2025 IEEE Biomedical Circuits and Systems Conference (BioCAS) | 979-8-3315-7336-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/BioCAS67066.2025.00067

Real-Time, Low-Power Purkinje Cell Spike Sorting for Headstage Systems in Freely Moving Mice

Kiran Fatima*, Vincenzo Romano‡, Christos Strydis‡§ and Muhammad Ali Siddiqi*‡§
*Electrical Engineering Department, Lahore University of Management Sciences, Pakistan
‡Department of Neuroscience, Erasmus University Medical Center, Rotterdam, The Netherlands
§Quantum and Computer Engineering Department, Delft University of Technology, The Netherlands
v.romano@erasmusmc.nl, c.strydis@erasmusmc.nl, m.siddiqi@lums.edu.pk

***Abstract*—Studying Purkinje cell activity is vital for understanding brain function and movement disorders. However, conventional wired headstage setups in mice restrict natural behavior, while transmitting full neural waveforms wirelessly is prohibitively power-intensive. This work presents a lightweight, real-time spike sorting system implemented on the EFM32PG28 microcontroller, capable of classifying *complex spikes* and *simple spikes* from Purkinje cells directly on the headstage. By extracting only relevant spike information, the system enables efficient wireless transmission or local storage. Leveraging knowledge distillation and Matrix Vector Processor (MVP) hardware acceleration, the spike sorter achieves an overall F1-score of 93.43% and completes detection and classification in 0.765 ms. With over 53 hours of continuous operation on a 50-mAh battery, the proposed solution is well-suited for long-duration, untethered cerebellar experiments in freely moving mice.**

***Index Terms*—Spike sorting, Purkinje cell, knowledge distillation, matrix vector processor, brain-machine interface**

I. Introduction

Purkinje cells are the sole output neurons of the cerebellar cortex and play a critical role in motor coordination and learning. Their unique spiking behavior—characterized by frequent Simple Spikes (SS) and infrequent, complex-shaped Complex Spikes (CS) (Figure 1B)—makes them particularly significant for neuroscientific studies [1]. In vivo electrophysiological experiments in mice often rely on high-density extracellular recordings to study these spikes. However, current experimental setups require *wired* headstage systems (Figure 1A), which restrict natural movement, introduce stress artifacts, and limit the fidelity of recorded data [2].

An ideal solution involves transitioning to wireless headstage devices. Yet, transmitting full-bandwidth neural signals wirelessly imposes unsustainable power and bandwidth demands, especially for small, battery-powered systems. A more efficient approach is to perform real-time *spike sorting* (i.e., spike detection and classification) directly on the headstage, and transmit or store only the instances when the spikes occurred. However, this introduces strict computational and energy constraints: the sorter must operate within a 1-ms latency budget (corresponding to the minimum inter-spike interval, as shown in Figure 1B), consume minimal energy, and run on embedded hardware with limited resources.

Existing spike sorting methods [3]–[6] are too power-hungry, require resources unavailable on constrained embedded platforms, and/or often rely on both local field potentials (LFPs) and single-unit action potentials (APs). Furthermore, most existing spike-sorting solutions are designed for generic neural data and do not explicitly account for the unique temporal and morphological features of Purkinje-cell spikes. In particular, distinguishing CS from SS in real time remains challenging due to the rare occurrence and morphological variability of CS events [2].

In this work, we propose a real-time, energy-efficient spike classification system specifically designed for Purkinje-cell recordings, enabling the use of a *tetherless* headstage in mice. Unlike previous approaches, our method performs accurate spike classification using only single-channel AP data, significantly reducing hardware complexity and power consumption without compromising accuracy. The system is implemented on the EFM32PG28 ultra-low-power microcontroller (MCU) and leverages knowledge distillation [7] from a high-performance deep learning model to a lightweight student model optimized for embedded inference. By utilizing the on-chip Matrix Vector Processor (MVP) accelerator of the MCU, the proposed system enables real-time and prolonged wireless operation on headstage devices and supports scalable neuroscience experiments in freely moving mice.

II. Related Work

Several commercial and research-grade systems have been developed for neural signal acquisition and processing in freely moving animals. These include solutions like FreeLynx [8], CerePlex Exilis [9], and eCube [10], as well as custom FPGA-based platforms [11]–[13]. These systems are either too bulky for use in mice or lack the autonomy needed to support experiments in mice lasting up to 24 hours. Consequently, they are unsuitable for cerebellar-specific studies in mice that require accurate discrimination between SS and CS.

