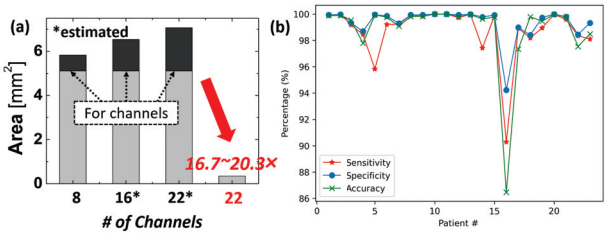


Fig. 8. (a) Comparison in area between *MulPi* and a patient-specific chip [14]. Increasing channels indicate the patient increment. (b) Accuracy according to patients by the proposed chip.

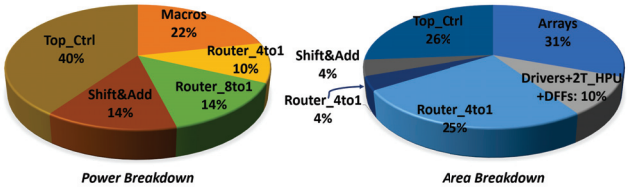


Fig. 9. Measured power and area breakdown.

V. CONCLUSION

This work presents *MulPi*, the first fabricated chip simultaneously satisfying multi-class and patient-independent seizure classification for authentic diagnosis. The proposed 5-layer *MulPiCNN* sets multiple classes up, targeting high accuracy across diverse patients. Our cross-layer solution, ISRW-CIM with 2T-HPU, tackles challenges in SRAM CIM and ensures efficiency. The proposed chip proves high accuracy and efficiency in area, time, and power despite the higher complexity of the algorithm. The chip can be further extended with more classes, thereby offering the potential for highly complex epilepsy diagnosis in the future.

REFERENCES

[1] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, and H. Adeli, "Deep convolutional neural network for the automated detection and diagnosis of seizure using eeg signals," *Computers in biology and medicine*, vol. 100, pp. 270–278, 2018.

[2] Y. Bengio, N. Léonard, and A. Courville, "Estimating or propagating gradients through stochastic neurons for conditional computation," *arXiv preprint arXiv:1308.3432*, 2013.

[3] Y. Chen, J. Mu, H. Kim, L. Lu, and T. T.-H. Kim, "Bp-scim: A reconfigurable 8t sram macro for bit-parallel searching and computing in-memory," *IEEE Transactions on Circuits and Systems I: Regular Papers*, 2023.

[4] A. Chua, M. I. Jordan, and R. Muller, "Soul: An energy-efficient unsupervised online learning seizure detection classifier," *IEEE Journal of Solid-State Circuits*, vol. 57, no. 8, pp. 2532–2544, 2022.

[5] Y.-Y. Hsieh, Y.-C. Lin, and C.-H. Yang, "A 96.2-nj/class neural signal processor with adaptable intelligence for seizure prediction," *IEEE Journal of Solid-State Circuits*, vol. 58, no. 1, pp. 167–176, 2022.

[6] T. M. Ingolfsson, U. Chakraborty, X. Wang, S. Beniczky, P. Ducouret, S. Benatti, P. Ryvlin, A. Cossettini, and L. Benini, "Epidenet: An energy-efficient approach to seizure detection for embedded systems," in *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2023, pp. 1–5.

[7] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2704–2713.

[8] B. Kim, S. Li, and H. Li, "Inca: Input-stationary dataflow at outside-the-box thinking about deep learning accelerators," in *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2023, pp. 29–41.

[9] U. Shin, L. Somappa, C. Ding, Y. Vyza, B. Zhu, A. Trouillet, S. P. Lacour, and M. Shoaran, "A 256-channel 0.227 μj/class versatile brain activity classification and closed-loop neuromodulation soc with 0.004 mm 2-1.51 μw/channel fast-settling highly multiplexed mixed-signal front-end," in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 65. IEEE, 2022, pp. 338–340.

[10] C.-W. Tsai, R. Jiang, L. Zhang, M. Zhang, L. Wu, J. Guo, Z. Yan, and J. Yoo, "Scicnn: A 0-shot-retraining patient-independent epilepsy-tracking soc," in *2023 IEEE International Solid-State Circuits Conference (ISSCC)*. IEEE, 2023, pp. 488–490.

[11] M. Um, M. Kang, H. Kwak, K. Noh, S. Kim, and H.-M. Lee, "An ecram-based analog compute-in-memory neuromorphic system with high-precision current readout," in *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2023, pp. 1–5.

[12] J. Van Assche, M. F. Carlino, M. D. Alea, S. Massaioli, and G. Gielen, "From sensor to inference: end-to-end chip design for wearable and implantable biomedical applications," in *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2023, pp. 1–5.

[13] J. Wang, S. Zhao, J. Yang, and M. Sawan, "An event-driven neural signal processor for closed-loop seizure prediction," in *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2023, pp. 1–5.

[14] Y. Wang, H. Luo, Y. Chen, Z. Jiao, Q. Sun, L. Dong, X. Chen, X. Wang, and H. Zhang, "A closed-loop neuromodulation chipset with 2-level classification achieving 1.5-vpp cm interference tolerance, 35-db stimulation artifact rejection in 0.5 ms and 97.8%-sensitivity seizure detection," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 15, no. 4, pp. 802–819, 2021.

[15] J.-H. Yoon, M. Chang, W.-S. Khwa, Y.-D. Chih, M.-F. Chang, and A. Raychowdhury, "29.1 a 40nm 64kb 56.67 tops/w read-disturb-tolerant compute-in-memory/digital rram macro with active-feedback-based read and in-situ write verification," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 64. IEEE, 2021, pp. 404–406.

[16] M. Zhang, L. Zhang, C.-W. Tsai, and J. Yoo, "A patient-specific closed-loop epilepsy management soc with one-shot learning and online tuning," *IEEE Journal of Solid-State Circuits*, vol. 57, no. 4, pp. 1049–1060, 2022.