An Integrated CMOS/Memristor Bio-Processor for Re-Configurable Neural Signal Processing

An Integrated CMOS/Memristor Bio-Processor for Re-configurable Neural Signal Processing

Grahame Reynolds, Xiongfei Jiang, Alexander Serb, Themis Prodromakis, Shiwei Wang
Centre for Electronics Frontiers, IMNS, School of Engineering, University of Edinburgh, EH9 3JL, UK
Email: {g.k.reynolds, xiongfei.jiang, aserb, t.prodromakis, shiwei.wang}@ed.ac.uk

***Abstract*—This paper proposes a bio-processor for neural signal analysis. The device architecture features an analogue Front-End and a Process Element, the latter can be scaled as an array. Rather than a single dedicated algorithm, the Process Element supports multiple analysis modes, utilising the analogue behaviour of memristors. When used as part of an array structure, each Process Element can be programmed independently and furthermore, the array elements can be electrically interconnected in an arbitrary manner. The device facilitates an inter-network of in-memory computation units, i.e. an inter-network of functions. This supports construction of a system that is highly scalable, re-configurable and thus adaptive. The device enables multi-functional neural recording and processing, for early stage signal exploration. The device has been implemented using a standard 180nm CMOS process with the addition of back-end-of-line (BEOL) memristor deposition. Although targeted at neural signal analysis, the device and the architecture described is considered general purpose and may find application within other disciplines.**

I. Introduction

The capture and analysis of neural signals from neural recording apparatus, together with the application of stimuli by neuromodulation, has become increasingly commonplace to decipher activity for diagnosis and/or to address medical conditions [1]. Neural analysis has been used to detect and investigate conditions such as epilepsy, the effects of spinal injury and neurodegenerative disorders such as Dementia, Parkinson's disease and Alzheimer's disease [2]. Such analysis has also been used to investigate the prospects for brain-controlled prostheses, mobility aids and appliances. Recent advances with neural signal processors have been enabled by implementation of machine learning (ML) algorithms on low-power integrated circuit chips, which facilitated low-latency detection of neurological disorders (especially the onset of epilepsy) on the edge [3]–[7]. However, most processors today achieve an improvement in energy efficiency by sacrificing versatility, support limited algorithms and are dedicated to specific application scenarios. When applied, especially within an implantable device, it is essential the processor can be easily reconfigured to adapt to the changes of the tissue-implant interface conditions and customised to patients. Such adaptation is preferably achieved through reprogramming of the device rather than by a surgical hardware upgrade.

Recent advances in memristor or resistive RAM (RRAM)-based in-memory computing (IMC) techniques have provided the opportunity to achieve both energy efficiency and versatility. The typical core structure is a crossbar of One-Transistor-One-Resistor (1T1R) cells, functioning as both memory and computational elements. This reduces excessive energy consumption during memory access. The advantages of energy efficiency from memristor/RRAM-based IMC have been demonstrated in various neural signal processor designs [9]–[12]. With a memristor/RRAM crossbar as the memory/compute core, a processor can not only perform general-purpose neural network computation [8] but also support various signal processing functions and modalities, such as finite impulse response (FIR) filtering [9], template matching (TM) [10], memristive integrating sensing (MIS)/analog conductance modulation [11], [12], and reservoir computing (RC) [13], [14] by using different peripheral circuits for memristor/RRAM interfacing and control. While the proof of concept for these processing functions and modalities have been demonstrated separately and on standalone memristor/RRAM arrays, a fully integrated chip that incorporates re-configurable processing capabilities for versatile neural signal analysis (e.g. as shown in Fig.1) has not yet been achieved.

