Closed-Loop Neural Interfaces with Embedded Machine Learning

Bingzhao Zhu[1,*], Uisub Shin[1,*], Mahsa Shoaran[2]
[1]School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA
[2]Institute of Electrical Engineering & Center for Neuroprosthetics, EPFL, 1202 Geneva, Switzerland
Email: bz323@cornell.edu; us52@cornell.edu; mahsa.shoaran@epfl.ch
[*]These authors contributed equally to this work.

***Abstract*—Neural interfaces capable of multi-site electrical recording, on-site signal classification, and closed-loop therapy are critical for the diagnosis and treatment of neurological disorders. However, deploying machine learning algorithms on low-power neural devices is challenging, given the tight constraints on computational and memory resources for such devices. In this paper, we review the recent developments in embedding machine learning in neural interfaces, with a focus on design trade-offs and hardware efficiency. We also present our optimized tree-based model for low-power and memory-efficient classification of neural signal in brain implants. Using energy-aware learning and model compression, we show that the proposed oblique trees can outperform conventional machine learning models in applications such as seizure or tremor detection and motor decoding.**

***Index Terms*—Neural interfaces, low-power, machine learning, oblique tree, disease detection, closed-loop stimulation.**

I. Introduction

Closed-loop neural interfaces enhance the therapeutic efficacy while alleviating side effects and improving the battery life in their open-loop counterparts. For example, adaptive stimulation in response to abnormal brain activity has shown promise in treating epileptic seizures [1] and motor symptoms of Parkinson's disease [2]. Recently, the application of machine learning (ML) in closed-loop neural interfaces and its ASIC implementation has gained interest among researchers. Real-time neural processing can be enabled through on-chip feature extraction and classification, followed by a closed-loop feedback (e.g., neurostimulation) to suppress a symptom, provide a sensory feedback, or control a prosthetic device in a brain-machine interface (BMI), as illustrated in Fig. 1. The ASIC realization of ML is particularly favorable in such implants, enabling real-time near-sensor processing, triggering a therapeutic feedback, lowering the data transmission rate, and alleviating security and privacy concerns.

Despite the promise and benefits of machine learning for closed-loop neural interfacing, stringent energy and area constraints on multi-channel neural implants pose significant challenges for the ASIC implementation of ML models. Therefore, it is critical to develop hardware-efficient ML solutions to overcome such limitations. This paper reviews the state-of-the-art neural interfaces with embedded classification and describes several techniques for energy, area and memory efficiency, including single-path inference, cost-aware learning, and model compression. A new class of oblique tree-based models suited for hardware-efficient realization is proposed and verified on three neural signal classification tasks.

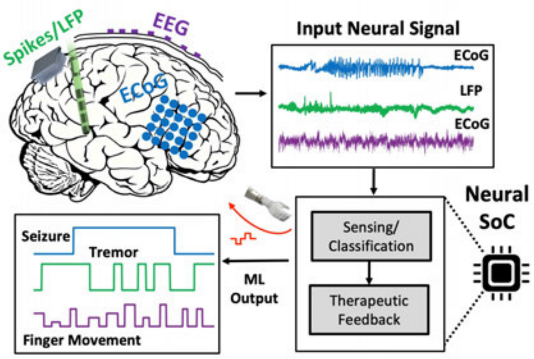

Fig. 1. Block diagram of a closed-loop neural interface; Neural signals are recorded with invasive or noninvasive electrodes, and an embedded classifier detects disease symptoms or decodes a movement. Closed-loop feedback is enabled to suppress an abnormal activity or control a prosthetic device.

II. Neural Interfaces with Embedded ML

Various machine learning algorithms and hardware architectures have been reported for neurological disease detection [3]–[13] and brain-machine interfacing [14]–[16] using either invasive or noninvasive electrodes, as summarized below.

A. Symptom Detection

Accurate detection of symptoms in neurological disorders is the first step toward an effective closed-loop stimulation therapy. A typical example is epileptic seizure detection, where a supervised machine learning algorithm could be used to detect *seizure* events from electrophysiological recordings. The majority of seizure detection SoCs in literature have adopted support vector machine (SVM) classifiers [3], [4], Fig. 2(a). SVMs typically require a large number of multiply-and-accumulate (MAC) and non-linear operations, while their computational and memory resources linearly scale with the number of channels and input features.

