Compressed Sensing Based Seizure Detection for an Ultra Low Power Multi-core Architecture

Roghayeh Aghazadeh[‡], Fabio Montagna[†], Simone Benatti[†], Javad Frounchi[‡], Davide Rossi[†]
[‡]University of Tabriz, [†]University of Bologna
{r_aghazadeh,jfrounchi}@tabrizu.ac.ir, {fabio.montagna, simone.benatti, davide.rossi}@unibo.it

Abstract—**Extracting information from brain signals in advanced Brain Machine Interfaces (BMI) often requires computationally demanding processing. The complexity of the algorithms traditionally employed to process multi-channel neural data, such as Principal Component Analysis (PCA), dramatically increases while scaling-up the number of channels and requires more power-hungry computational platforms. This could hinder the development of low-cost and low-power interfaces which can be used in wearable or implantable real-time systems. This work proposes a new algorithm for the detection of epileptic seizure based on compressively sensed EEG information, and its optimization on a low-power multi-core SoC for near-sensor data analytics: Mr. Wolf. With respect to traditional algorithms based on PCA, the proposed approach reduces the computational complexity by 4.4x in ARM Cortex M4-based MCU. Implementing this algorithm on Mr.Wolf platform allows to detect a seizure with 1 ms of latency after acquiring the EEG data for 1 s, within an energy budget of 18.4 μJ. A comparison with the same algorithm on a commercial MCU shows an improvement of 6.9x in performance and up to 18.4x in terms of energy efficiency.**

Keywords—**EEG, BMI, compressing sensing, SVM, embedded systems, ultra-low power, multi-core**

I. Introduction

Epilepsy is one of the most frequent brain diseases, affecting at least 50 million subjects, one third of whom suffer unpredictable recurrent seizures despite treatment. This results in major medical, social and economic burdens. Nowadays, most efficient approaches in epileptic seizure detection rely on video-EEG monitoring. The EEG is recorded for a prolonged period, supported by continuous video recording. This procedure requires a protected environment as well as expensive medical equipment and significant intervention of clinicians. Furthermore, it negatively impacts on the patients quality of life and the obtrusiveness of the recording setup hinders continuous monitoring of subjects status.

Detecting seizures in an automated fashion represents a paramount challenge to improve both the life quality of patients and the level of the treatment. The analysis of the brain signals, leveraging machine learning algorithms, has been widely investigated over the last years, helped also by the availability of wide public EEG datasets [1], [2]. Several algorithms were investigated and tested [1], [3], delivering high performance in terms of accuracy, sensitivity and specificity. Unfortunately, most of these techniques are computationally intensive and require bench top systems, like personal computers or other high-end computational platforms.

Recent advancements in embedded low-power electronics [4] offer major new opportunities to tackle this issue. In fact, the design of energy-efficient multi-core digital platforms allows to run machine learning algorithms to enable the real-time seizure detection maintaining high energy efficiency and a wearable form factor [5]. Several works show implementation of seizure detection algorithms based on dimensionality reduction, conventional feature extraction techniques and statistical learning classifiers [5]–[7]. However, the computational complexity of these algorithms significantly increases with the number of channels, especially for what concerns the dimensionality reduction stage [5].

To reduce the complexity of the preprocessing, a promising approach is to leverage the Compressive Sensing (CS) techniques. For instance, in [8], authors implemented a seizure detection algorithm on a low-power domain-specific many-core platform. As opposed to multi-channel approaches based on traditional dimensionality reduction algorithms [5], in [8] compressed sensing reduced the computational complexity, but with poor classification accuracy. On the other hand, in this work we demonstrate that coupling preprocessing based on compressed sensing with powerful feature extraction and classification techniques allows to achieve classification performance results similar to traditional approaches like PCA coupled with lightweight execution typical of CS approaches.