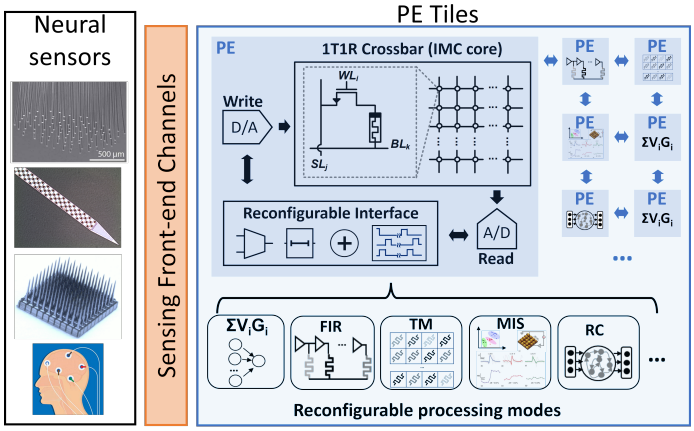

Fig. 1. The proposed re-configurable neural signal processor based on memristor/RRAM IMC. In each PE, an 1T1R crossbar array performs energy efficient IMC for various neural signal processing tasks (e.g. general-purpose vector-matrix multiplication for neural network acceleration, FIR filtering, template matching, memristive integrating sensing, and reservoir computing, etc.) by re-configuring the peripheral circuits. The processing algorithms can be flexibly constructed by chaining multiple PEs together achieving multi-step processing involving one or more processing tasks.

Herein we describe an integrated, re-configurable and adaptive system. The architecture is presented in Section II, operating modes in III, the completed design in IV and a discussion concerning scaling and further work in V.

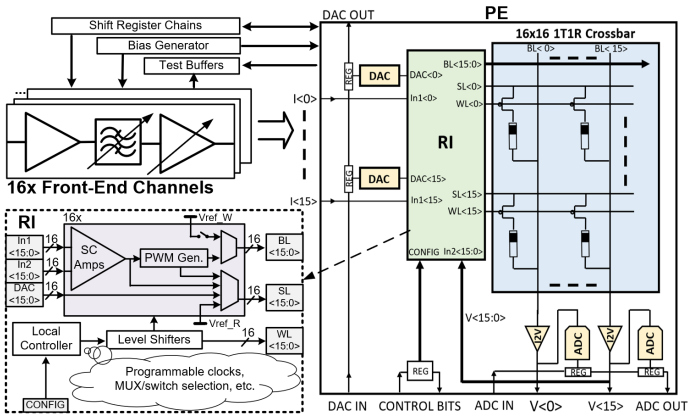

Fig. 2. Chip architecture. The front-end circuit conditions the incoming sensor signals for onward processing [15]. It contains 16 analogue channels each consisting of a low-noise amplifier, band-pass filter and variable-gain amplifier. The bandwidth and gain of the channels are both programmable, making it feasible to record multiple biomarkers such as action potentials (APs), local field potentials (LFPs), and intracranial/extracranial electroencephalography (EEG) recordings. The front-end also supports a pass-through mode allowing the input signals to be processed by the PE directly to perform multi-step neural signal processing by chaining multiple processor chips together. The PE performs the processing functions using a memristor crossbar as a fundamental computation structure to which the stimuli are controlled by a Reconfigurable Interface (RI). The PE supports both analogue or digital inputs and analogue or digital outputs. The inputs can be sourced from the analogue front-end in gain/filter mode, the front-end in pass-through mode or from a digital stream applied to a Digital-to-Analogue Converter (DAC). The front-end outputs or signals obtained from the DAC are presented to the memristor crossbar. The crossbar outputs are converted to analogue outputs by a dedicated current-to-voltage (I2V) circuit per bit-line and may also be converted to a digital output, using a dedicated Analogue-to-Digital Converter (ADC) per bit-line.

II. Chip Architecture

In this paper we present the design of an integrated CMOS/Memristor bio-processor using memristor-based processing elements (PEs) with re-configurable peripheral circuits to support multiple neural signal processing modalities including FIR, TM, MIS, and RC as shown in Fig 2. It also includes a 16-channel neural sensing front-end. A single PE is implemented in this design with a flexible data interface such that multiple PEs can be inter-connected via chip-chip interconnections at printed circuit board level. The sub-blocks of the PE are described as follows.

1) DAC: The DAC consists of an 8-bit R-2R design, where the output voltage linearly spans the amplitude between two externally applied reference voltages. The digital input is applied serially using a double-buffered register. The first stage being a serial shift register, the second stage being a static register. The second stage is loaded on demand from the first stage, hence the first stage can be reloaded without disturbing the state of the second stage. A PE has 16 8-bit DAC registers.