Recently, ensembles of decision trees (DTs) such as gradient boosting trees and random forests have emerged as an accurate yet hardware-friendly solution for resource-constrained platforms. With simple comparisons applied to input features, tree-based models enable low-complexity hardware architectures. In [5], an ensemble of eight gradient-boosted DTs achieved an energy efficiency of 41.2nJ/class for 32-channel intracranial EEG (iEEG)-based seizure detection, Fig. 2(b). A sequential feature extraction approach enabled the use of a single feature extraction engine per tree, thus significantly reducing the hardware cost for multi-channel implants. Another tree-based

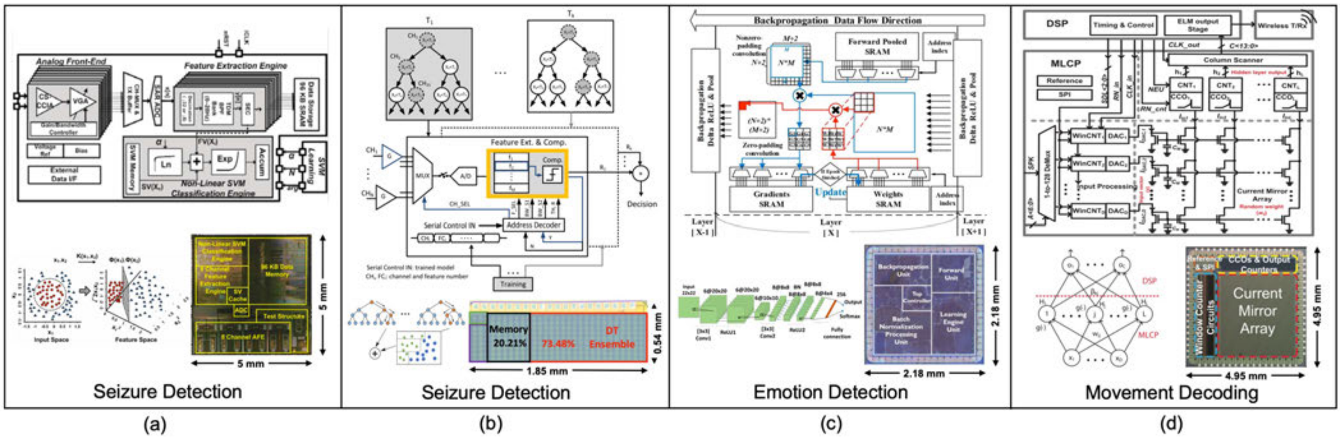

Fig. 2. Hardware block diagram and chip micrograph of learning models for different applications: (a) non-linear SVM for seizure detection [4], (b) gradient-boosted trees for seizure detection [5], (c) CNN for emotion detection [11], and (d) ELM for motor intention decoding [14].

TABLE I
COMPARISON OF MACHINE LEARNING SOCS

Parameter	JETCAS'18 [5]	ISSCC'13 [4]	JSSC'13 [8]	ISSCC'20 [7]	JETCAS'19 [11]	TCAS-I'19 [13]	TBioCAS'16 [14]	TBioCAS'17 [15]	TVLSI'19 [16]
Process	65 nm	180 nm	180 nm	65 nm	28 nm	180 nm	350 nm	130 nm††	65 nm
Classifier	XGB DT	Non-Lin SVM	LLS	AdaBoost DT	CNN	NN-based DT	ELM	DT	K-means
Application	Seizure Det.	Seizure Det.	Seizure Det.	Seizure Det.	Emotion Det.	Sleep Staging	Motor Decoding	Spike Sorting	Spike Sorting
Signal Modality	iEEG	EEG	iEEG	iEEG	EEG	EEG, EMG	Monkey LFP	Rat LFP	Synthetic Spikes
No. of Channels	32	8	8	8	6	2	128	32	128
Energy Eff. (or Power)	41.2 nJ/class.	1.22 μJ/class.**	77.9 μJ/class.	36 nJ/class.	76.61 mW	0.149 mJ/epoch	16.2 nJ/class.	24 μW	22.4 μW
Memory	1 kB	N.A.	N.A.	N.A.	11.8 kB	6.4 kB	N.A.	0.56 kB	4.88 kB
Area*	0.8 mm^2	5.63 mm^2	4.85 mm^2	0.71 mm^2	3.47 mm^2	8.77 mm^2	17.4 mm^2	0.73 mm^2	0.41 mm^2
Sensitivity	83.7%	95.1%	92%	96.7%	83.4%‡	81%‡	99.3%‡	~77%‡	72/86%‡§
Specificity	88.1%	0.27 FA/h†	N.A.	0.8 FA/h†	N.A.	N.A.	N.A.	N.A.	N.A.
Latency	1.79 s	2 s	0.8 s	N.A.	0.45 s	N.A.	N.A.	N.A.	N.A.

