An Accurate and Hardware-Efficient Dual Spike Detector for Implantable Neural Interfaces

Xiaorang Guo*†‡, MohammadAli Shaeri* and Mahsa Shoaran*
xiaorang.guo@tum.de {mohammad.shaeri, mahsa.shoaran}@epfl.ch
*Institute of Electrical and Micro Engineering, Center for Neuroprosthetics, EPFL, 1202 Geneva, Switzerland
†Faculty of Electrical and Computer Engineering, Technische Universität Dresden, 01069 Dresden, Germany
‡Chair of Computer Architecture and Parallel Systems, Technische Universität München, 85748 Garching, Germany

2022 IEEE Biomedical Circuits and Systems Conference (BioCAS) | 978-1-6654-6917-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/BioCAS54905.2022.9948602

***Abstract*—Spike detection plays a central role in neural data processing and brain-machine interfaces (BMIs). A challenge for future-generation implantable BMIs is to build a spike detector that features both low hardware cost and high performance. In this work, we propose a novel hardware-efficient and high-performance spike detector for implantable BMIs. The proposed design is based on a dual-detector architecture with adaptive threshold estimation. The dual-detector comprises two separate TEO-based detectors that distinguish a spike occurrence based on its discriminating features in both high and low noise scenarios. We evaluated the proposed spike detection algorithm on the Wave_Clus dataset. It achieved an average detection accuracy of 98.9%, and over 95% in high-noise scenarios, ensuring the reliability of our method. When realized in hardware with a sampling rate of 16kHz and 7-bits resolution, the detection accuracy is 97.4%. Designed in 65nm TSMC process, a 256-channel detector based on this architecture occupies only $682\mu m^2$/Channel and consumes $0.07\mu W$/Channel, improving over the state-of-the-art spike detectors by 39.7% in power consumption and 78.8% in area, while maintaining a high accuracy.**

***Index Terms*—Spike detection, dual-detector, on-chip, neural signal processing, high-density, brain-machine interface (BMI)**

I. Introduction

In a brain network, neurons mainly communicate through brief electrical pulses known as action potentials or spikes [1]. Traditionally, a brain-machine interface (BMI) records multi-channel (100–200) neural activity from brain regions associated with a target task. Next, spike occurrences are detected as the preliminary processing step for neural decoding. Future-generation BMIs, on the other hand, will be implantable high-density systems [2] with the capability of on-chip data processing [2]–[5]. The detection accuracy should be high in various signal scenarios to be reliable for clinical use. Furthermore, the design should be low-power and area-efficient to be integrable on next-generation highly-miniaturized neural microchips [6].

To date, various spike detectors have been proposed for implantable BMIs [7]–[16]. The simple *absolute thresholding (AT)* method compares the signal amplitude with a predefined threshold level [10]. Despite being hardware-efficient, this method is highly sensitive to noise, which leads to poor performance at low signal-to-noise ratios (SNRs). To address this issue, the *dual vertex threshold (DVT)* algorithm uses a positive and a negative threshold for spike detection [15], [16]. Detectors based on data transformation techniques have also been proposed to improve performance. In [9], the moving average energy (MAE), which calculates the instant signal energy over a sliding window, was proposed for spike detection. The *Teager energy operator (TEO)* (*a.k.a.*, nonlinear energy operator, NEO) and its variants [11]–[13] are among the popular data transformation techniques widely used in spike detectors. Moreover, time-frequency analysis methods such as *discrete wavelet transform (DWT)* and *stationary wavelet transform (SWT)* have been reported for spike detection [7], [17], [18]. Other designs combined wavelets and TEO to achieve a high performance [8], [19]. However, such methods are computationally complex for integration on implantable devices with limited hardware resources [20]–[22].