A small but growing body of work has specifically addressed Purkinje cell spike sorting. Sedaghat-Nejad et al. [5], Markanday et al. [3], Zur et al. [4], and Michikawa et al. [6] have proposed software-based solutions that classify Purkinje spikes—particularly complex spikes—with promising accuracy. However, these approaches rely on offline processing using GPUs or high-performance PCs, making them unsuitable for real-time or embedded applications.

Siddiqi et al. [2] recently proposed a promising lightweight ASIC solution for Purkinje spike classification. While their de-

979-8-3315-7336-2/25/$31.00 ©2025 IEEE
DOI 10.1109/BioCAS67066.2025.00067

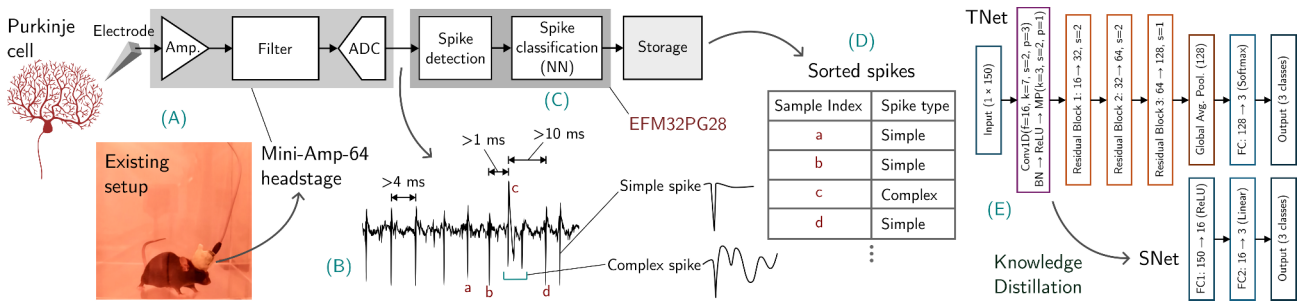

Fig. 1. Overview of the proposed approach. (A) Conventional wired headstage setup used in cerebellar experiments on mice. (B) Purkinje cell AP waveform showing simple and complex spikes. (C) Spike classification performed on EFM32PG28 using its MVP accelerator, following spike detection. (D) Spike-sorting output stored as spike type and sample index in on-board memory. (E) A CNN-based teacher model trained for spike classification, distilled into a lightweight student model for real-time, low-power deployment. *f*: no. of filters, *k*: kernel size, *s*: stride, *p*: padding, *MP*: MaxPooling, *FC*: fully connected layer.

sign demonstrates potential, it has not yet been fabricated and lacks reconfigurability. Therefore, developing new headstage technologies for mice that are lightweight, energy-efficient, reconfigurable, wireless, and capable of real-time processing remains a timely and important research priority.

III. Proposed Solution

This work addresses the aforementioned challenges by presenting a low-power, real-time Purkinje cell spike-sorting system based on a knowledge-distilled neural network deployed on a 32-bit MCU.

A. Experimental Setup

The single-channel AP dataset comprised 1,748,869 SS and 25,713 CS from 98 Purkinje cells in mice, recorded using a wired setup with the Mini-Amp-64 headstage (Cambridge NeuroTech) [14]. The Mini-Amp-64 incorporates silicon neural probes and the Intan RHD2000 chip, which integrates amplifiers, an ADC, and filters [15]. The ADC operates at a 24.414 kHz sampling rate with 10-bit resolution.

B. Implementation Platform

The hardware platform used for implementation and testing the spike-sorting system is the Silicon Labs EFM32PG28 Pro Kit. The EFM32PG28 MCU features a 32-bit ARM Cortex-M33 core and a hardware AI/ML accelerator (MVP), enabling faster inference at the edge with reduced power consumption [16]. It also includes a range of standard peripherals and can interface seamlessly with the aforementioned ADC via an SPI (Serial Parallel Interface) connection.

C. System Architecture

After the MCU receives the digitized neural data via SPI, it passes the signal through a detector function that identifies the occurrence of a potential spike (see Figures 1C and 2). The detected spike is then forwarded to an embedded neural network (NN), which classifies it in real time as either a Complex Spike (CS) or a Simple Spike (SS). The classification result can either be stored in on-board non-volatile memory or transmitted wirelessly. For the purposes of this paper, we assume the former (see Figure 1D).