∗ Estimated area of feature extractor and classifier from chip micrographs (excluding pads)
† Number of false alarms per hour
‡ Accuracy metric
∗∗ Estimated from power breakdown
†† Post-synthesis results
§ For unsupervised and semisupervised modes, respectively

on-chip classifier was recently reported for epileptic seizure detection [7], where 1024 decision stumps (i.e., trees of depth one) were aggregated using AdaBoost technique. With bit-serial operation and on-chip weight regeneration, the 8-channel SoC reported an energy efficiency of 36nJ/class. Moreover, DT ensembles have shown superior performance in other neural tasks such as Parkinsonian tremor detection using local field potentials (LFP) [9], [17] and migraine state classification from somatosensory evoked potentials (SSEP) [10]. ML models have also been explored for neural signal classification in applications such as emotion detection [11], [12], sleep stage classification [13], and for predicting memory dysfunction [18] and mental fatigue [19] (to potentially trigger a neurostimulation therapy). In [11], a convolutional neural network (CNN) SoC with online training capability was implemented for emotion recognition, Fig. 2(c). To reduce the memory and area utilized by batch processing, training and acceleration were executed in four phases through re-using minibatch data and hardware, at the cost of increased training time. Combined with an external feature extraction processor, the CNN classifier achieved an accuracy of 83.36% in a binary emotion detection task. A 4-layer neural network classifier was recently reported for emotion detection in autistic children [12]. This 2-channel EEG processor achieved a classification accuracy of 85.2% while consuming 10.1μJ/class.

B. Brain-Machine Interfaces

BMIs provide a communication channel between the human brain and external environment for paralyzed patients. Similar to implants for disease detection, BMIs also operate on a resource-constrained platform, making it crucial to design hardware-friendly movement decoders for fully implantable BMIs. A variety of signal modalities such as EEG, ECoG, spikes and LFP can be used as input to a BMI, providing various degrees of motor control and invasiveness. An intra-cortical decoder based on extreme learning machine (ELM) was reported in [14], exploiting the CMOS device mismatch for random weight generation and sub-threshold current-mode operation, as shown in Fig. 2(d). The ELM processor combined with a DSP (performing additional MAC operations and spike sorting) consumed 16.2nJ/class. A scalable DT-based spike sorting processor for high-channel-count BMIs was reported in [15], while [16] presents an area-efficient spike sorting processor with online K-means clustering. The hardware efficiency of the processor was improved using multiplier-less spike detection and feature extraction, time-multiplexed registers, and low-voltage SRAM.

The hardware cost and performance of state-of-the-art classifiers in neural interface SoCs are summarized in Table I. As shown in this table, DTs provide an attractive solution for low-power and area-limited applications. It should be noted that when comparing different ML SoCs, various factors such as classification task, channel count, feature type, and signal modality need to be taken into account. A comparison of different classifiers for iEEG and EEG-based seizure detection can be found in [5], [20], while the hardware complexity of DTs and deep neural networks is compared in [21].

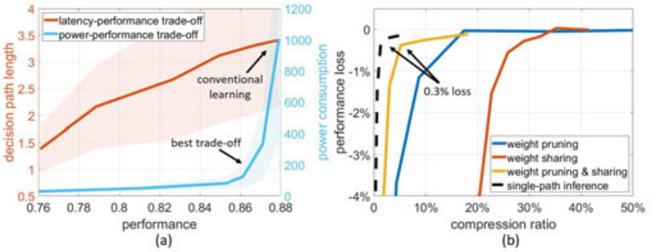

Fig. 3. (a) Trade-off between classification performance and latency/power for seizure detection. The inference power is normalized to the power of most efficient feature (line-length). Shading areas indicate standard deviation across patients; (b) Model compression with weight pruning and sharing. In the single-path scheme, only the parameters used during inference are included. This experiment was conducted for an oblique tree on the MNIST dataset.

