Design Methodology for Energy-constrained AI Edge Inference in Implantable Medical Devices

José Sales Filho, Jianxiong Xu, Mustafa Kanchwala, Gerard O'Leary, José Zariffa, Roman Genov
University of Toronto, Toronto, ON, CA

2023 IEEE Biomedical Circuits and Systems Conference (BioCAS) | 979-8-3503-0026-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/BioCAS58349.2023.10388749

***Abstract*—This paper provides design guidelines for enabling edge-computing machine learning blocks in implantable medical devices. Energy consumption is a critical design factor for such implants, as the dissipated heat can damage the surrounding tissue and battery size is limited. For wirelessly powered devices, the power budget is also often limited by safety guidelines for specific absorption rates imposed on the radiated electromagnetic field. The paper examines several case studies of closed-loop neural implantable systems and analyzes the design choices for placing inference computation blocks along the implant-wearable-handheld-cloud chain. The case studies cover examples of our recent work for both the central nervous system (CNS) and peripheral nervous systems (PNS). Depending on the data and the algorithm's complexity, the examples use either in-implant or remote inference, or a hybrid approach that combines both. For each example, the paper shows trade-offs between local and remote computation along the signal path.**

***Index Terms*—AI, machine learning, neural interfaces, implants, edge computing.**

I. Introduction

Implantable medical devices, and particularly neurostimulation devices, have transformed the healthcare field by providing innovative solutions for treating acute and chronic disorders. These devices can be applied to both the central (CNS) and peripheral nervous systems (PNS), and have emerged as an effective approach which can improve patient outcomes and quality of life. These devices can operate in open-loop or closed-loop modes, with the latter being more promising due to its responsive nature and effectiveness in delivering optimally-timed therapy [1]. Many such devices, however, may not have been used to their full potential, as their computational capabilities often cannot support the computational demands imposed by modern high-efficacy machine learning (ML) inference algorithms.

One such example is neurostimulators for intractable epilepsy. Epilepsy is a common neurological disorder that affects over 65 million people worldwide. However, about one-third of them - over 20 million people suffer from intractable epilepsy, i.e. they are resistant to existing anti-seizure medications [2]. Implanted closed-loop neurostimulators, which interface with the CNS or the PNS, are often used to deter seizures from spreading. However, these devices have some limitations such as low average seizure freedom and high false stimulation rates [3].

Closed-loop neurostimulators can also be used to help restore sensory and/or motor functionality in patients who have suffered spinal cord injuries such as paraplegia or quadriplegia [4][5]. However, recent cases still face challenges in terms of low efficacy, mainly due to the lack of computational capabilities to implement methods for selective recording and stimulation of function-specific neural pathways.

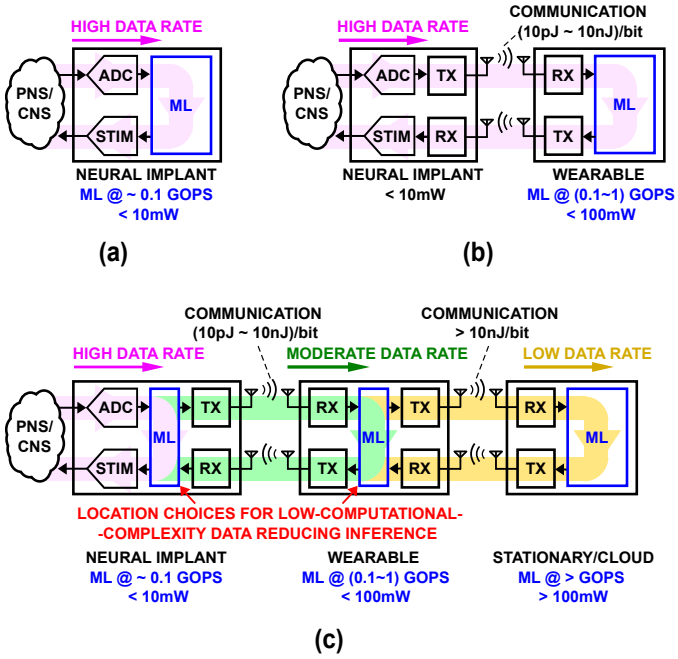

Fig. 1. Strategies for enabling AI inference in energy-constrained implantable medical devices: (a) in-implant inference; (b) remote inference; (c) hybrid inference.