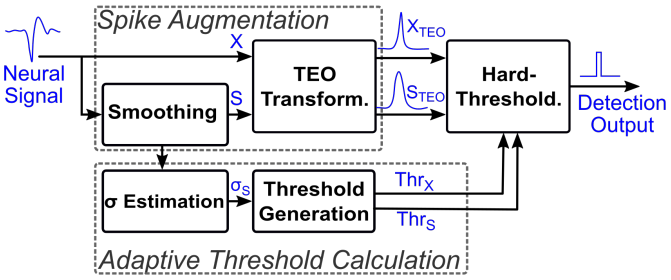

Fig. 1. The generic block diagram of the proposed dual spike detector.

Here, we present the design and implementation of a hardware-efficient dual spike detector with online threshold estimation. The dual detector smooths the neural signal to reduce its high-frequency noise content. Next, TEO transformation is applied to the original and smoothed signals to enhance spike detectability. A spike is detected if either of the aforementioned signals exceeds the associated threshold level. The proposed dual detector achieves a high detection accuracy even in high-noise scenarios thanks to the dual-path detection method. To show the scalability of the design, we implemented a 256-channel digital spike detector based on this architecture, achieving state-of-the-art accuracy and hardware efficiency. The remainder of this paper is organized as follows. We introduce the proposed detection algorithm and adaptive thresholding method in Section II. Section III presents the hardware implementation. Simulation results are presented in Section IV, followed by a conclusion in Section V.

II. Spike Detection Method

The proposed dual-detection method comprises two separate spike detectors working in parallel. In one path, the TEO of

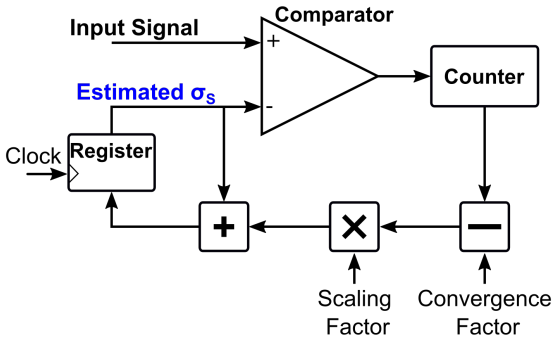

Fig. 2. Block diagram of the on-chip standard deviation estimator for adaptive threshold calculation.

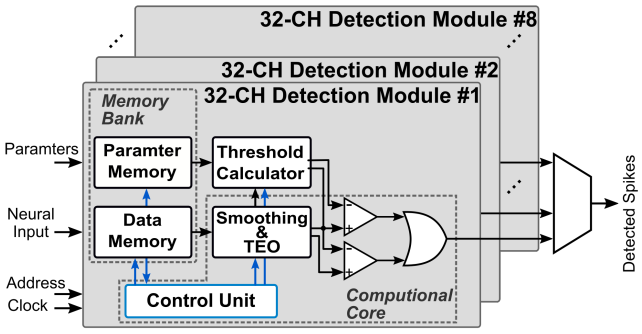

Fig. 3. The hardware architecture of the 256-channel dual spike detector.

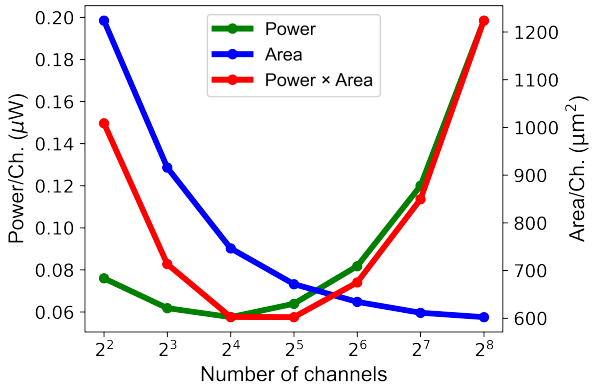

Fig. 4. Hardware design optimization via area and power co-analysis.