In this work, we present a novel framework for seizure detection, using CS to reduce the complexity of the processing chain, Least-Squared Spectral Analysis (LSSA) as feature extraction and Support Vector Machine (SVM) to classify the features. Therefore, the multi-modal approach that combines parallel processing and near threshold computing on a multi-core Parallel Ultra-Low Power (PULP) platform [9] allows us to detect a seizure with an energy consumption of 18.45 μJ. The proposed algorithm obtains an improvement of 4.4x in terms of energy with respect to traditional algorithms when implements on the ARM Cortex M4-based MCU. Moreover, we compared the results of implementation on the PULP with a commercial ARM Cortex M4-based MCU gaining an improvement in energy efficiency of 18.4x. It is noteworthy that the proposed framework is based on a fully programmable multi-core architecture, showing a highly versatility and it can scale between number of channels and cores and can be tailored to a wide range of applications and architectures.

II. Material and Methods

A. Compressed Sensing-based Seizure Detection

During the last years, CS is gaining ground as a promising framework to tackle energy consumption and computational

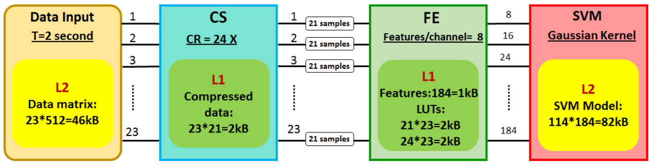

Fig. 1. Seizure detection algorithm processing chain and memory requirement for each block.

complexity issues in high dimensional and long-term monitoring of physiological signals. Nevertheless, CS-based acquisition techniques suffer from the reconstruction process, since it relies on the solution of an optimization problem that can reach $O(N^3)$ complexity. Extracting the desired information directly from compressively sensed signals avoids this problem and reduces the required energy budget of the application. In CS, signals are sub-sampled at a smaller sampling rate w.r.t. the Nyquist theorem. For this reason, a signal $\mathbf{X}_{N\times 1}$ is multiplied by a projection matrix $\mathbf{\Phi}_{M\times N}$, as shown in

$$\mathbf{Y} = \mathbf{\Phi} X \tag{1}$$

This results in a sub-sampled signal $\mathbf{Y}_{M\times 1}$, where $M \ll N$. To ensure a successful compression the signals should be sparse in some bases. Moreover, the projection matrix and these sparse basis have to be incoherent [10]. The amount of data reduction is quantified through the compression ratio (CR), which can be defined as

$$CR = \frac{N}{M} \tag{2}$$

where N indicates the dimension of the Nyquist sampled signal and M is the dimension of the compressed signals.

The feasibility of applying CS principles for EEG signals has been demonstrated in [11]. The EEG signals are sparse in several basis like Gabor, Discrete Wavelet and Discrete Cosine Transform (DCT). Random sensing matrices with a sufficient number of samples exhibit a low coherence with these basis. Pamula et al. [12] show that using a projection matrix containing only one nonzero entry per each row at random positions is a valid low complexity approach, that still provides good integrity of the compressed sparse signal. This matrix can be created by randomly choosing a subset of the rows of an identity matrix. The same projection matrix can be reused for compressing every epoch of the signal.

As shown in Figure 1 each channel of the EEG data is compressed over a window of 2 seconds with 50% overlapping between windows (1 second latency) with a Compression Ratio (CR) of 24x. In this paper, the Least-squares Spectral Analysis (i.e. Lomb-Scargle periodogram) is used to extract the features of compressively sampled signals. Lomb-Scargle periodogram is a well-known procedure to generate a power spectrum and to detect the periodic component in random and unevenly sampled dataset. Given an N sample signal $\mathrm{X_k}$ sampled at $\mathrm{t_k}$ times, where k goes from 1 to N. The Lomb-Scargle periodogram (LSP) is defined in [13] and its final formula is shown in Eq. (3)