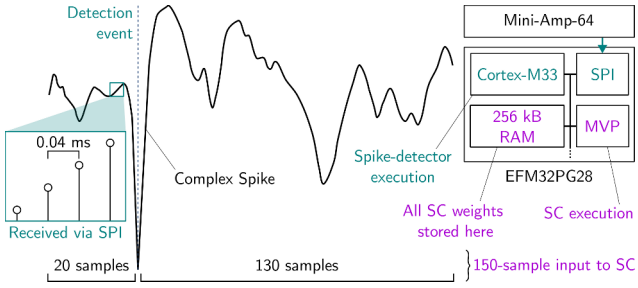

Fig. 2. The NN spike classifier (SC) receives 20 samples before and 130 after the detection point as input. Right: Functional mapping on EFM32PG28.

TABLE I
Comparison of Teacher Networks

TNet*	Architecture	# of Parameters	SS F1-score	CS F1-score
1	150, 16, 32, 64, 128, 256, GAP, 3	439.8K	0.99	0.96
2	150, 16, 32, 64, 128, GAP, 3	109.4K	0.99	0.95
3	150, 16, 32, 64, GAP, 3	26.5K	0.99	0.94

'GAP': Global Average Pooling.
* Teacher NNs are CNNs with some residual blocks.

D. Spike Detector

The spike detector employed in this work is based on the design proposed by Yang et al. [17] for real-time operation. It first applies an exponential infinite impulse response (IIR) filter to the digitized neural signal to attenuate high-frequency noise. Spike detection is then carried out using the Non-Linear Energy Operator (NEO) [18], which estimates the instantaneous energy of the signal. This output is compared against a dynamically updated threshold, and a spike is detected when the NEO response exceeds this threshold.

E. Classifier Design and Knowledge Distillation

The classification system adopts a knowledge distillation framework [7], where a high-capacity *teacher network* (TNet) is trained for accuracy, and a compact *student network* (SNet) is optimized for deployment efficiency (see Figure 1E). The final SNet is quantized and implemented using TensorFlow Lite for MCUs (TFLM) and EFM32PG28's *on-chip* MVP

TABLE II
COMPARISON OF STUDENT NETWORKS

SNet	Architecture	# of Params	F1-score SS	F1-score CS	CR	Arena Size*
1 (CNN)	150, 16, 32, 64, 128, SE, GAP, 3	20.3K	0.98	0.93	5.4×	9.6 MB
2 (CNN)	150, 8, 16, 32, GAP, 3	5.6K	0.98	0.90	19.5×	419 KB
3 (MLP)	150, 64, 32, 16, 3	12.3K	0.98	0.90	8.9×	54.3 KB
4 (MLP)	150, 16, 3	2.5K	0.97	0.89	43.7×	13.9 KB

'SE': Squeeze-and-Excitation block
'CR': Compression Ratio (# of TNet Parameters / # of SNet Parameters)
* Memory required for neural-network inference on the MCU.

(Figure 2) to enable fast, low-power inference on embedded hardware.

IV. RESULTS

A. Classifier Design Space Exploration

Multiple Convolutional Neural Network (CNN)-based TNet architectures were designed and evaluated to achieve an optimal balance between classification performance and model complexity. Overly large models were avoided to ensure that the distilled SNet generalizes well. Each CNN has 150 input neurons, corresponding to 150 samples from a spike window (see Figure 2), and produces 3 outputs representing classifications for SS, CS, and false detections.

The neural dataset was partitioned into 40% for training, 20% for validation, and 40% for testing. Training was conducted using the Adam optimizer with an initial learning rate of 0.001, which was decayed by a factor of 0.1 every 10 epochs. A weighted cross-entropy loss function was used to address class imbalance (Class SS: 0.1, Class CS: 10), with a batch size of 128 over 50 epochs.

Table I summarizes the performance and complexity of the different TNet variants on *test* data. Among these, `TNet-2` was selected for its favorable trade-off between accuracy and complexity and was therefore used to train compact SNets via knowledge distillation in a one-teacher, one-student setup [7]. Each SNet was trained over 20 epochs using the Stochastic Gradient Descent optimizer with a momentum coefficient of 0.7, a batch size of 32, and a learning rate of 0.0001. Kullback-Leibler loss with a temperature of 2.0 and a weighting factor of 1.0 was employed to guide the distillation process. To ensure deployability on MCUs, the SNets were designed with minimal layers and parameters to meet stringent memory constraints. These models were made quantization-friendly by avoiding complex activation functions and leveraging depth-wise separable convolutions, grouped convolutions, and bottleneck layers to minimize computational load.