2) Memristor Crossbar: A bespoke 16 x 16 TiO_2, PMOS, 1T1R structure is used for the crossbar, the memristor being integrated by deposition onto the surface of the silicon wafer [16]. The memristor device has a bipolar switching characteristic; ideal for both digital and analogue application.

3) I2V: The crossbar bit-lines are connected to a dedicated I2V. This is an operational-amplifier based transimpedance amplifier (TIA) with adjustable range.

4) ADC: Each I2V output is connected to a dedicated ADC. The ADC consists of an 8-bit ramp design, where the output code linearly spans the amplitude between two reference voltages. By adjusting the reference voltages, the ADC may be used to provide a threshold or activation function, such as Rectified Linear Unit (ReLU). The ADC result is read using a double-buffered register. The first stage being a static register, the second stage being a serial shift register. The second stage is loaded on demand from the first stage, hence the second stage may be read without disturbing the ADC operation while a conversion is in progress. A PE has 16 8-bit ADC registers.

5) RI: As shown in Fig. 2, The RI comprises primarily of a local controller and 16 units of switched-capacitor (SC) amplifiers, pulse-width modulation (PWM) generators, selection switches and multiplexers. The local controller receives configuration bits from the global shift register chains and provides programmable clocks and control signals for the SC amplifiers and other circuits, according to the operating mode of the PE. A more detailed circuit diagram and the clock schemes of the RI is shown in Fig.3 and the RI operation principles are described together with the re-configurable neural signal processing modes in Section III.

III. Re-configurable Operation Modes

The configurations of the processor for different operating modes are illustrated in Fig. 3. The basic functions include electroforming and readout modes. Using these two basic functions in an interleaved manner, write-and-verify iterations can be applied to program the memristors to the desired resistance states. The supported neural signal processing modalities are as follows.

1) General-Purpose Vector Matrix Multiplication (GP VMM): GP VMM is an essential operation in neural network computation. The trained weights are stored as memristor conductance and the VMM operation is performed by applying the input signal vectors on the SLs. The input vectors can be from either the DACs (digital inputs) or the SC amplifiers through their embedded analog drivers (analog inputs). The VMM calculation results are read on the BLs where the I2V/TIAs and ADCs convert the summed currents into analog or digital outputs. Each PE can process up to 16 input neurons and support 16 output neurons in a single layer. Deeper or larger scale neural networks can be constructed using the PE in several iterations or using multiple PEs in parallel.

2) Finite Impulse Response (FIR) filtering: FIR filtering is commonly used to pre-process neural signals by extracting features in the frequency domain. Brain wave patterns have been associated with different brain states and these occur in specific frequency bands [9], [17]. The PE configuration for FIR is largely the same as for VMM, but the timing of the SC amplifier clocks in the RI differ to implement a 16-sample delay line, whereby each SC amplifier and each row of the crossbar serve as one FIR tap. The conductance of the memristors is defined by the FIR filter coefficients in this case. Each column of the crossbar serves as one frequency band. Each PE can support up to 16-channel, 16-tap FIR filtering.