Implantable neurostimulators, such as the two examples above, typically have strict power and size constraints, which make implementing computationally-expensive algorithms by way of on-chip signal processing challenging. This severely limits the kinds of ML algorithms that can be implemented on-chip. On the other hand, performing signal processing on a wearable or stationary computer, where one does not face the same area and size constraints, requires in-implant high-throughput wireless transmitters that can send out all the recorded raw data. The latter approach adds a significant cost of shipping each bit of data out, and also may add significant delays, which increases latency and may lead to reduced efficacy. The challenges faced by these two scenarios for implementing complex and with low latency ML algorithms (i.e. on-device processing or remote-computer-based processing) necessitate a deeper exploration of various trade-offs related to them as well as other possible solutions.

A brute-force approach to overcome these challenges is to aggressively optimize the energy efficiency of the signal processing circuits. For example, recent research in low-power biomedical integrated circuits have explored energy-efficient ways to do so in the implant [6][7] or on a wearable platform [8][9].

Another recent approach is to implement ML inference blocks at one or more points along the signal path of the implant's closed loop, in order to reduce and process data, while carefully considering the energy cost of computation and communication at each such point [10].

In this paper, we analyze the trade-offs involved and offer considerations for choosing where to close the loop within the signal path of the implant-wearable-handheld-cloud signal path. To illustrate these concepts, we present case studies from our recent work to illustrate and explain the design choices related to them.

II. Where to Close the Loop?

A crucial design choice for an implantable device that operates under tight energy constraints is the location of the inference-based feedback loop along the signal path. This decision depends on three key factors: (1) The latency required for the physiological function response; (2) The power budget available and allocated to each part of the implant signal path; (3) The computational complexity of the inference algorithm.

One possible option is to implement the closed-loop operation entirely in the implant, as shown in Fig. 1(a). Implantable devices usually have a power budget of less than 10 mW, assuming wireless power operation [11], and a maximum computational throughput of 0.1 GOPS (Giga Operations Per Second). This option is ideal for minimizing closed-loop latency, as the inference is done at the edge. However, this also means that any algorithm in the implant would need an energy efficiency of 0.1 pJ/bit, which demands highly-efficient hardware design, often supported by on-chip hardware acceleration.

Another option is to send the recorded data from an implant to a remote processor, such as a wearable device, as shown in Fig. 1(b). Here, the power and computational throughput constraints are less strict to run the inference algorithm remotely. Around 100 mW, assuming a battery-powered wearable. However, the energy cost of transmitting a bit over the wireless channel is at least two orders of magnitude higher (10 pJ/bit to 10nJ/bit). This is feasible if either the system operates with low data rate, or an energy-efficient data transmitter is used at the implant side. Naturally, the communication delay between the implant and the wearable needs to be considered in the closed-loop latency response.

A third option is a hybrid inference approach, shown in Fig. 1(c), where the ML computational blocks are split among the implant, wearable and possible stationary platforms. This method allows for more flexibility to balance the trade-off between energy cost of computation and throughput in data communication. By doing an initial inference step, the data transmission rate to the wearable is greatly reduced. Moreover, the feedback from the subsequent inference stage can be potentially used for online training, which is challenging to do at the implant stage.