the input neural signal is calculated to highlight sharp peaks, followed by an adaptive thresholding procedure to distinguish the spikes from background noise. The TEO transformation is formulated as follows:

$$\mathscr{T}\{X\}[k] = X[k]^2 - X[k+1]X[k-1], \tag{1}$$

where $X[k]$ and $\mathscr{T}\{X\}[k]$ represent the input and TEO signals, respectively [18]. Since TEO accentuates sharp signal variations, it is only helpful when the high-frequency background noise is limited. To detect spiking activity even in the presence of high-frequency noise, we included a second data path in our design. In the second path, neural signals are first smoothed to reduce the high-frequency noise. The smoothed signal is then fed to a TEO-based spike detector. The larger the window size, the better the smoothing function can be performed. However, a large window size demands a larger memory that increases hardware utilization. Considering the trade-off between performance and hardware cost, we set the window size to two samples in this design. Combining the results from both paths, a spike event will be raised if either detector captures above-threshold activities. Fig. 1 shows the block diagram of the proposed method. As shown later in this paper, the dual-detection strategy leads to a significant improvement in detection accuracy, particularly in high-noise scenarios.

In a neural interface, the recorded signals are susceptible to drift over time. In order to adjust the detection threshold to various changes in the signal, we designed an adaptive threshold estimator to calculate the threshold level in an online fashion on the chip. Due to the different characteristics of TEO-augmented signals (X_{TEO}, S_{TEO}), specific threshold levels are computed for each signal. We formulate the threshold levels as follows:

$$\begin{aligned} Thr_X &= C_1 \times \sigma_S, \\ Thr_S &= C_2 \times \sigma_S + C_3 \times \sigma_S^2, \end{aligned} \tag{2}$$

where Thr_X and Thr_S indicate the threshold levels for X_{TEO} and S_{TEO} signals (shown in Fig. 1), respectively. σ_S denotes the estimated standard deviation of signal 'S', C_1, C_2 and, C_3 are the coefficients calculated via co-optimization of the hardware and detection accuracy. The typical values of these coefficients are in the form of powers of 2 and can be implemented with only a few shifts and additions rather than complex multiplications.

To calculate the statistical 'standard deviation' or std, many signal samples are required which could increase the hardware complexity. The number of data samples higher than std in a large neural dataset is proportional to the statistical standard deviation [7]. Therefore, based on this concept, we modified the estimation method introduced in [7] to improve the hardware efficiency for online std calculation. Fig. 2 illustrates the block diagram of the std estimator. The smoothed signal is first compared with an initial, predefined value of σ_S. The counter then calculates the number of samples higher than σ_S. This one-bit stream is subsequently accumulated every 256 clock cycles, and the result is subtracted by a 'convergence factor' that is determined based on the empirical data distribution. Following scaling, σ_S will be updated by the difference between the number of σ_S-exceeded samples and the associated convergence factor through the feedback loop, as shown in Fig. 2. In other words, this feedback loop is designed to ensure the convergence of the number of samples exceeding σ_S to the convergence factor. The small values of 'convergence factor' and 'scaling factor' could make the σ_S-estimation process more stable, but at the cost of higher latency. Here, we chose 20 and 0.001 for the convergence and scaling factors, respectively. Following convergence, we calculate two separate threshold values based on Eq. 2.

III. Hardware Implementation

Fig. 3 illustrates the modular architecture of the proposed 256-channel dual spike detector. In this design, arithmetic

