A 41.2 nJ/class, 32-Channel On-Chip Classifier for Epileptic Seizure Detection

Milad Taghavi, Benyamin A. Haghi, Masoud Farivar, Mahsa Shoaran, Azita Emami

***Abstract*—A 41.2 nJ/class, 32-channel, patient-specific on-chip classification architecture for epileptic seizure detection is presented. The proposed system-on-chip (SoC) breaks the strict energy-area-delay trade-off by employing area and memory-efficient techniques. An ensemble of eight gradient-boosted decision trees, each with a fully programmable Feature Extraction Engine (FEE) and FIR filters are continuously processing the input channels. In a closed-loop architecture, the FEE reuses a single filter structure to execute the top-down flow of the decision tree. FIR filter coefficients are multiplexed from a shared memory. The 540 × 1850 μm^2 prototype with a 1kB register-type memory is fabricated in a TSMC 65nm CMOS process. The proposed on-chip classifier is verified on 2253 hours of intracranial EEG (iEEG) data from 20 patients including 361 seizures, and achieves specificity of 88.1% and sensitivity of 83.7%. Compared to the state-of-the-art, the proposed classifier achieves 27× improvement in Energy-Area-Latency product.**

I. Introduction

Recently, classification techniques have enabled data-driven solutions for closed-loop therapeutic and prosthetic devices [1-4]. These medical devices have benefited a broad range of neurological disorders such as epilepsy [1-3], sleep staging [4,5]. Prior findings confirm the presence of statistical and mathematical biomarkers in Electroencephalography (EEG) of such patients [6-8]. In epileptic seizures, an abrupt change in these biomarkers advance the clinical onset of the seizure. The interval varies from 0.5s to 10s [9]. Therefore, a fast and efficient classifier, which can predict a seizure incident in advance would enable drug delivery or can alarm patients or caregivers to take proper actions [10]. To assess our proposed on-chip machine learning classifier, epilepsy is chosen as the target disease due to the availability of continuous recordings from many patients. This architecture, however, can potentially benefit many other on-chip sensing applications.

Researchers have employed a variety of classification algorithms and analog/digital circuit design techniques to implement low-power and area-efficient SoCs for seizure detection [1-2]. Time-division multiplexing band-pass filter architecture together with the log-linear Gaussian basis function classifier presented in [1] achieves one of the best energy efficiencies so far (1.31 μJ/class) with a latency of 2s and occupies 7mm^2 working with 8 channels. Entropy-and-spectrum-aided method in [2] achieved a low latency of 0.8s with 6.5mm^2 area and energy efficiency of 77.9 μJ/class for 8 channels. In particular achieving a latency of < 2s with low energy consumption and small area is challenging [1].

To improve the strict energy-area-delay trade-off and increase the number of channels, we employed a patient-specific prediction model in the form of an ensemble of 8 decision trees (Figure 1a), trained by the gradient-boosting machine learning algorithm. The implemented SoC can support up to 32 channels. One fully-programmable FEE unit is used per tree and controlled by Tree Control Unit (TCU) in a closed-loop system to extract biomarkers [11,12]. Figure 1b shows a Mealy FSM implementation of the closed-loop system. This technique substantially reduces the power and area overhead. To extract spectral density features (biomarkers), a single FIR filter structure is used and its coefficients are multiplexed according to the feature being processed. The programmable-FIR structure reduces FEE area. As a result, the proposed closed-loop hardware architecture for decision tree based prediction models achieves an energy efficiency of 41.2 nJ/class with a small area of 1mm^2.

II. Machine Learning Algorithm and Data Description

Gradient-boosting [13] is one the most successful machine learning techniques for classification. This algorithm produces a prediction model in the form of ensemble of weak learners. By exploiting a gradient-based optimization and boosting, the algorithm trains the prediction model to classify abnormalities in future feature vectors.

Prior works [6-9] have studied optimal features for seizure onset detection. Based on the reported results, we chose a set of two time domain and nine Fourier transform domain measures as our feature set[1]. Since ripple, fast-ripple and high-frequency oscillation (HFO) are features from relatively higher frequency bands, we require our patient data to be sampled at comparatively higher sampling rates. The iEEG collaborative database [14] supports recordings both at high and low sampling rates (500-5000 samples/s) for epilepsy studies.