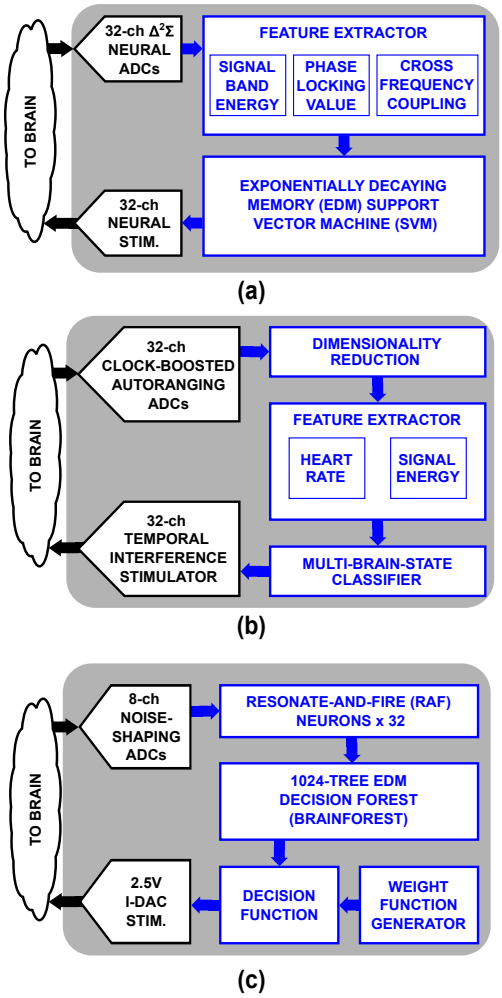

Fig. 2. Examples of closed-loop CNS interfaces with in-implant inference: (a) Neural Interface Processor (NURIP) implant IC, employed for epiletic seizure detection (Adapted from [12]); (b) second version of the NURIP IC, employed for seizure and abnormal sleep activity detecion (Adapted from [13]); and (c) energy-efficient brain implant powered by the BrainForest decision-tree-based classifier (Adapted from [7]).

A. In-implant Inference

The Neural Interface Processor (NURIP) is an example of in-implant inference by using energy efficient hardware in epilectic seizure detection [6][12]. A simplified block diagram of such system is depicted in Fig. 2(a). It performs spatial filtering and dimensionality reduction on the digitized data stream from the $\Delta^2\Sigma$ ADCs using an auto-encoder neural network. Moreover, it extracts three frequency-dependent features from the data: (1) signal-band energy, (2) phase locking value and (3) cross-frequency coupling. These features are then fed into an Exponentially-decaying Memory Support Vector Machine (EDM-SVM) accelerator that classifies the brain state among three classes: (1) seizure, (2) ictal or (3) inter-ictal.

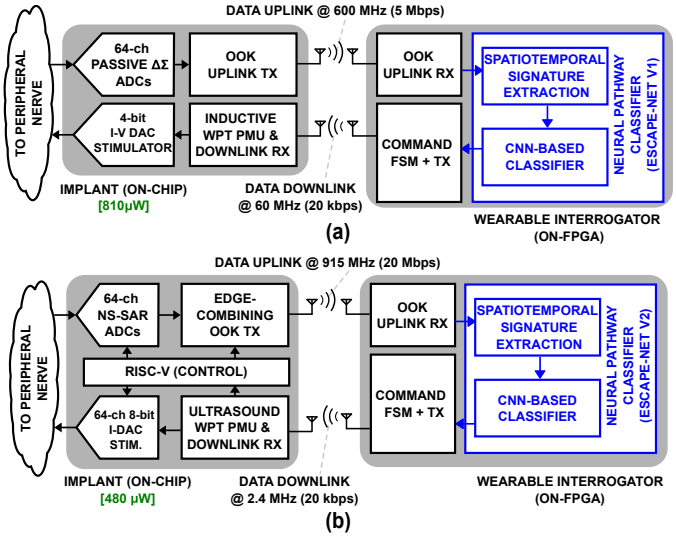

Fig. 3. Examples of closed-loop PNS interfaces with remote, on-wearable inference: (a) ultra-low power inductive powered active PNS probe IC, that operates in tandem with the on-interrogator CNN-based ESCAPE-NET V1 neural pathway classifier (Adapted from [8]; (b) fascicle-selective ultrasound powered peripheral nerve interface IC, which employs an the ESCAPE-NET V2 classifier (Adapted from [9]).

A key contribution of this design is its ability to reuse existing on-chip hardware and use simple circuits such as circular buffers, FIR filters with CORDIC blocks, and shift and add operations for efficient implementation of the EDM-SVM. This approach enables a small form factor implementation and achieves 168.8 μJ/classification and latency smaller than 0.1 s, requiring 96kB of SRAMs.