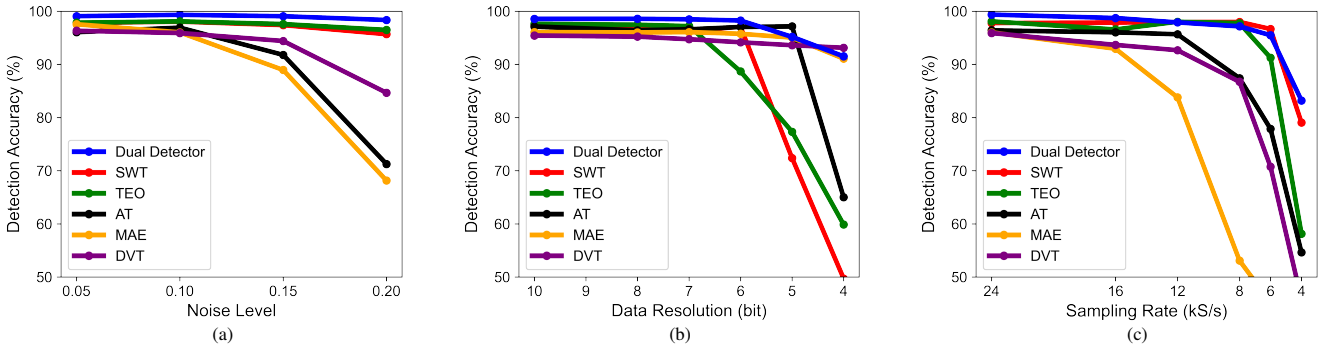

Fig. 5. The mean detection accuracy vs. (a) noise level, (b) data resolution, and (c) sampling rate for the proposed dual-detector and previous methods (algorithm-level simulation). For (b) and (c) the noise level is fixed at 0.1.

operations are executed sequentially on the incoming channel data (i.e., 256 digitized channels). In order to reduce the hardware cost, the computational blocks such as TEO and threshold calculator are shared among every 32 channels. Here, five register banks store the channel data and threshold levels. Also, a multiplexer is used to consolidate the outputs of the 32-channel modules. As a result, the 1-bit data stream (detected spikes) will be generated via multiplexing the module outputs in the time domain. The original (X) and smoothed (S) neural signals are truncated to 7 and 6 bits, respectively, to further improve the hardware efficiency. Moreover, the TEO outputs are truncated to 8-bit (X_{TEO}) and 9-bit (S_{TEO}) signals in order to reduce the memory requirements for thresholds calculation. This can significantly improve the hardware efficiency at the cost of near 1% drop in detection accuracy.

As shown in Fig. 4, there is a trade-off between the occupied area and power consumption of the proposed 256-channel spike detector. Hardware sharing decreases the area per channel as the channel count increases. Similarly, the power per channel decreases as a function of channel count at low channel densities (i.e., <16). This trend changes at higher channel counts where the dynamic power begins to dominate the system power. By optimizing the area-power product, we can find the optimal number of channels for hardware sharing (32). The 256-channel detector thus contains eight modules operating in parallel, each handling 32 channels. The system is clocked at 4MHz, while the 32-channel detectors are activated alternatively to reduce dynamic power.

IV. Results

We used the Wave_Clus dataset to assess the performance of our spike detector [23]. The Wave_Clus dataset includes four different subsets named Easy1, Easy2, Difficult1, and Difficult2 sampled at 24kS/s. Each subset has different noise levels, ranging from 0.05 to 0.2 (for Easy1, it can be up to 0.4). The noise level is defined as the standard deviation of noise relative to the average magnitude of spikes. To evaluate the spike detection performance, we used the ratio of correctly detected spikes (true positives, TP) to the total number of detected (TP+FP) and missed (FN) spikes as accuracy criteria.

Fig. 5(a) shows the detection accuracy of the proposed algorithm and previous methods reported in literature versus the noise level of the input signal. When the noise level is at its minimum (0.05), all methods could detect the spikes with high accuracy (>95%). However, by increasing the noise level, the performances of MAE, AT, and DVT considerably decreased. The mean accuracies of MAE, AT, and DVT are 87.7% (97.5-68.1%), 89.0% (96.1-71.2%), and 92.8% (96.3-84.7%), respectively. TEO and SWT achieved a high mean accuracy of 97.5% (97.8-96.5%), and 97.2% (97.8-95.7%). The proposed dual detection method (with a mean accuracy of 98.9%) outperformed all other methods, with accuracies ranging from 99.1% (at high SNR) to 98.4% (at low SNR).