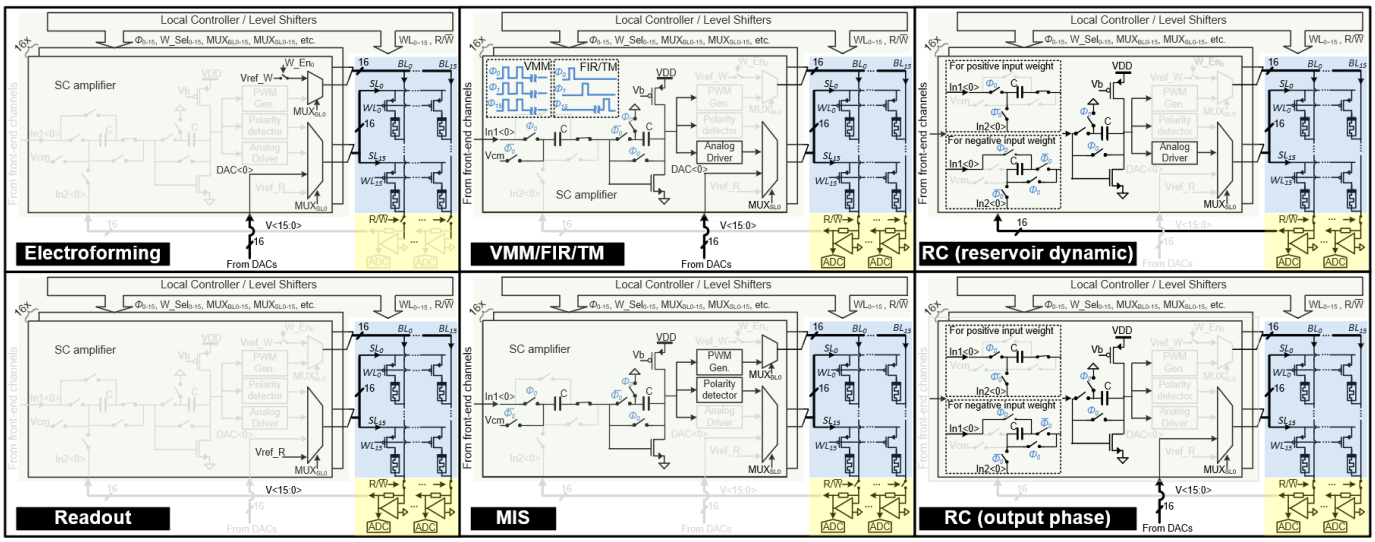

Fig. 3. PE configurations for different processing modalities. During electroforming, most of the circuit blocks in the PE are disabled and/or bypassed except the DACs and a few switches and multiplexers. The electroforming signals are supplied from the DAC through the source lines (SL_{0-15}) with the reference voltage ($Vref\_W$) applied through the bit lines (BL_{0-15}). The memristors are electroformed in an interleaved manner by activating the selectors through the word lines (WL_{0-15}) which are controlled by the RI. In readout mode, the resistance of the memristors are measured by applying a 0.2V voltage difference (which is below the resistive switching threshold) across the devices through the SLs and BLs. The signal $R/\bar{W}$ (controlled by the local controller) is used to connect or disconnect the I2V/ADC circuits with the crossbar array when reading from or writing to the memristors. During VMM, FIR, and TM, the 16 SC amplifiers in the RI are activated which generate the stimuli voltages on the SLs based on the input signals. In VMM the 16 SC amplifiers are clocked synchronously with each SC amplifier representing one input neuron, and each I2V/ADC on the BLs represents one output neuron. In FIR and TM, the 16 SC amplifiers are clocked in a time-interleaved fashion creating a 16-sample delay line. In FIR mode each row of the memristors represents the coefficients of one FIR filter tap and each column represents one frequency band, while in TM each row represents the template coefficients and each column represents one waveform template. In MIS, the SCA amplifiers are PWM modulated to drive the memristors with higher energy efficiency. The PWM pulses are applied on the BLs and bi-phasic pulses are achieved by applying the polarity information through the SLs. The SC amplifiers are clocked in time-interleaved fashion similar to FIR/TM. In RC, the SC amplifiers can be configured to apply a positive or negative sign (weight) to the inputs. The reservoir internal connection coefficients are implemented on the WLs which can be easily re-configured to support multiple reservoir topologies (e.g. simple cycle reservoir, delay line reservoir, etc). The I2Vs readout the current memristor resistance states weighted by WLs and feed the results back to the SC amplifiers which sum the results with the sign-weighted inputs and generate the voltage stimuli to update the memristor resistance states. Each column of memristors represents one 16-node reservoir and each PE consists of 16 reservoirs to process 16-channel inputs. Each row of memristors represents one internal node in the reservoirs which process one sample of the input time-series. To process 16 channels, the SC amplifiers are time-division multiplexed. After all memristors are updated (reservoir dynamics finish), the outputs are calculated by applying the DAC signals (trained weights) on the SLs and reading out the VMM results on BLs.

3) Template Matching (TM): The TM mode compares the inputs against waveform templates represented by the memristor resistance states. The configuration of PE for TM is mostly the same as for FIR; the only difference being that memristors are programmed to store the template coefficients instead of FIR coefficients. The matching results can be read out on the BLs either all at once or one cell at a time, to allow flexible data normalisation required in TM.