Table II presents the *tested* SNets, their classification performance, and memory footprints. Among these, only `SNet-3` and `SNet-4`, both implemented as Multilayer Perceptrons (MLPs), fit within the 256-kB RAM constraint of the EFM32PG28 MCU. `SNet-4` was selected over `SNet-3`, as the latter's marginal accuracy gain did not justify its higher memory usage. `SNet-4` was quantized to 8-bit precision and deployed on EFM32PG28 and its MVP using TFLM, enabling energy-efficient, real-time spike classification on-device.

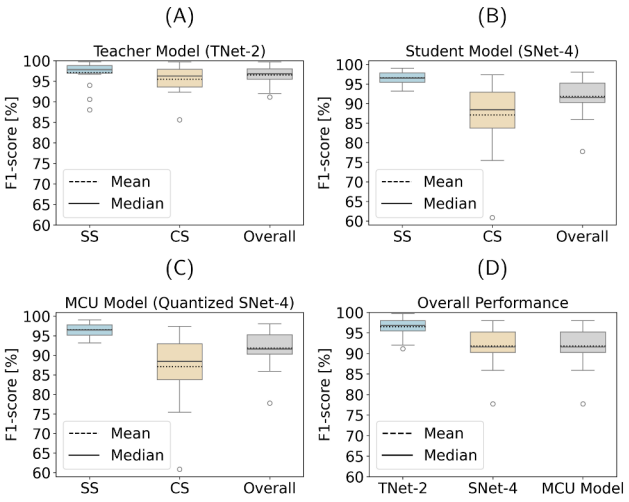

Fig. 3. F1-score distribution across spike types and NN models.

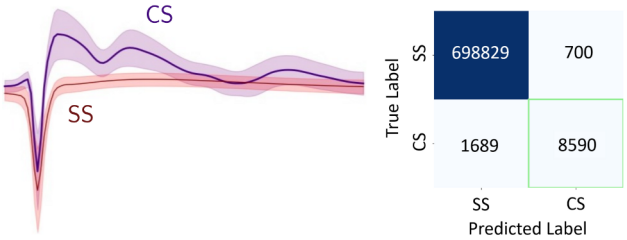

Fig. 4. Left: Overlay plots of spikes classified by the quantized `SNet-4`. Right: Confusion matrix of the quantized `SNet-4` classifier (SS vs. CS).

TABLE III
SPIKE-SORTING PERFORMANCE WITH/WITHOUT MVP

Freq. (MHz)	With MVP I_{avg} (mA)	With MVP t_I (ms)	With MVP E_I (μJ)	Without MVP I_{avg} (mA)	Without MVP t_I (ms)	Without MVP E_I (μJ)
7	0.63	1.40	1.66	0.66	8.18	9.72
13	**0.83**	**0.73**	**1.09**	0.85	4.39	6.72
16	0.92	0.61	1.01	0.95	3.57	6.10
19	1.01	0.48	0.87	1.04	3.02	5.65
26	1.20	0.37	0.80	1.24	2.19	4.89
32	1.37	0.31	0.76	1.42	1.80	4.60
39	1.73	0.24	0.74	1.79	1.41	4.54

I_{avg}: Average current consumption
t_I: Inference time
E_I: Energy consumption per inference, at a supply of 1.8 V

B. Classifier Evaluation

The primary challenge in Purkinje cell spike classification lies in accurately detecting complex spikes due to their intricate morphology and lower occurrence frequency compared to simple spikes. Despite this, our teacher model achieves a high overall mean F1-score of **96.5%** across individual Purkinje cells, as shown in Figure 3A, demonstrating consistently strong performance across both spike types. Figure 3B illustrates the F1-score distribution for the student model, which, despite being **43.7**× smaller in complexity, maintains a high overall F1-score of **91.8%**, albeit with a reduction in CS F1-

TABLE IV
COMPARISON WITH THE STATE OF THE ART

	[8]	[19]	[9]	[10]	[11]	[12]	[13]	[5]	[3]	[4]	[6]	[2]	This work
CS-classification F1-score (%)	N/A_1							–	92.5	–	96.6	90.15	86.65
SS-classification F1-score (%)	N/A_1							–	–	–	–	94.93	96.48
Autonomy (hrs)	3	1.6	2.5	0.5	1.75	72	24	N/A_2				24	53.7
Targeted animal	rats	mice	rats	mice	mice	cats	monkeys	N/A_2				mice	mice