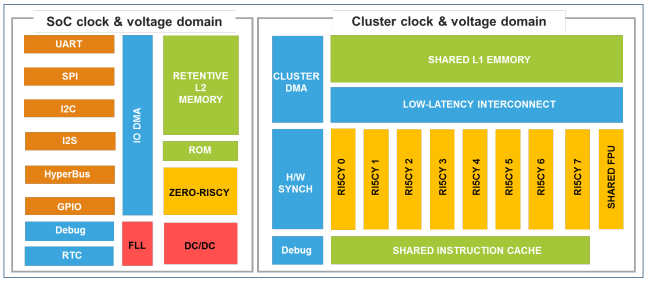

Fig. 2. Mr. Wolf SoC block diagram.

$$\begin{aligned} P_{LS}(f) = \frac{1}{2\delta^2} & \frac{[\sum_{k=1}^{N}(x_k - \widetilde{x})\cos(2\pi f(t_k-\tau))]^2}{\sum_{k=1}^{N} cos^2(2\pi f(t_k-\tau))} \\ & + \frac{[\sum_{k=1}^{N}(x_k - \widetilde{x})\sin(2\pi f(t_k-\tau))]^2}{\sum_{k=1}^{N} sin^2(2\pi f(t_k-\tau))} \end{aligned} \tag{3}$$

where $\widetilde{x}$ and δ^2 are respectively the mean and the variance of the data and the τ is calculated from Eq. (4):

$$\tau = \frac{1}{4\pi f}\arctan\left(\frac{\sum_{k=1}^{N}\sin(2(2\pi f)t_k)}{\sum_{k=1}^{N}\cos(2(2\pi f)t_k)}\right) \tag{4}$$

The amount of LSP had been calculated for each epoch of compressed data in 8 frequency bands between 0.5 - 25 Hz.

The feature vectors obtained by the LSP are classified as seizure/non-seizure using the SVM, a supervised classifier which aims to find the optimal separation hyperplane between 2 classes. The separation hyperplanes represent the boundary between the 2 classes and it is defined by Support Vectors (SVs), weights (α_i), bias (b) parameters. Such values are used to classify a feature vector X according to the following equation:

$$C = \sum_i \alpha_i K(S_i, X) + b \tag{5}$$

where k denotes the kernel function. The model for the classification was computed off-line on MATLAB and the parameters of the training function were tuned through a k-fold cross-validation, leading to the choice of Gaussian kernel, c parameter between 10-100 and sigma between 10-20 (depending on the subject).

B. *Mr.Wolf SoC*

In this work we target the implementation of the aforementioned algorithm on a Parallel Ultra-Low-Power (PULP) SoC [9] implemented in CMOS 40nm technology: Mr.Wolf. The SoC, shown in Figure 2, is a multi-core programmable processor coupling an advanced MCU controlled based on a tiny (12 Kgates) RISC-V processor (zero-risky) [14] accelerated by a powerful 8-processors compute cluster leveraging the flexible and powerful DSP extensions available on the RI5CY processor [14]. The SoC contains a full set of peripherals, including a Quad SPI (QSPI) interface suitable to connect to a multi-channel EEG acquisition system. Data transfers from peripherals are autonomously managed by a multi-channel I/O DMA to minimize the amount of interactions

	Accuracy(%)		Sensitivity(%)		Specificity(%)	
Subject	This work	Work in [5]	This work	Work in [5]	This work	Work in [5]
1	91.4	92.9	96.0	98.0	91.4	92.9
2	96.0	94.2	93.5	95.5	95.9	94.2
3	94.5	90.5	94.5	92.6	94.3	90.4
4	96.5	97.4	96.1	100.0	96.5	97.4
5	98.4	99.0	100.0	94.9	98.4	99.0
Average	95.4	94.8	96.8	95.4	95.3	94.8

TABLE I
ACCURACY, SENSITIVITY AND SPECIFICITY USING DATA FROM 5 RANDOMLY CHOSEN SUBJECTS OF THE CHB-MIT DATASET.