4) Memristive Integrated Sensing (MIS): MIS is an emerging method for low-latency and low-power neural signal detection and classification based on the principle that each neural signal leaves a distinguishable signature on the resistance state when applied to a memristor [11], [12]. The MIS mode requires writing to the memristors continuously. Therefore, the SC amplifier output is PWM modulated for improved energy efficiency, allowing memristors to be driven using logic gates. Similarly to FIR/TM, the SC amplifiers are clocked in the time-interleaved fashion with each row of the memristors processing one sample in a 16-sample time window. Each column of the memristors process inputs from one of the 16 front-end channels. The PE is re-configured into readout mode after each time window so that the memristor states are read for further processing.

5) Reservoir Computation (RC): RC is a type of Recurrent Neural Network (RNN) whereby the input data is transformed into spatiotemporal patterns in a high-dimensional space by an RNN within the reservoir itself [19]. It is especially efficient for pattern analysis of signals with rich temporal dynamic features such as neural signals, and physical RCs can be easily implemented using memristors [13], [14]. The PE can function as a physical RC equivalent to the 'minimum complexity echo state network' [18] governed by the following equations:

$$x_{t+1} = H(Vs_{t+1} + Wx_t) \tag{1}$$

$$y_{t+1} = Ux_t \tag{2}$$

where x is the reservoir internal states mapped to the resistance of the memristors, W is the reservoir internal connection weights (0 or 1) and mapped to the WL selection bits, H is reservoir activation function mapped to the nonlinear memristor V-R dynamics, V is the input connection weights (fixed unity value for all inputs, with random signs) and mapped to the SC amplifier sign configurations, U is the trained weights for the output layer and mapped to DAC signals on the SLs,

finally y is the reservoir output. The RC operations requires two phases: in phase I (governed by Eq. 1), the reservoir dynamic is achieved by updating the memristor states x_{t+1} using stimulus generated from the RI/SC amplifiers which sum the next-state input vectors s_{t+1} weighted by random signs V and the current node states x_t weighted by the internal connection coefficients. In phase II (governed by Eq. 2), the reservoir outputs are obtained by taking the VMM results (trained output layer weight vectors multiplied by the reservoir state matrix) on the BLs and read out by the I2Vs and ADCs.

IV. Results

The processor core layout is shown in Fig. 4. The chip

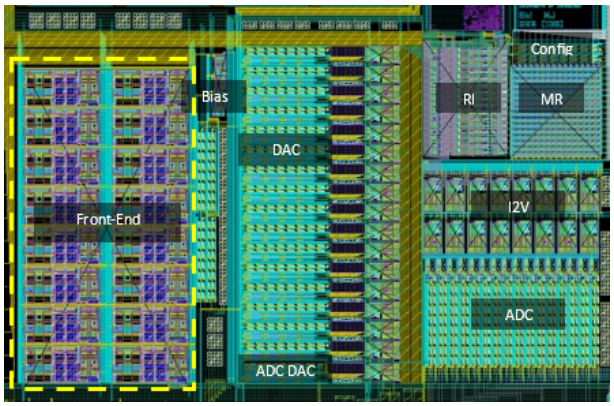

Fig. 4. Processor core. The Front-End (marked in yellow) measures 0.6x1.2mm, the PE 1.6x1.4mm. Block "MR" is the memristor array.

has been implemented using a standard 180nm CMOS BCD technology (Fig. 4), with the memristors to be integrated in house through post-CMOS processing on the back-end-of-line. As shown in Fig. 5, simulations in Cadence Virtuoso demonstrate the processor operates in multiple modes in one signal simulation run. The design is compared with state-of-the-art memristor-based bio-signal processors in Table. I.

TABLE I
Comparison with Memristor-based Bio-signal Processors

Metric	[9]	[10]	[11]	[12]	This work
Fully-integrated chip	N	N	N	N	Y
Memristor array size	1k	16x16	32x32	32x32	16x16
Number of channels	1	32	1	16	16
Processing function	FIR + SLP[1]	TM	MIS	MIS	FIR/TM/ MIS/RC
Re-configurable	N	N	N	N	Y

[1]Single-layer perceptron

V. Discussion and Conclusions

As described, the architecture of the processor includes a fundamental, re-usable PE element. Up-scaling may be accomplished by increasing the PE count. Thus a single PE instance may be integrated as a die with multiple such die then interconnected as a 2D or 3D array. Similarly, multiple instances of PE may be arrayed (integrated) on a die with multiple such die arranged in a 2D or 3D array.