III. Hardware-Algorithm Optimized DTs

Compared to the conventional approach of transmitting raw neural data for off-the-body processing, neural implants with embedded ML avoid the use of power-demanding transmitters. Yet, efficient implementation of ML is pivotal to minimize heat dissipation and battery usage. Moreover, small area of the implant and large number of channels require the use of minimal silicon and memory resources. On-chip classifiers enable a fast inference and provide a timely feedback to the patients, leading to a rapid activation of therapeutic stimulation or prosthetic control. Overall, embedded ML models are required to consume low power and small area, while providing a low detection latency and high classification accuracy.

Among the ML algorithms widely used in embedded neural SoCs, tree-based models are compatible with a single-path inference scheme [5], [22] where a single root-to-leaf path is visited to make predictions (i.e., a small portion of the model). This lightweight inference is a significant advantage, considering the large number of channels in modern neural interface platforms. DTs can be further optimized for power- and memory-efficient implementation through a wise co-design of algorithm and hardware, as described below.

A. Cost-Aware Learning

During inference, the major hardware cost (e.g., power, area or latency) of tree-based models is associated with feature extraction [5]. As the inference time increases proportional to the length of a decision path, sequential processing may raise a concern on detection latency. In [5], we used an asynchronous approach to reduce the latency due to single-path sequential processing. A cost-aware learning scheme may reduce the power and/or latency by incorporating these cost factors into objective function as a regularization term [23], [24]: $\min \sum_i L\left(y_i, f\left(\boldsymbol{x}_i\right)\right) + \lambda \Psi\left(f, \boldsymbol{x_i}\right)$. Here, we seek to learn a decision function $f(x)$ that minimizes the loss $L\left(y_i, f\left(\boldsymbol{x}_i\right)\right)$ in conjunction with the computational cost $\Psi\left(f, \boldsymbol{x_i}\right)$. As an example, we study the impact of optimizing the model for power and latency. The power consumption for extracting various features was estimated for a standard digital implementation in 65nm CMOS process [22]. Taking seizure detection task as an example, the line-length feature ($\frac{1}{d}\sum_d |x[n] - x[n-1]|$, $d =$ window size) consumes a negligible power, whereas band power features are more power demanding due to the FIR filtering stage. For latency estimation, we used the length of the decision path, which is upper bounded by the maximum depth of the tree.

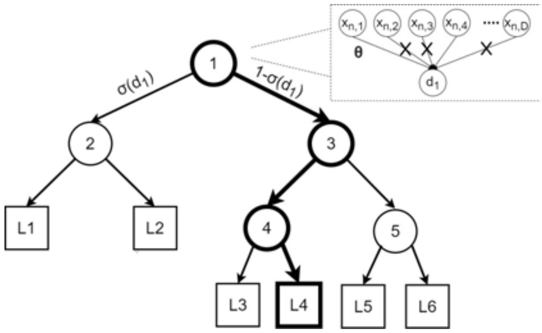

Fig. 4. Proposed oblique tree with probabilistic routing. In the inference phase, the test samples follow the most probable path. Inside internal nodes, the decision functions are represented by a two-layer neural network [22].

Figure 3(a) shows the trade-off between classification performance (F1 score) and cost metrics (latency, power). An ensemble of axis-aligned trees was used to detect epileptic seizures in 10 patients [22]. The number of trees and maximum depth were optimized using 5-fold cross-validation on each subject. As shown in Fig. 3(a), the classification performance degrades by shortening the decision path, thus requiring an asynchronous learning scheme to reduce the latency [5]. On the other hand, a power-efficient performance region is observed in Fig. 3(a), where we can drastically reduce the power consumption with only a marginal performance loss.