Furthermore, we investigated the impact of input data resolution and sampling rate on the detection performance. Fig. 5(b) shows the detection accuracy versus resolution at a noise level of 0.1. The performances of TEO, AT and SWT were high at high resolutions, with the performance rapidly dropping for lower resolutions. This analysis shows that the proposed method, DVT and MAE are less sensitive to this parameter, obtaining accuracies above 90% even at 4-bit resolution. Fig. 5(c) further illustrates the detection accuracy versus sampling rate for various spike detection methods. As shown in this figure, our proposed method and SWT are less sensitive to sampling rate compared to other methods.

Table. I summarizes the design specifications, detection accuracy, and hardware-level performance of the proposed method and state-of-the-art spike detectors. References [9], [10], [24], [25] present measured results, while the others report simulated hardware results. We designed the 256-channel dual spike detector in a 65nm TSMC process, with an

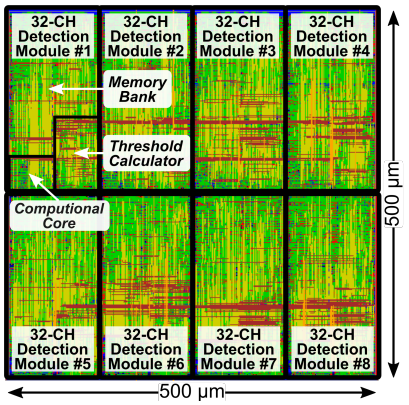

Fig. 6. The physical layout of the 256-channel dual spike detector. The area of each 32-channel module is about 0.031mm^2.

TABLE I: The performance summary of the proposed method and state-of-the-art spike detectors.

Spike Detector	This work	[10]	[7]	[9]	[12]	[13]	[14]	[15]	[24]	[25]
CMOS Process (nm)	65	130	130	180	130	180	40	40	65	22
Method	Dual-detection	AT	SWT	MAE	TEO	ED[§]	PBOTM[*]	DVT	N/A	TEO
Adaptive Threshold	✓	✓	✓	×	✓	✓	×	×	N/A	×
Channel Count	256	10	16	8	64(max)	N/A	16	16	1024	16
Resolution (bits)	7	16	6	8	10	N/A	N/A	12	10	9
Sample Rate (kS/s)	16	20	25	30	20	16	12	24	20	25
Power per Channel (μW/Ch)	0.07	2.96 [†]	1.71	0.116	0.05[‡‡]	5.1	19.0[††]	8.09[†]	2.72	0.29
Area per Channel (mm^2/Ch)	6.82×10^{-4}	0.08[††]	0.014	0.27	1.6×10^{-3}[‡‡]	0.018	0.0175[††]	N/A	0.02	3.22×10^{-3}
Accuracy	97.4%	97.8% [‡]	98%∼99%[¶]	97%[‡]	95%[‡]	95%	98.3%[**]	98.12%[**]	N/A	N/A

[‡] A different dataset was used.
[§] ED represents the *energy of derivative* method that is an approximated calculation of TEO designed in analog.
[*] PBOTM refers to the *preselection Bayes optimal template matching* that simultaneously performs spike detection and spike sorting.
[¶] Overlapping spikes were excluded in the calculation of detection accuracy.
[‡‡] Memory blocks were excluded in the estimation of power and area.
[**] False positives were not considered in the calculation of detection accuracy.
[†] Power is estimated based on the power breakdown in the paper.
[††] Power or area was reported for the whole design, which also includes the spike sorting block, compression block or others.

active area of 0.25mm^2 and power consumption of 17.55μW at a 1.1V supply. Fig. 6 shows the layout of the 256-channel dual spike detector introduced in Fig. 3. The simulated dynamic and leakage powers are 10.8μW and 6.74μW at 16kHz, respectively. The area and power breakdowns are shown in Fig. 7. The register banks (i.e., memory) are the dominant block in terms of both power and area usage, consuming over 68% of the hardware resources in this design.