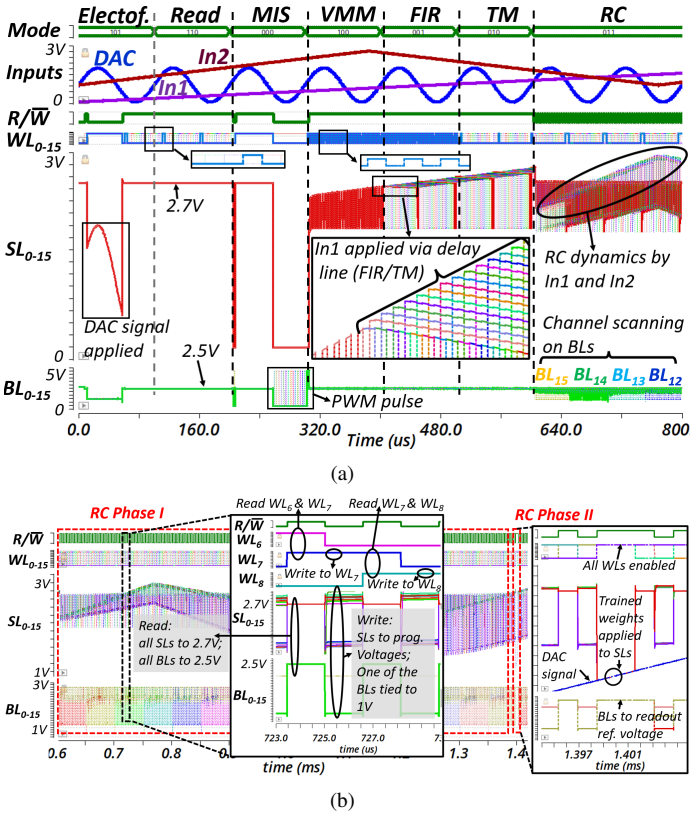

Fig. 5. (a) Sweeping the 7 operation modes in one transient simulation. (b) Zoom-in view of the waveforms during RC operation

The interconnection schemes allow different parts of a PE array to be used for different purposes. For example, a first sub-array of PE elements may be used to acquire neural data using MIS mode. A second sub-array may perform FIR in parallel. A third sub-array may use RC to perform classification. Furthermore, since PE can be controlled independently, a PE

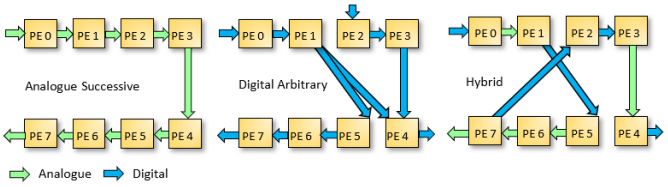

Fig. 6. PE interconnection schemes. Each PE in an array can be configured independently and with analogue or digital inputs and/or outputs. Analogue connections are always concatenated successively (connection is parallel), digital connections can be concatenated arbitrarily (connection is serialised), analogue and digital can be mixed, digital connections may be split or combined at bit level.

array may be used to execute data capture using MIS. From observations of the collected data, a signal artefact of interest may become evident. The PE array may then be reprogrammed to execute template matching for that specific artefact.

In conclusion, this paper has presented an integrated, re-configurable system which represents a platform on which to develop adaptive neural analysis systems.

REFERENCES

[1] L. Drew, "Decoding the business of brain–computer interfaces," Nature Electronics, Volume 6, February 2023, pp. 90-95.

[2] D. Pei, R. Vinjamuri, Advances in Neural Signal Processing, September 2020, EBOOK ISBN 978-1-83968-396-1.

[3] C. -W. Tsai et al., "SciCNN: A 0-Shot-Retraining Patient-Independent Epilepsy-Tracking SoC," 2023 IEEE International Solid- State Circuits Conference (ISSCC), San Francisco, CA, USA, 2023.