The design depicted in Fig. 2(b) is a second version of the implant based on the NURIP classifier [13]. Main changes include a feature extractor to measure heart rate changes and signal-band energy for sleep state classification, and an auto-ranging analog front-end resistant to stimulation artifacts and DC offsets. The system also showcases current DACs deploying temporal interfering stimulation (TIS) for non-invasive targeted stimulation, and wireless power and data blocks to enable fully wireless implantable operation.

The inference block in this system has similar performance specifications as the previous implementation depicted in Fig. 2(a). Although consuming little enough power for implantable operation, it is worth stating that the majority of the energy consumption of the classifier resides in the SRAM memory access.

The architecture of the BrainForest classifier[7], depicted in the implantable system in Fig. 2(c), does not need on-chip SRAMs or memory access operations, unlike NURIP. It uses a neuromorphic feature-extraction engine based on resonate-and-fire (RAF) neurons. To perform bandpass filtering, it applies a two-stage approach: first, a low-performance power-of-two coefficients IIR filter that uses shift and add operations in hardware; second, a non-linear filtering logic that performs half-wave detection by counting the samples between the min and max points of the filtered waveform. It encodes the amplitude with a bit-serial firing pattern before sending the spike data stream to the classifier. BrainForest is an ultra-low power classifier which consists of a 1024-tree exponentially decaying memory decision forest (EDM-DF). The decision trees in this model have a decision depth of 1, which makes the memory access deterministic. The model parameters are locally integrated in the distributed compute elements and the weights are regenerated on-chip to avoid using SRAMs or memory accesses.

This design introduces energy efficient circuits that surpass conventional methods in the following aspects: (1) The feature extractor based on RAF employs a low-performance but resource-efficient IIR filter stage that can detect signal band energy more sensitively at lower ADC resolutions than the conventional FIR based design. This enables the use of low-resolution ADCs that consume less power, and demonstrates the potential of developing innovative energy efficient circuits for edge-AI processing; (2) The memory-optimized energy efficient EDM-DF model for classification reduces the power and area requirements by several orders of magnitude compared to decision forest models with SRAMs. The achieved real-time energy consumption of 36 nJ per classification for an on-chip closed-loop operation makes this approach suitable for applications in epilepsy and one of the best in its class for edge-AI in neural implants.

B. Remote Inference

In the first example of remote inference (Fig. 3(a)), an ultra-low power active probe IC operates in conjunction with a wearable interrogator [10]. The implant uses a highly energy-efficient OOK data transmitter to send the raw digitized data output from the passive-$\Delta\Sigma$-based recording front-end channels. As a result, all the digital signal processing and inference occur on the wearable interrogator side, implemented on an FPGA.

One potential application for this active nerve probe is minimally invasive extraneural interfaces for the sciatic nerve. Neural activity from the nerve reflects sensory and motor functions related to lower limbs, and it can be utilized in prostheses for individuals with impairments. To achieve neural pathway recording selectivity, a CNN-based classifier called Extraneural Spatiotemporal Compound Action Potential Extraction (ESCAPE-NET) [14] is used in this system.

The ESCAPE-NET classifier distinguishes among three different neural pathways from sciatic nerve recordings (Fig.4(a)): dorsiflexion, plantarflexion, and heel pricking. The CNN processes spatiotemporal maps, highlighting spatial and temporal features similar to conventional images. The order of electrode scanning can be adjusted to prioritize either spatial or temporal emphasis, as shown in Fig.4(b).

The classifier implemented on the wearable interrogator (Fig.3(a)) is illustrated in Fig.4(c). It consists of two parallel paths for spatial and temporal features, each with three convolutional layers and max pool layers, connected to an output layer for activity classification.