B. Model Compression

Compression techniques such as fixed-point quantization, weight pruning and sharing have been widely used to reduce the model size in DNNs [25], [26]. Similar techniques can be used to develop hardware-friendly DTs that are more efficiently deployable on ASIC. Importantly, conventional tree ensembles may suffer from a large model size due to the large number of trees required in non-trivial classification tasks [7], [10], [27]. To partially alleviate this issue, we quantized the tree parameters to reduce model size and allow fixed-point arithmetic [22], [24]. In a boosting framework, threshold values and leaf weights were quantized with 10 and 3 bits for seizure detection, reducing the model size by 2.4× compared to a gradient boosting ensemble with floating point weights.

Unlike axis-aligned decision trees, oblique trees involve multiple features in their internal nodes and can generate accurate predictions with a single tree (Fig. 4) [22]. Interestingly, within a probabilistic routing scheme, oblique trees can be trained using gradient-based optimization, similar to a neural network. Here, the decision functions in the internal nodes can be represented by a two-layer neural network, for which weight pruning and sharing techniques can be used to create sparse connections. Figure 3(b) illustrates the trade-off between classification performance and model size for an oblique tree (OT) with a maximum depth of 4, trained on the MNIST dataset. With weight pruning and sharing, the model size was compressed by 20× with only a marginal performance loss (0.3%). Moreover, with single-path inference only a small portion (26.3%) of the parameters were used, improving the hardware efficiency during inference.

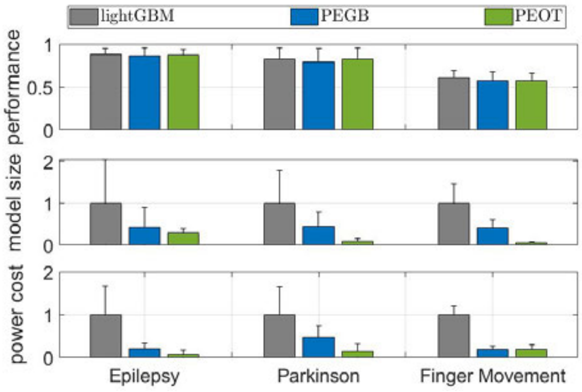

Fig. 5. Comparison of tree-based models on three neural tasks. Ensemble of gradient boosted trees (lightGBM [28]), power-efficient gradient boosted trees with fixed-point quantization (PEGB [23]), and power-efficient oblique trees with weight pruning and sharing (PEOT [22]) were compared. Average model sizes and power costs are normalized to the performance of lightGBM. Error bars represent the standard deviation among subjects.

Combining compression with cost-aware learning, we built power-efficient oblique trees (PEOT) and benchmarked them against conventional (lightGBM [5], [28]) and power-efficient gradient boosted trees (PEGB [23]). Testing on three different neural tasks including seizure detection (iEEG, 10 patients), tremor detection (LFP, 12 patients), and finger movement classification (ECoG, 9 subjects), PEOT reduced the model sizes by 10.5× and the power cost by 8.8×, Fig. 5. The PEOT model also achieved average reduction factors of 4.4× in model size and 2.3× in power cost compared to PEGB.

IV. Conclusion

We reviewed a recent trend in the development of closed-loop neural interfaces that embed ML on chip. Algorithm and hardware approaches for ML SoCs in various neural applications were discussed. We proposed a power-efficient oblique tree model which integrates cost-aware learning, weight pruning and sharing. Testing on three neural classification tasks, the proposed model improved the energy and memory efficiency while maintaining the classification performance.

References

[1] M. J. Morrell, "Responsive cortical stimulation for the treatment of medically intractable partial epilepsy," *Neurology*, vol. 77, no. 13, pp. 1295–1304, 2011.

[2] S. Little, A. Pogosyan *et al.*, "Adaptive deep brain stimulation in advanced parkinson disease," *Annals of neurology*, vol. 74, no. 3, pp. 449–457, 2013.

[3] K. H. Lee and N. Verma, "A low-power processor with configurable embedded machine-learning accelerators for high-order and adaptive analysis of medical-sensor signals," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 7, pp. 1625–1637, 2013.

[4] M. A. B. Altaf, J. Tillak *et al.*, "A 1.83 μJ/classification nonlinear support-vector-machine-based patient-specific seizure classification SoC," in *2013 ISSCC*. IEEE, 2013, pp. 100–101.

[5] M. Shoaran, B. A. Haghi *et al.*, "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.