with the controlling core when performing IO. 512 kB of L2 memory are available on the SoC, accessible from the parallel cluster through a dedicated DMA controller. The cluster is equipped with a single cycle latency, multi-banked L1 memory, enabling shared-memory parallel programming models such as OpenMP. Two floating-point units (FPU) are shared among the 8 processors of the cluster, necessary to accelerate computations when highly dynamic data is present such as in EEG processing algorithms. Fast event management, parallel thread dispatching, and synchronization are supported by a dedicated hardware block (HW Sync), enabling very fine-grained parallelism and hence high energy efficiency in parallel workloads. To maximize power efficiency, the SoC contains an internal DC/DC converter that can be directly connected to an external battery that can deliver voltages in the range of 0.8 V to 1.1 V. When the system is in sleep mode this regulator is turned off and a lowdropout (LDO) regulator powers the real-time clock fed by a 32 kHz crystal oscillator, which controls programmed wake-up and, optionally, part of the L2 memory allowing retention of application state for fast wake-up. When in deep sleep the current consumption is reduced to 72 μW (from VBAT) assuming the RTC is active and no data retention, and 108 μW assuming full L2 retention.

III. Experimental Results

We evaluate our algorithm on 5 of the 23 subjects included in the CHB-MIT [1] dataset. Data are obtained with a 23 electrodes setup on the scalps of epileptic pediatric subjects following the 10-20 System [15]. The EEG signals are sampled at 256 Hz, through a 16-bit ADC. The subjects used for the test were randomly chosen among the dataset. The processing chain was first tuned and executed on MATLAB, verifying the accuracy, sensitivity and specificity.

For the training of the classifier, the 25% of seizure epochs and the same length of non-seizure epochs are used. The remaining epochs are used for testing and validating the model. As shown in Table I, the algorithm reaches on average 95.4% accuracy and 96.8% and 95.3% respectively for sensitivity and specificity, computed as

$$\begin{aligned} Sensitivity &= \frac{TP}{FN+TP} * 100 \\ Specificity &= \frac{TN}{FP+TN} * 100 \end{aligned} \tag{6}$$

where FN, FP, TP and TN stand for False Negative, False Positive, True Positive and True Negative. After the validation, the seizure detection algorithm was implemented and tested on Wolf and on a ARM Cortex M4-based commercial MCU. A block diagram of the seizure detection algorithm is shown in Figure 1, providing details of the whole processing chain and in Table II the execution time (clock cycles) and energy analysis of the seizure detector on the target platforms are summarized.

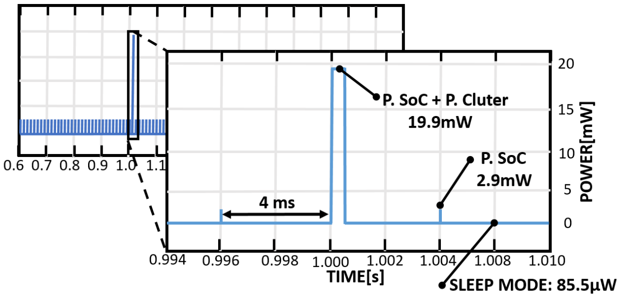

Fig. 3. Mr. Wolf (8-cores) Power consumptions: idle (deep sleep), only the acquisition (SoC), acquisition and elaboration (SoC+Cluster).

The CS kernel requires a small amount of clock cycles to be executed. In fact it represents only 1% of the overall execution time. The input data with the size of 46 kB is stored in L2 memory. The compressive sensing block randomly takes 21 samples out of the 512 input data for each channel (21x23=2 kB). From the Table II, it is noticeable that, when executing the CS kernel over multiple cores, even if the workload is perfectly balanced among the cores, the speed-up tends to saturate. This is mainly due to the fact that random accesses to L2 memory (15 clock cycles for each access) are required. Thus, due to the randomness of these accesses, the double buffering would have no impact (or even negative) on the performance.

After CS, the Feature Extraction (FE) kernel is performed. This kernel constitutes around the 20% of the total load. The frequency power of compressed data are calculated using LSP function implemented for the frequency bands between 0.5-25 Hz. Moreover, since the frequency and the time stamps are constant in this application, the sine and cosine functions of LSP are implemented with Look Up Tables (LUT). The dimension of these LUTs are 24x23 and 21x23 (4 kB). The FE kernel demonstrates a strong predisposition to scale among all the cores, allowing to exploit parallelism. In fact, passing from the execution with 1-core to the 8-cores Wolf a 7.3x improvement is obtained.