For each patient, iEEG recordings are partitioned randomly ten times (80% training set, 20% validation set). Training sets are processed offline for feature extraction. For the purpose of feature extraction and training, time series divided into windows of 1s, and all features from all channels

Milad Taghavi, Benyamin A. Haghi, Azita Emami are with the Electrical Engineering Department, California Institute of Technology
Email: {mtaghavi, benyamin.a.haghi, azita}@caltech.edu
Masoud Farivar is with Google, Mountain View, CA, USA
Email: masoudf@google.com
Mahsa Shoaran is with the School of Electrical and Computer Engineering, Cornell University
Email: shoaran@cornell.edu

[1] delta: 1-4 Hz, theta: 4-8Hz, alpha: 8-13Hz, beta: 13-30Hz, gamma: 30-80Hz, ripple: 80-200Hz, fast-ripple: 200-250Hz, HFO: 80-250Hz or 80-600Hz

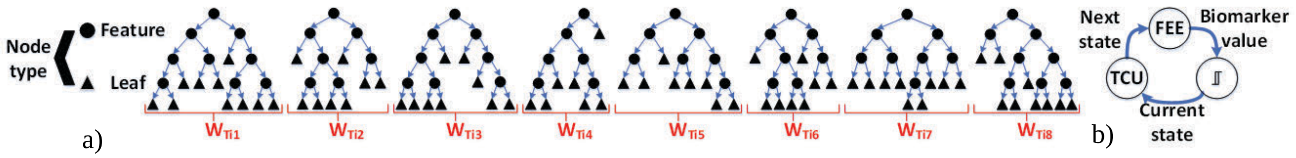

Figure 1: a) An ensemble of 8 trees trained for Patient ID# 3, b) Mealy FSM flowchart of closed-loop system

are extracted for each window. Then, the extracted features from training sets are fed to the algorithm for training the DT ensemble. The trained prediction model, which is the output from the gradient-boosting algorithm, includes full information on tree structures in the ensemble such as each leaf values, thresholds and selected features. In the proposed classifier, seizure vs non-seizure events are indicated whenever the Decision Function (DF) for the ensemble is positive (1).

$$\mathrm{DF} = \sum_{n=1}^{8} \mathrm{W}_{\mathrm{Ti}_n} \rightarrow \begin{cases} \mathrm{DF}>0 & \text{seizure} \\ \mathrm{DF}<0 & \text{non-seizure} \end{cases} \qquad (1)$$

In this prediction model, the longest possible update interval of DF cannot be longer than the longest path in the ensemble. This update interval determines the latency of the system. To minimize latency, in validation set tests, features are extracted from minimum possible sub-window[2] size of time series.

In the proposed architecture, DTs trained while working freely and in parallel regardless of the sub-window size of nodes. Finally, to avoid long latencies, results of completed and incomplete decision trees been averaged. Detailed discussion of the classification algorithm can be found on [11,12,15].

III. Proposed Architecture

Figure 2 shows the main blocks of the implemented Mealy FSM for the SoC: i. Ensemble of 8 DTs; ii. Memory Control Unit (MCU); and iii. Asynchronous Trees Reset Control (ATRC). Functional description of these blocks is explained in the following subsections.

A. DT ensemble

The ensemble includes 8 decision tree structures with maximum depth of 4 (15 nodes). For each tree, TCU will set the next state's memory pointer based on current state, comparator status and other internal flags. At each state transition, if the 'end_flag' is not active, the 'start' command will activate the FEE. When feature extraction and comparison is done, TCU resets all internal registers and wait for new node settings. At the last processing node of DT, TCU sends out the 'tree_end' flag and final node info to ATRC. In each node, according to node information of the current state, a multiplexer selects a channel among 32 channels of input data. This channel is then fed to FEE.

Processing of each selected feature is done in FEE module. A decoder activates/deactivates its sub-modules according to the current node's selected feature. Sub-blocks process input data within the time the 'start' command is available. After feature extraction is done, 'FEE_done' flag becomes available for comparator and TCU. The final value of FEE will be held until the next update.

B. Memory Control Unit

The MCU monitors read/write access to memory. In the write mode, a decoder activates different memory sub-modules for programming through serial input. It first sends out two reset pulses to the target sub-module. After these two pulses are cleared, the next data packet will be stored in the selected sub-module. For each DT, four sub-memory blocks with depth of 15 are storing the tree structure. These units include each node's feature/channel selection, threshold values and etc. The fully programmable memory enables patient-specific seizure detection. In read mode, MCU receives pointers and commands from each DT, and sends