[6] M. Shoaran, M. Farivar, and A. Emami, "Hardware-friendly seizure detection with a boosted ensemble of shallow decision trees," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2016, pp. 1826–1829.

[7] G. O'Leary, J. Xu *et al.*, "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," in *2020 ISSCC*. IEEE, 2020, pp. 402–404.

[8] W.-M. Chen, H. Chiueh *et al.*, "A fully integrated 8-channel closed-loop neural-prosthetic CMOS SoC for real-time epileptic seizure control," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 1, pp. 232–247, 2014.

[9] L. Yao, P. Brown, and M. Shoaran, "Improved detection of Parkinsonian resting tremor with feature engineering and Kalman filtering," *Clinical Neurophysiology*, vol. 131, no. 1, pp. 274–284, 2020.

[10] B. Zhu, G. Coppola, and M. Shoaran, "Migraine classification using somatosensory evoked potentials," *Cephalalgia*, vol. 39, no. 9, pp. 1143–1155, 2019.

[11] W.-C. Fang, K.-Y. Wang *et al.*, "Development and validation of an EEG-based real-time emotion recognition system using edge AI computing platform with convolutional neural network system-on-chip design," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 4, pp. 645–57, 2019.

[12] A. R. Aslam, T. Iqbal *et al.*, "A 10.13μJ/classification 2-channel deep neural network-based SoC for emotion detection of autistic children," in *2020 IEEE Custom Integrated Circuits Conference (CICC)*. IEEE, 2020, pp. 1–4.

[13] S.-Y. Chang, B.-C. Wu *et al.*, "An ultra-low-power dual-mode automatic sleep staging processor using neural-network-based decision tree," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 9, pp. 3504–3516, 2019.

[14] Y. Chen, E. Yao, and A. Basu, "A 128-channel extreme learning machine-based neural decoder for brain machine interfaces," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 3, pp. 679–692, 2016.

[15] Y. Yang, S. Boling, and A. J. Mason, "A hardware-efficient scalable spike sorting neural signal processor module for implantable high-channel-count brain machine interfaces," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 11, no. 4, pp. 743–754, 2017.

[16] A. T. Do, S. M. A. Zeinolabedin *et al.*, "An area-efficient 128-channel spike sorting processor for real-time neural recording with 0.175 μW/channel in 65-nm CMOS," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 1, pp. 126–137, 2019.

[17] L. Yao, P. Brown, and M. Shoaran, "Resting tremor detection in Parkinson's disease with machine learning and Kalman filtering," in *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2018, pp. 1–4.

[18] Y. Ezzyat, P. A. Wanda *et al.*, "Closed-loop stimulation of temporal cortex rescues functional networks and improves memory," *Nature Communications*, vol. 9, no. 1, pp. 1–8, 2018.

[19] L. Yao, J. L. Baker *et al.*, "Mental fatigue prediction from multi-channel ecog signal," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 1259–1263.

[20] A. Page, C. Sagedy *et al.*, "A flexible multichannel eeg feature extractor and classifier for seizure detection," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 62, no. 2, pp. 109–113, 2015.

[21] M. Taghavi and M. Shoaran, "Hardware complexity analysis of deep neural networks and decision tree ensembles for real-time neural data classification," in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE, 2019, pp. 407–410.

[22] B. Zhu, M. Farivar, and M. Shoaran, "Resot: Resource-efficient oblique trees for neural signal classification," *IEEE Transactions on Biomedical Circuits and Systems*, 2020.

[23] B. Zhu, M. Taghavi, and M. Shoaran, "Cost-efficient classification for neurological disease detection," in *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE, 2019, pp. 1–4.

[24] B. Zhu and M. Shoaran, "Hardware-efficient seizure detection," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2019, pp. 2040–2043.

[25] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," in *International conference on learning representation*, 2016.

[26] D. Lin, S. Talathi, and S. Annapureddy, "Fixed point quantization of deep convolutional networks," in *International Conference on Machine Learning*, 2016, pp. 2849–2858.

[27] L. Yao and M. Shoaran, "Enhanced classification of individual finger movements with ECoG," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2019, pp. 2063–2066.

[28] G. Ke, Q. Meng *et al.*, “Lightgbm: A highly efficient gradient boosting decision tree,” in *Advances in neural information processing systems*, 2017, pp. 3146–54.