'–': Not provided, 'N/A_1': Not applicable—does not target Purkinje cells; 'N/A_2': Not applicable—software-only, unsuitable for headstage use.

score to **86.6%**. Figure 3C summarizes the performance of the quantized student model deployed on the EFM32PG28 MCU and its MVP using TFLM. The hardware implementation achieves performance comparable to its software counterpart, with no observable degradation following 8-bit quantization. Finally, Figure 3D compares the teacher, student, and MCU-deployed models. While the teacher model demonstrates the highest overall accuracy, the observed drop in CS performance for the student model reflects a deliberate design trade-off made to enable an ultra-low-power, real-time implementation on the embedded EFM32PG28 platform.

Figure 4 (left) shows the superimposed waveforms of the classified CS and SS spikes by the quantized `SNet-4`, illustrating the high quality of classification. The consolidated confusion matrix is presented in Figure 4 (right).

C. Real-Time Performance

Purkinje cell spikes occur approximately every 10 ms on average [2]. However, the interval between a simple spike and a subsequent complex spike can be as short as 1 ms, necessitating that the spike sorting system complete both detection and classification within this window. To meet these real-time constraints, current consumption and inference time were evaluated across various MCU clock frequencies. As shown in Table III, a clock frequency of 13 MHz was selected for the final implementation, offering an optimal trade-off between power efficiency and processing speed. At this frequency, the spike detector requires a maximum of 0.035 ms—well below the 0.04-ms sample period shown in Figure 2—while the classifier completes execution in 0.73 ms, resulting in a total inference time of 0.765 ms. This ensures reliable operation even in the worst-case scenario of a 1-ms inter-spike interval.

Table III also presents a comparison with an implementation that does not utilize the MVP accelerator and instead runs entirely on the CPU (Cortex-M33). The results show that the EFM32PG28's MVP significantly improves matrix operation efficiency by acting as a dedicated processing unit that offloads computations from the main CPU. This shortens neural-network inference time, and enables the system to meet stringent real-time constraints with minimal energy consumption.

D. Device Autonomy

The next step is to evaluate the battery life of the complete headstage to demonstrate its suitability for long-duration experiments (i.e., exceeding 24 hours). Figure 5 shows the estimated operational time based on the combined current consumption of the Intan-based ADC (see Section III-A),

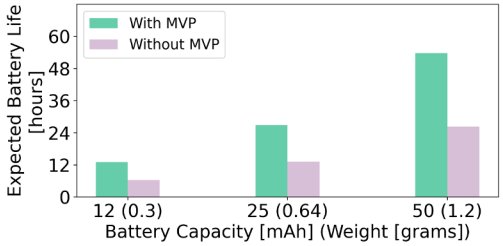

Fig. 5. Battery life of the headstage at typical miniature battery capacities.

ultra-low-power non-volatile memory (STT-RAM [20]), and the MCU running both the detector and classifier. Results indicate that the proposed system can operate continuously for over 53 hours when powered by a lightweight 50 mAh miniature battery, underscoring its viability for sustained use in resource-constrained headstage devices. Moreover, utilizing the MVP approximately **doubles** battery life compared to a Cortex-M33-only solution, which requires a higher clock frequency (48 MHz) to satisfy real-time constraints.

Table IV compares the proposed solution with related work.

V. CONCLUSIONS

This work presents a real-time, energy-efficient Purkinje-cell spike-sorting system tailored for embedded operation on headstage devices used in small animal experiments. By leveraging knowledge distillation and hardware acceleration via the MVP, the proposed system achieves a strong balance between performance and efficiency. It delivers competitive F1-scores of 86.65% for complex spikes and 96.48% for simple spikes, while maintaining a total inference time under 0.8 ms. Additionally, the complete system demonstrates a battery autonomy of over 53 hours on a lightweight 50 mAh cell, highlighting its suitability for long-duration, tetherless cerebellar recordings in freely moving mice. These features position the solution as a practical step toward realizing fully wireless electrophysiology platforms that can support advanced neuroscience research in naturalistic settings.

ACKNOWLEDGMENTS

This work was supported by the Faculty Initiative Fund (FIF-876) at the Lahore University of Management Sciences and by the DBI2 project under the NWO Gravitation Programme (no. 024.005.022).

REFERENCES

[1] V. Romano, A. L. Reddington, S. Cazzanelli, R. Mazza, Y. Ma, C. Strydis, M. Negrello, L. W. Bosman, and C. I. De Zeeuw, "Functional convergence of autonomic and sensorimotor processing in the lateral cerebellum," *Cell reports*, vol. 32, no. 1, p. 107867, 2020.