For a fair comparison against the state-of-the-art, the per-channel area and power consumption of the proposed detector and other similar designs are presented in Table I. The area and power of the dual-detector are 682μm^2/Channel and 70nW/Channel, respectively. These results show 78.8% and 39.7% improvements in area utilization and power consumption, respectively, over the state-of-the-art spike detectors [9], [25]. This comparison confirms that our proposed spike detector is the most hardware-efficient design reported so far. This is a result of extensive hardware optimizations (e.g., block sharing and data truncation). It is worth mentioning that the power consumption of 50nW/Channel reported in [12] did not include the power consumed by memory banks, which could be significant in high-density BMIs.

From the standpoint of detection performance, the proposed dual detector achieved an overall accuracy of 97.4% after hardware optimization. This performance is comparable to the state-of-the-art spike detectors while being achieved at a significantly lower hardware cost. The high accuracy of ∼99% reported in [7] was achieved by excluding the overlapping spikes in the dataset. Moreover, the detection accuracies of [14] and [15], which are slightly higher than our detector, were calculated with a different definition (i.e., the ratio of true positives to the sum of true positives and false negatives). Thus, false positives were discarded in their calculations. In addition, these performances were achieved at significantly higher hardware costs compared to our proposed detector.

V. Conclusion

This paper presents a novel spike detector that benefits from a dual detection architecture. This algorithm achieved nearly 99% accuracy in software simulations, and 97.4% using an efficient hardware implementation. To improve the hardware efficiency, we optimized our system for low sampling rate and data resolution as well as optimal channel count for hardware sharing. In 65nm TSMC process, the dual-detector occupies only 682μm^2/Channel and consumes only 0.07μW/Channel, making it a proper candidate for implantable BMIs. The proposed spike detector outperforms current state-of-the-art spike detectors in terms of hardware efficiency.

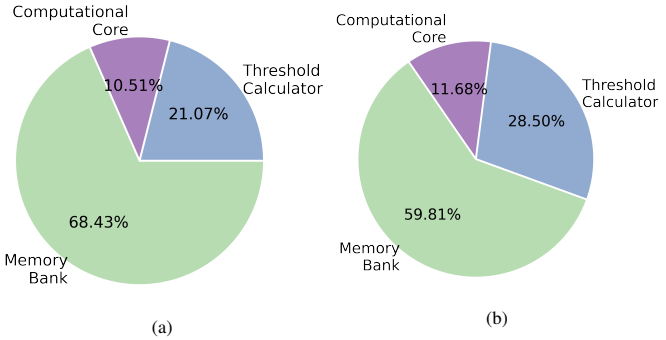

Fig. 7. (a) The area, and (b) power breakdowns of the proposed dual detector. The memory bank contains register banks for storing the internal values, clock gating, and threshold estimation. The computational units represent the hardware responsible for computations such as TEO, smoothing, and comparators. The threshold calculator contains the logic to implement the architecture in Fig. 2 (excluding the registers).

References

[1] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, and A. J. Hudspeth, *Principles of Neural Science*, 5th ed. McGraw-Hill, October 2012.

[2] 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 mm2-1.51 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.

[3] M. Shaeri and A. M. Sodagar, "A framework for on-implant spike sorting based on salient feature selection," *Nature Communications*, vol. 11, no. 3278, pp. 1–9, June 2020.

[4] J. Yoo and M. Shoaran, "Neural interface systems with on-device computing: Machine learning and neuromorphic architectures," *Current opinion in biotechnology*, vol. 72, pp. 95–101, 2021.

[5] M. Shaeri, A. Afzal, and M. Shoaran, "Challenges and opportunities of edge ai for next-generation implantable BMIs," *IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, 2022.

[6] B. Zhu, U. Shin, and M. Shoaran, "Closed-loop neural prostheses with on-chip intelligence: A review and a low-latency machine learning model for brain state detection," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 15, no. 5, pp. 877–897, 2021.

[7] Y. Yang, C. S. Boling, A. M. Kamboh, and A. J. Mason, "Adaptive threshold neural spike detector using stationary wavelet transform in cmos," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 23, no. 6, pp. 946–955, 2015.

[8] N. Nabar and K. Rajgopal, "A wavelet based Teager energy operator for spike detection in microelectrode array recordings," in *TENCON 2009 - IEEE Region 10 Conference*, 02 2009, pp. 1–6.