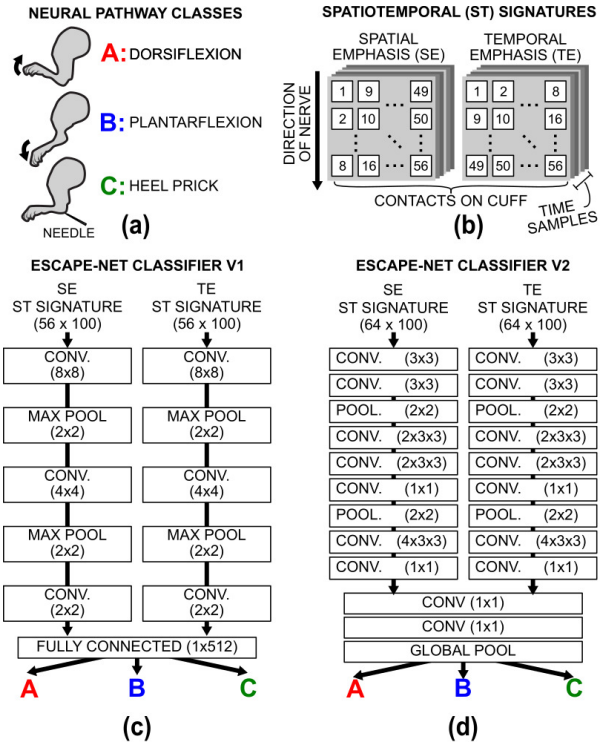

Fig. 4. (a) Neural pathway outputs for the ESCAPE-NET CNN classifier, which represents three classes of hind limb motor and sensory functions; (b) Definition of spatiotemporal signatures with spatial and temporal emphases; (c) Diagram of the ESCAPE-NET CNN classifier (V1) [14], realized on the interrogator in [8]; (d) Diagram of the second version of the ESCAPE-NET CNN classifier (V2), using more intermediate layers but with overall fewer filter parameters, employed on the interrogator in [9].

The second case study is a second version of the active probe in [8] was presented in [9], and its simplified diagram is depicted in Fig. 3(b). The enhancements to the on-chip implant include: (a) ultrasound powering to extend the operating range towards the nerve; (b) an on-chip RISC-V processor for control purposes; (c) fascicle-selective stimulation paradigm to target deep regions within the nerve; (d) and an energy-efficient edge-combining data transmitter that achieves a higher data rate of 20 Mbps.

A similar version of the ESCAPE-NET classifier is implemented on the interrogator side of this system, as shown in Fig. 4(d). This CNN classifier utilizes more intermediate convolutional and pooling layers, resulting in fewer parameters per layer and reduced area consumption on the FPGA fabric while maintaining comparable classification performance.

C. Hybrid Inference

The system presented in Fig. 5 illustrates an example of the hybrid inference approach [10]. The wearable unit in this system acts as a middle layer to detect abnormal data among the raw EEG recorded signals provided by the implant, that can be fed to a subsequent classification stage at the stationary layer. The classifier implemented on the FPGA comprises three main blocks: (1) an one-class (OC) SVM classifier that determines the occurrence of abnormal seizure events; (2) a feature extractor, identical to the one used in the implanted device, for debugging purposes; (3) an ARM Cortex-A9 soft processor for data handling, hosting the TCP/IP server and facilitating SVM training weight updates.

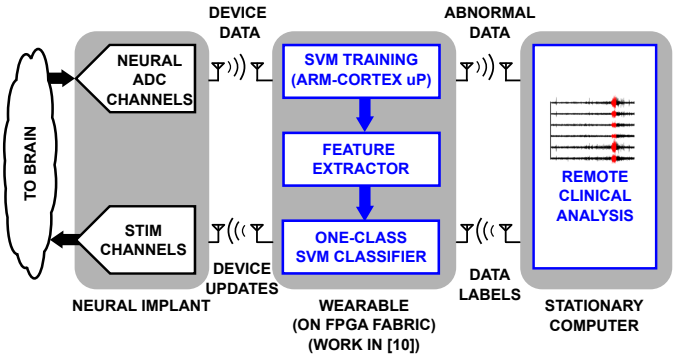

Fig. 5. Example of a closed-loop CNS interface hybrid inference split between implant and wearable: ML microserver using an one-class support vector machine classifier. This system sends abnormal EEG events for clinical assessment, and re-trains the implantable device (Adapted from [10]).

This system demonstrates the efficacy OC-SVMs in accurately labeling intricate intracranial EEG recordings and training supervised learning models, based on either the clinician's feedback or a further ML inference block implemented on the stationary computer. The introduction of patient-localized wearable units addresses the imperative need for continual learning in personalized biomedical devices.