The most time consuming kernel of the processing chain is the SVM on both ARM Cortex M4 and Mr.Wolf (71.6% and 79.4% respectively). The model of the SVM derived from the off-line training on MATLAB is stored in L2 memory (82 kB) and it is transfered from L2 to L1 memory via DMA with double buffering. SVM reaches 6.6x speedup with 8-core Mr.Wolf platform. The non-ideal speed-up of the SVM is caused by workload unbalance and by the final classification stage which is executed sequentially due to the lightweight workload (i.e. few hundreds of cycles), not justifying the overhead of an additional OpenMP parallel region.

Fig. 3 shows the power contributions during different operative states of the system. The total memory footprint for the implementation of the algorithm on Mr.Wolf platform is

	ARM Cortex M4 - work in [5]			ARM Cortex M4 - this work			1-core Mr.Wolf			8-cores Mr.Wolf		
Kernel	kCyc	ld%	E[μJ]^b	kCyc	ld%	E[μJ]^b	kCyc	ld%	E[μJ]^c	kCyc	sp^a	E[μJ]^c
PP	2600.00	82.00	1234.00	7.40	1.00	3.51	3.60	0.50	0.26	1.00	3.60	0.18
FE	192.00	6.00	91.10	195.60	27.30	92.83	140.00	20.00	10.29	19.20	7.29	3.40
C	369.00	12.00	175.10	512.00	71.60	243.00	553.00	79.40	40.66	83.90	6.59	14.87
TOT	3165.00	100.00	1502.21	715.00	100.00	339.34	696.60	100.00	51.17	104.10	6.69	18.45

^{a}Speed-Up with respect to single-core Mr.Wolf paltform ^{b} 168MHz@1.8V ^{c} 110MHz@0.8V

TABLE II

PROPOSED IMPLEMENTATION ON BOTH ARM CORTEX M4 [5] AND MR.WOLF. PP, FE, C AND TOT STAND FOR PRE-PROCESSING, FEATURE EXTRACTION, CLASSIFICATION AND TOTAL, WHILE KCYC, LD, SP, E STAND FOR THOUSANDS OF CYCLES, LOAD, SPEED-UP AND ENERGY.

equal to 135 kB, hence we need 3 blocks of L2 memory to store all the data and the code. The amount of power used in sleep mode is 72 μW for the RTC and 3x4.5=13.5 μW for memory retention, which leads to a total power of 85.5 μW. The sampling frequency is equal to 256Hz, thus the system requires 2.9 mW (SoC Power) to acquire a new sample each 4ms. Furthermore, after the first second of execution, the cluster elaborates the acquired samples in 1.04 ms. At this stage, the total power is around 19.9 mW and it derives from the sum of both Soc and Cluster Power.

Table II shows the performance and the energy consumption measurements on the ARM Cortex M4-based (we targeted our implementation on a STM32F4-DISCOVERY board) and Mr.Wolf. In terms of execution time, we obtain a 4.4x improvement comparing the previous work with PCA and the one presented in this paper. The ARM Cortex M4 operates at 168MHz at 1.85 V, thus in both the implementations the entire execution is performed with a latency less than 20ms. The most important aspect to consider is represented by 4.4x gain in the energy consumption obtained changing the pre-processing kernel (i.e. from PCA to CS). Moreover, passing from ARM Cortex M4 to 1-core Mr.Wolf give us a great improvement in term of energy consumption (6.6x at 110 MHz@0.8 V), even if the number of clock cycles are similar. This gap increases exploiting the parallel programming, leading to a 2.8x passing from 1 to 8-cores Mr.Wolf and 18.4x with respect to ARM Cortex M4.