[9] R. Fiorelli, M. Delgado-Restituto, and Á. Rodríguez-Vázquez, "Charge-redistribution based quadratic operators for neural feature extraction," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 3, pp. 606–619, 2020.

[10] G. Gagnon-Turcotte, I. Keramidis, C. Ethier, Y. De Koninck, and B. Gosselin, "A wireless electro-optic headstage with a 0.13-μm CMOS custom integrated DWT neural signal decoder for closed-loop optogenetics," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 13, no. 5, pp. 1036–1051, 2019.

[11] Z. Zhang and T. G. Constandinou, "Adaptive spike detection and hardware optimization towards autonomous, high-channel-count BMIs," *Journal of Neuroscience Methods*, vol. 354, p. 109103, 2021.

[12] 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, 2017.

[13] S. Dwivedi and A. K. Gogoi, "A novel adaptive real-time detection algorithm for an area-efficient CMOS spike detector circuit," *AEU - International Journal of Electronics and Communications*, vol. 88, pp. 87–97, 2018.

[14] H. Xu, Y. Han, X. Han, J. Xu, S. Lin, and R. C. Cheung, "Unsupervised and real-time spike sorting chip for neural signal processing in hippocampal prosthesis," *Journal of Neuroscience Methods*, vol. 311, pp. 111–121, 2019.

[15] C. Seong, W. Lee, and D. Jeon, "A multi-channel spike sorting processor with accurate clustering algorithm using convolutional autoencoder," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 15, no. 6, pp. 1441–1453, 2021.

[16] P. Li, M. Liu, X. Zhang, and H. Chen, "Efficient online feature extraction algorithm for spike sorting in a multichannel FPGA-based neural recording system," in *2014 IEEE Biomedical Circuits and Systems Conference (BioCAS) Proceedings*, 2014, pp. 1–4.

[17] M. A. Shaeri and A. M. Sodagar, "A method for compression of intra-cortically-recorded neural signals dedicated to implantable brain–machine interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 23, no. 3, pp. 485–497, 2015.

[18] M. Shaeri and A. M. Sodagar, "Data transformation in the processing of neuronal signals: A powerful tool to illuminate informative contents," *IEEE Reviews in Biomedical Engineering*, pp. 1–17, February 2022.

[19] F. Lieb, H.-G. Stark, and C. Thielemann, "A stationary wavelet transform and a time-frequency based spike detection algorithm for extracellular recorded data," *Journal of Neural Engineering*, vol. 14, no. 3, p. 036013, 2017.

[20] B. Zhu, U. Shin, and M. Shoaran, "Closed-loop neural interfaces with embedded machine learning," in *2020 27th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*. IEEE, 2020, pp. 1–4.

[21] M. Shoaran, B. A. Haghi, M. Taghavi, M. Farivar, and A. Emami-Neyestanak, "Energy-efficient classification for resource-constrained biomedical applications," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 4, pp. 693–707, 2018.

[22] B. Zhu, M. Farivar, and M. Shoaran, "Resot: Resource-efficient oblique trees for neural signal classification," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 4, pp. 692–704, 2020.

[23] R. Q. Quiroga, Z. Nadasdy, and Y. Ben-Shaul, "Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering," *Neural computation*, vol. 16, no. 8, pp. 1661–1687, 2004.

[24] D.-Y. Yoon, S. Pinto, S. Chung, P. Merolla, T.-W. Koh, and D. Seo, "A 1024-channel simultaneous recording neural SoC with stimulation and real-time spike detection," in *2021 Symposium on VLSI Circuits*, 2021, pp. 1–2.

[25] S. M A. Zeinolabedin, *et al.*, "A 16-channel fully configurable neural SoC with 1.52W/Ch signal acquisition, 2.79W/Ch real-time spike classifier, and 1.79 TOPS/W deep neural network accelerator in 22nm FD-SOI," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 16, no. 1, pp. 94–107, 2022.