III. Conclusion

This paper has illustrated the design trade-offs involved in implementing AI-based inference algorithms in energy-constrained implantable medical devices. We discussed several examples where ML blocks are implemented either in-implant, remotely on-wearable, or using a hybrid approach. This discussion provides valuable insights for designers to navigate the trade-offs related to latency, computational complexity, and energy consumption, which are commonly encountered in closed-loop implantable neural interfaces.

References

[1] M. T. Salam, H. Kassiri, R. Genov, and J. L. Perez Velazquez, "Rapid brief feedback intracerebral stimulation based on real-time desynchronization detection preceding seizures stops the generation of convulsive paroxysms," en, *Epilepsia*, vol. 56, no. 8, pp. 1227–1238, Aug. 2015, ISSN: 00139580. DOI: 10.1111/epi.13064. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/epi.13064 (visited on 06/16/2023).

[2] G. K. Bergey, M. J. Morrell, E. M. Mizrahi, A. Goldman, D. King-Stephens, D. Nair, S. Srinivasan, B. Jobst, R. E. Gross, D. C. Shields, G. Barkley, V. Salanova, P. Olejniczak, A. Cole, S. S. Cash, K. Noe, R. Wharen, G. Worrell, A. M. Murro, J. Edwards, M. Duchowny, D. Spencer, M. Smith, E. Geller, R. Gwinn, C. Skidmore, S. Eisenschenk, M. Berg, C. Heck, P. Van Ness, N. Fountain, P. Rutecki, A. Massey, C. O'Donovan, D. Labar, R. B. Duckrow, L. J. Hirsch, T. Courtney, F. T. Sun, and C. G. Seale, "Long-term treatment with responsive brain stimulation in adults with refractory partial seizures," eng, *Neurology*, vol. 84, no. 8, pp. 810–817, Feb. 2015, ISSN: 1526-632X. DOI: 10.1212/WNL.0000000000001280.

[3] R. S. Fisher, P. Afra, M. Macken, D. N. Minecan, A. Bagić, S. R. Benbadis, S. L. Helmers, S. R. Sinha, J. Slater, D. Treiman, J. Begnaud, P. Raman, and B. Najimipour, "Automatic Vagus Nerve Stimulation Triggered by Ictal Tachycardia: Clinical Outcomes and Device Performance–The U.S. E-37 Trial," eng, *Neuromodulation: Journal of the International Neuromodulation Society*, vol. 19, no. 2, pp. 188–195, Feb. 2016, ISSN: 1525-1403. DOI: 10.1111/ner.12376.

[4] S. Raspopovic, M. Capogrosso, F. M. Petrini, M. Bonizzato, J. Rigosa, G. Di Pino, J. Carpaneto, M. Controzzi, T. Boretius, E. Fernandez, G. Granata, C. M. Oddo, L. Citi, A. L. Ciancio, C. Cipriani, M. C. Carrozza, W. Jensen, E. Guglielmelli, T. Stieglitz, P. M. Rossini, and S. Micera, "Restoring Natural Sensory Feedback in Real-Time Bidirectional Hand Prostheses," *Science Translational Medicine*, vol. 6, no. 222, 222ra19–222ra19, Feb. 2014, Publisher: American Association for the Advancement of Science. DOI: 10.1126/scitranslmed.3006820. [Online]. Available: https://www.science.org/doi/10.1126/scitranslmed.3006820 (visited on 06/23/2023).

[5] P. P. Vu, Z. T. Irwin, A. J. Bullard, S. W. Ambani, I. C. Sando, M. G. Urbanchek, P. S. Cederna, and C. A. Chestek, "Closed-Loop Continuous Hand Control via Chronic Recording of Regenerative Peripheral Nerve Interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 2, pp. 515–526, Feb. 2018, Conference Name: IEEE Transactions on Neural Systems and Rehabilitation Engineering, ISSN: 1558-0210. DOI: 10.1109/TNSRE.2017.2772961.

[6] G. O'Leary, M. R. Pazhouhandeh, M. Chang, D. Groppe, T. A. Valiante, N. Verma, and R. Genov, "A recursive-memory brain-state classifier with 32-channel track-and-zoom Δ^2 Σ ADCs and Charge-Balanced Programmable Waveform Neurostimulators," *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, ISSN: 2376-8606, Feb. 2018, pp. 296–298. DOI: 10.1109/ISSCC.2018.8310301.