Comparing our implementation with a similar work described in [8], which uses compressing sensing and linear features with different classifiers in multi-core platform (PENC), our algorithm shows better results in term of energy consumption and accuracy. In fact, they achieve a sensitivity of 81.8% and specificity of 93.9% for Nyquist-domain seizure data which degrade 2.07% and 2.97% for a CR=16x respectively. In addition, the energy consumption of their design when they use LR and SVM as classifiers are about 1.175 mJ and 0.0018 mJ with CR of 16x, respectively.

IV. CONCLUSION AND FUTURE WORK

This paper presents a real-time, scalable and flexible seizure detection algorithm implemented on a Parallel Ultra-Low-Power (PULP) platform. We take the advantage of extracting the feature vector directly from compressively-sensed data to reduce the computational complexity. Extracting proper features in compressed domain and using an SVM classifier lead to an average sensitivity of 96.8% and accuracy of 95.4%. Using CS instead of the PCA dimensionality reduction approach, leads to an improvement of 4.4x in performance and energy consumption. Furthermore, implementing the algorithm on multi-core platform speed-ups the seizure detection algorithm by 6.7x with respect to sequential implementation. Moreover, the amount of energy consumption degrades 2.8x using the 8-cores Mr.Wolf. As the SVM kernel is the most time consuming part of the processing chain (around the 80% of the overall load), implementing a configurable accelerator for this kernel can further reduces the system power consumption and it can be done as a future work.

V. ACKNOWLEDGMENTS

This work was supported by the European project EuroCPS (grant n. 644090).

REFERENCES

[1] A. Shoeb *et al.*, "Application of machine learning to epileptic seizure detection," in *ICML-10*, 2010, pp. 975–982.

[2] R. Andrzejak *et al.*, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state," *Physical Review E*, 2001.

[3] N. Ahammad *et al.*, "Detection of epileptic seizure event and onset using eeg," *BioMed research international*, vol. 2014, 2014.

[4] S. Benatti *et al.*, "Multiple biopotentials acquisition system for wearable applications." in *BIODEVICES*, 2015, pp. 260–268.

[5] Benatti *et al.*, "Scalable eeg seizure detection on an ultra low power multi-core architecture," in *Biomedical Circuits and Systems Conference (BioCAS), 2016 IEEE*. IEEE, 2016, pp. 86–89.

[6] K. H. Lee *et al.*, "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*, 2013.

[7] F. Montagna *et al.*, "Flexible, scalable and energy efficient bio-signals processing on the pulp platform: A case study on seizure detection," *Journal of Low Power Electronics and Applications*, 2017.

[8] A. Kulkarni *et al.*, "Sketching-based high-performance biomedical big data processing accelerator," in *Circuits and Systems (ISCAS), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 1138–1141.

[9] D. Rossi *et al.*, "Energy-efficient near-threshold parallel computing: The pulpv2 cluster," *IEEE Micro*, vol. 37, no. 5, pp. 20–31, September 2017.

[10] E. J. Candes *et al.*, "Near-optimal signal recovery from random projections: Universal encoding strategies?" *IEEE transactions on information theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[11] M. Shoaib *et al.*, "A 0.6–107 μw energy-scalable processor for directly analyzing compressively-sensed eeg," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 61, no. 4, pp. 1105–1118, 2014.

[12] V. R. Pamula *et al.*, "A 172μw compressively sampled photoplethysmographic (ppg) readout asic with heart rate estimation directly from compressively sampled data," *IEEE transactions on biomedical circuits and systems*, vol. 11, no. 3, pp. 487–496, 2017.

[13] N. R. Lomb, "Least-squares frequency analysis of unequally spaced data," *Astrophysics and space science*, vol. 39, no. 2, pp. 447–462, 1976.

[14] P. D. Schiavone *et al.*, "Slow and steady wins the race? a comparison of ultra-low-power risc-v cores for internet-of-things applications," in *PATMOS*, Sept 2017, pp. 1–8.

[15] G. Klem *et al.*, "The ten-twenty electrode system of the international federation," *Electroencephalogr Clin Neurophysiol*, 1999.