[2] M. A. Siddiqi, D. Vrijenhoek, L. P. Landsmeer, J. van der Kleij, A. Gebregiorgis, V. Romano, R. Bishnoi, S. Hamdioui, and C. Strydis, "A lightweight architecture for real-time neuronal-spike classification," in *Proceedings of the 21st ACM International Conference on Computing Frontiers*, 2024, pp. 32–40.

[3] A. Markanday, J. Bellet, M. E. Bellet, J. Inoue, Z. M. Hafed, and P. Thier, "Using deep neural networks to detect complex spikes of cerebellar purkinje cells," *Journal of neurophysiology*, vol. 123, no. 6, pp. 2217–2234, 2020.

[4] G. Zur and M. Joshua, "Using extracellular low frequency signals to improve the spike sorting of cerebellar complex spikes," *Journal of Neuroscience Methods*, vol. 328, p. 108423, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027019302808

[5] E. Sedaghat-Nejad, M. A. Fakharian, J. Pi, P. Hage, Y. Kojima, R. Soetedjo, S. Ohmae, J. F. Medina, and R. Shadmehr, "P-sort: an open-source software for cerebellar neurophysiology," *Journal of neurophysiology*, vol. 126, no. 4, pp. 1055–1075, 2021.

[6] T. Michikawa, K. Isobe, and S. Itohara, "A spike-sorting method based on hierarchical clustering to discriminate extracellularly recorded simple spikes and complex spikes from cerebellar purkinje cells," *bioRxiv*, pp. 2021–09, 2021.

[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[8] I. NeuraLynx, *FreeLynx - User Manual*, 2021, accessed: 2025-02-18.

[9] L. Blackrock Microsystems, *CerePlex Exilis - Instructions for Use*, 2020, accessed: 2025-02-18.

[10] L. White Matter, "Details of the ecube standalone and stacking headstages," 2020, accessed: 2025-02-18. [Online]. Available: https://docs.white-matter.com/docs/ecube/hardware/headstages/

[11] G. Gagnon-Turcotte, Y. LeChasseur, C. Bories, Y. Messaddeq, Y. De Koninck, and B. Gosselin, "A wireless headstage for combined optogenetics and multichannel electrophysiological recording," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 11, no. 1, pp. 1–14, 2017.

[12] L. Grand, S. Ftomov, and I. Timofeev, "Long-term synchronized electrophysiological and behavioral wireless monitoring of freely moving animals," *Journal of Neuroscience Methods*, vol. 212, no. 2, pp. 237–241, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027012004189

[13] S. Luan, I. Williams, M. Maslik, Y. Liu, F. De Carvalho, A. Jackson, R. Q. Quiroga, and T. G. Constandinou, "Compact standalone platform for neural recording with real-time spike sorting and data logging," *Journal of Neural Engineering*, vol. 15, no. 4, p. 046014, may 2018. [Online]. Available: https://dx.doi.org/10.1088/1741-2552/aabc23

[14] Cambridge NeuroTech, *Mini-Amp-64 User Guide – Version 1.0*, 2020.

[15] Intan Technologies, LLC, *RHD2000 Series Digital Electrophysiology Interface Chips*, 2023.

[16] Silicon Labs, "Efm32pg28 datasheet," Mar. 2023, [Online; accessed 19-Feb-2025]. [Online]. Available: https://www.silabs.com/documents/public/data-sheets/efm32pg28-datasheet.pdf

[17] Y. Yang and A. J. Mason, "Hardware efficient automatic thresholding for neo-based neural spike detection," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 4, pp. 826–833, 2016.

[18] S. Mukhopadhyay and G. Ray, "A new interpretation of nonlinear energy operator and its efficacy in spike detection," *IEEE Transactions on biomedical engineering*, vol. 45, no. 2, pp. 180–187, 1998.

[19] G. Bilodeau, G. Gagnon-Turcotte, L. L. Gagnon, I. Keramidis, I. Timofeev, Y. De Koninck, C. Ethier, and B. Gosselin, "A wireless electro-optic platform for multimodal electrophysiology and optogenetics in freely moving rodents," *Frontiers in Neuroscience*, vol. 15, 2021. [Online]. Available: https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2021.718478

[20] Everspin Technologies. (2024) Spin-transfer Torque MRAM Technology. [Online]. Available: https://www.everspin.com/