[4] U. Shin et al., "NeuralTree: A 256-Channel 0.227-μJ/Class Versatile Neural Activity Classification and Closed-Loop Neuromodulation SoC," in IEEE Journal of Solid-State Circuits, vol. 57, no. 11, pp. 3243-3257, Nov. 2022.

[5] A. Chua, M. I. Jordan and R. Muller, "SOUL: An Energy-Efficient Unsupervised Online Learning Seizure Detection Classifier," in IEEE Journal of Solid-State Circuits, vol. 57, no. 8, pp. 2532-2544, Aug. 2022.

[6] Y. Wang, Q. Sun, H. Luo, X. Chen, X. Wang and H. Zhang, "26.3 A Closed-Loop Neuromodulation Chipset with 2-Level Classification Achieving 1.5Vpp CM Interference Tolerance, 35dB Stimulation Artifact Rejection in 0.5ms and 97.8% Sensitivity Seizure Detection," 2020 IEEE International Solid- State Circuits Conference - (ISSCC), San Francisco, CA, USA, 2020, pp. 406-408.

[7] G. O'Leary, D. M. Groppe, T. A. Valiante, N. Verma, and R. Genov, "NURIP: Neural Interface Processor for Brain-State Classification and Programmable-Waveform Neurostimulation," IEEE J. Solid-State Circuits, vol. 53, no. 11, pp. 3150–3162, 2018.

[8] W. Wan et al., "A compute-in-memory chip based on resistive random-access memory," Nature, vol. 608, no. 7923, pp. 504–512, 2022.

[9] Z. Liu et al., "Neural signal analysis with memristor arrays towards high-efficiency brain–machine interfaces," Nature Commun., vol. 11, no. 1, pp. 1–9, 2020.

[10] Y. Shi et al., "High Throughput Neuromorphic Brain Interface with CuOx Resistive Crossbars for Real-time Spike Sorting," 2021 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2021, pp. 16.5.1-16.5.4.

[11] I. Gupta, A. Serb, A. Khiat, R. Zeitler, S. Vassanelli, and T. Prodromakis, "Real-time encoding and compression of neuronal spikes by metal-oxide memristors," Nature Commun., vol. 7, no. 1, p. 12805, Nov. 2016.

[12] Z. Liu et al., "Multichannel parallel processing of neural signals in memristor arrays," Sci. Adv., vol. 6, no. 41, pp. 2–10, 2020.

[13] J. Moon et al., "Temporal data classification and forecasting using a memristor-based reservoir computing system," Nature Electron., vol. 2, no. 10, pp. 480–487, Oct. 2019.

[14] Y. Zhong, J. Tang, X. Li, B. Gao, H. Qian, and H. Wu, "Dynamic memristor-based reservoir computing for high-efficiency temporal signal processing," Nature Commun., vol. 12, no. 1, p. 408, Jan. 2021.

[15] X. Jiang, C. Sbandati, G. Reynolds, C. Wang, C. Papavassiliou, A. Serb, T. Prodromakis and S. Wang, "A Neural Recording System With 16 Reconfigurable Front-end Channels and Memristive Processing/Memory Unit," 2023 IEEE NEWCAS conference.

[16] A. Mifsud, J. Shen, P. Feng, L. Xie, C. Wang, Y. Pan, S. Maheshwari, S. Agwa, S. Stathopoulos, S. Wang, A. Serb, C. Papavassiliou, T. Prodromakis and T. Constandinou, "A CMOS-based Characterisation Platform for Emerging RRAM Technologies," 2022 IEEE International Symposium on Circuits and Systems (ISCAS).

[17] P. Abhang, S Mehotra, "Technological Basics of EEG Recording and Operation of Apparatus," Introduction to EEG- and Speech-Based Emotion Recognition, 2016.

[18] A. Rodan and P. Tiňo, "Minimum complexity echo state network," IEEE Trans. Neural Networks, vol. 22, no. 1, pp. 131–144, 2011, DOI: 10.1109/TNN.2010.2089641.

[19] G. Tanaka, T. Yamane, J. Benoit Héroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, A. Hirose, "Recent advances in physical reservoir computing: A review," Neural Networks, Jul 2019.