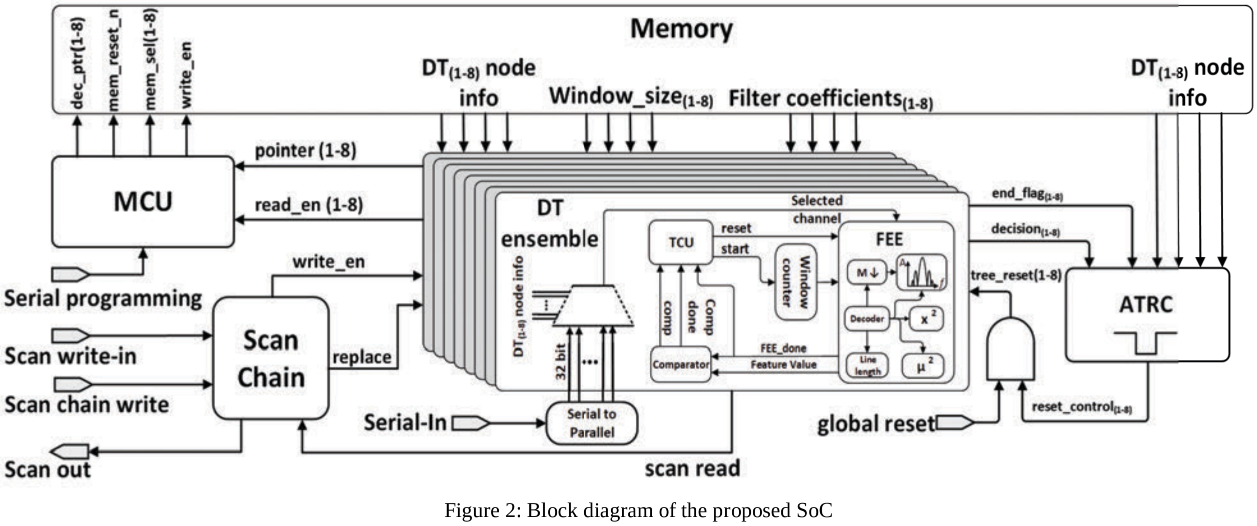

Figure 2: Block diagram of the proposed SoC

[2] delta: 2s, theta: 0.65, alpha: 0.5s, beta: 0.45s, gamma: 0.4s, ripple, fast-ripple, HFO, total power, line length, variance: 0.1s

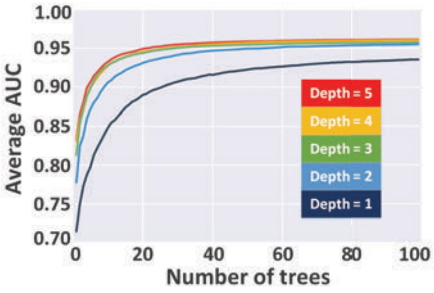

Figure 3: Average AUC at various depths versus number of tress in an ensemble

back the requested information. MCU also activates/deactivates the dedicated filter coefficient outputs from memory to DTs according to their node info. After the outputs were updated, MCU sends out “ready_flag” to TCU.

The total size of register type memory is 1kB. Shared filter coefficients take 262B. Each DT has a dedicated 690b memory for its node information.

C. Asynchronous Trees Reset Control

To capture all the descriptive abnormalities in the recordings, each tree works independently, i.e. when the ‘tree_end’ flag of a tree is available, ATRC stores tree status and resets the tree to its initial state. The reset is held for 2 clock cycles. After reset is cleared, the tree will start processing of input data. ATRC holds tree status until the next available ‘tree_end’ flag. Finally, ATRC assigns each tree’s respective leaf values to calculate DF according to (1).

IV. Hardware Description

Careful system-level analysis and hardware optimizations were employed to reduce the total power and area of the design. Following subsections cover the details of the implemented chip.

A. *Ensemble size*

Higher number of DTs in the prediction model would result in a larger area and higher power consumption. To find the minimum number of DTs in our prediction model while maintaining the accuracy, Area Under the Curve (AUC) performance of the ensemble versus the number of DTs for different values of the depth is simulated. As shown in Figure 3, the performance is not significantly improved (<5%) for depth values of 4 and higher. Also, an ensemble size of 8 would achieve an average performance of greater than 90%.

B. *Input bit precision*

The input bit precision have to be high enough to ensure the detectability of high-frequency features. On the other hand, lower bit resolution is preferred to reduce area and power. In order to find the optimum precision, the error in extracted HFOs from different channels of various patients was calculated with 9-12 bit-precisions. Figure 4 shows the error percentages with respect to floating point representation of input. A set of 400 time samples (each 1s) were randomly selected from iEEG database. As shown, 12-bit precision ensures less than 0.1% error in extracted HFO.

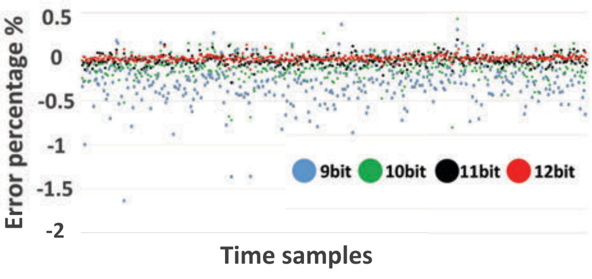

Figure 4: Error in extracted HFO for 9-12 bit.