[7] G. O'Leary, J. Xu, L. Long, J. S. Filho, C. Tejeiro, M. ElAnsary, C. Tang, H. Moradi, P. Shah, T. A. Valiante, and R. Genov, "26.2 A Neuromorphic Multiplier-Less Bit-Serial Weight-Memory-Optimized 1024-Tree Brain-State Classifier and Neuromodulation SoC with an 8-Channel Noise-Shaping SAR ADC Array," *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, ISSN: 2376-8606, Feb. 2020, pp. 402–404. DOI: 10.1109/ISSCC19947.2020.9062962.

[8] M. ElAnsary, J. Xu, J. S. Filho, G. Dutta, L. Long, A. Shoukry, C. Tejeiro, C. Tang, E. Kilinc, J. Joshi, P. Sabetian, S. Unger, J. Zariffa, P. Yoo, and R. Genov, "28.8 Multi-Modal Peripheral Nerve Active Probe and Microstimulator with On-Chip Dual-Coil Power/Data Transmission and 64 2nd-Order Opamp-Less $\Delta\Sigma$ ADCs," *2021 IEEE International Solid- State Circuits Conference (ISSCC)*, ISSN: 2376-8606, vol. 64, Feb. 2021, pp. 400–402. DOI: 10.1109/ISSCC42613.2021.9365856.

[9] J. Xu, J. S. Filho, S. Nag, L. Long, C. Tejeiro, E. Hwang, G. O'Leary, Y. Huang, M. Kanchwala, M. Abdolrazzaghi, C. Tang, P. Liu, Y. Sui, X. Liu, G. Eleftheriades, J. Zariffa, and R. Genov, "Fascicle-Selective Bidirectional Peripheral Nerve Interface IC with 173dB FOM Noise-Shaping SAR ADCs and 1.38pJ/b Frequency-Multiplying Current-Ripple Radio Transmitter," *2023 IEEE International Solid- State Circuits Conference (ISSCC)*, ISSN: 2376-8606, Feb. 2023, pp. 31–33. DOI: 10.1109/ISSCC42615.2023.10067626.

[10] G. O'Leary, A. O. Abraham, A. K. Kamath, D. Groppe, T. A. Valiante, and R. Genov, "Machine learning microserver for neuromodulation device training," en, *2017 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, Torino, Italy: IEEE, Oct. 2017, pp. 1–4, ISBN: 978-1-5090-5803-7. DOI: 10.1109/BIOCAS.2017.8325548. [Online]. Available: http://ieeexplore.ieee.org/document/8325548/ (visited on 06/15/2023).

[11] K.-W. Yang, K. Oh, and S. Ha, "Challenges in Scaling Down of Free-Floating Implantable Neural Interfaces to Millimeter Scale," *IEEE Access*, vol. 8, pp. 133 295–133 320, 2020, Conference Name: IEEE Access, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.3007517.

[12] 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 Journal of Solid-State Circuits*, vol. 53, no. 11, pp. 3150–3162, Nov. 2018, Conference Name: IEEE Journal of Solid-State Circuits, ISSN: 1558-173X. DOI: 10.1109/JSSC.2018.2869579.

[13] M. R. Pazhouhandeh, G. O'Leary, I. Weisspapir, D. Groppe, X.-T. Nguyen, K. Abdelhalim, H. M. Jafari, T. A. Valiante, P. Carlen, N. Verma, and R. Genov, "22.8 Adaptively Clock-Boosted Auto-Ranging Responsive Neurostimulator for Emerging Neuromodulation Applications," *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, ISSN: 2376-8606, Feb. 2019, pp. 374–376. DOI: 10.1109/ISSCC.2019.8662458.

[14] R. G. L. Koh, M. Balas, A. I. Nachman, and J. Zariffa, "Selective peripheral nerve recordings from nerve cuff electrodes using convolutional neural networks," eng, *Journal of Neural Engineering*, vol. 17, no. 1, p. 016 042, Jan. 2020, ISSN: 1741-2552. DOI: 10.1088/1741-2552/ab4ac4.