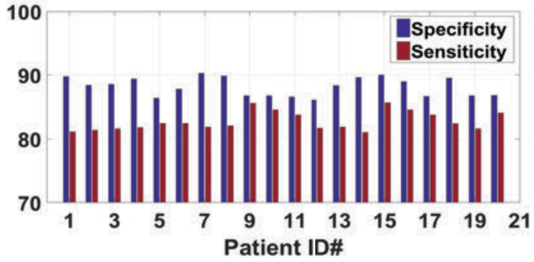

Figure 5: Sensitivity and specificity among patients

C. *Programmable FIR filters*

To calculate spectral density features, a cascade of two FIR filters were implemented. The first stage decimates input samples. The second stage provides band-pass filtering. Each stage may also be bypassed according to feature selection. Since at each node of the tree only one feature is being processed, a single filter structure with programmable coefficients can be used. This would relax the area-power constraint in feature extraction. FIR filters have Type-1 direct symmetric structures with 7 and 35 taps for first and second stage, respectively. A high number of taps would lead to extra power and area consumption in FEE and memory. To select optimal number of taps, extensive analysis made on accuracy of extracted HFO. Filter architectures and length were chosen to ensure lower than 5% error in feature extraction for HFO over all the training and validation datasets.

V. Measurement Results

For each patient, DT ensemble is programmed according to ensemble structure of his/her trained prediction model [see section II]. Then, the validation set partition of iEEG data for each patient is loaded on SoC for feature extraction and classification. Using the recorded classifier labels from DF, specificity and sensitivity are calculated according to (4) and (5) respectively:

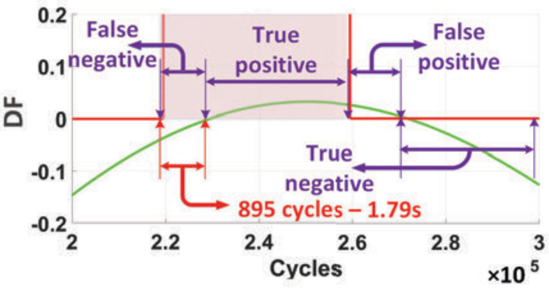

Figure 6: Variations of DF next to the worst-case latency case for patient with ID#7

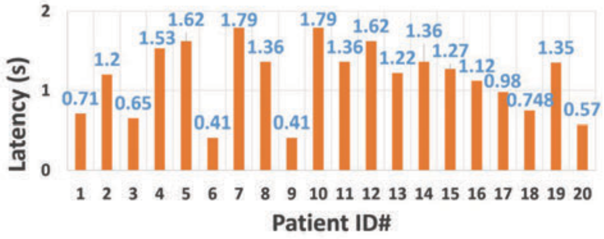

Figure 7: Worst-case latency for each patient

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \tag{4}$$

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{5}$$

The proposed classifier achieves an average specificity of 88.1% and an average sensitivity of 83.7%. The drop in sensitivity and specificity are due to limited frequency response of on-chip FIR filters in feature extraction. Figure 5 shows the specificity and sensitivity of each patient. Figure 6 shows variations of DF near ictal-marked data labels of Patient ID#7. As shown, the accumulated sum becomes positive 895 cycles after seizure incident. Figure 7 shows the worst case latency for each patient in this study.

The total static and dynamic power of classifier are 40.4 μW and 166 μW, respectively. The total dynamic power of memory (read/write) is 142 μW. Each tree, when active, consumes 3 μW of dynamic power. Power measurements were all made at worst-case scenarios where all the internal registers are switching and FEE is saturated (i.e. electrical onset of seizure incident).

The SoC operates at 0.8 V and at maximum frequency of 3.2MHz (serial input). Energy Efficiency of SoC at worst case power consumption (sampling rate of 5000 in this study) is 41.2nJ/class. Chip micrograph with area breakdown is provided in Figure 8. Each tree, together with its dedicated and shared memory allocations, takes 11.25% of the die area.

Table 1 summarizes the performance of the proposed design compared to the state-of-the-art seizure detection systems.

Table 1

Parameter	This Work	JSSC'17 [4]	TBCAS'16 [1]	JSSC'14 [2]
Signal	iEEG	EEG	EEG	iEEG
Application	Seizure	Sleep	Seizure	Seizure
Classifier	Gradient-boosted DT	DT	NLSVM	LLS
#Channels	32	1	8	8
Patient Specific	Yes	No	Yes	No
Sensitivity	83.70%	100%	95.1%	92%
Specificity	88.10%	100%	> 96.2%	N.A.
Latency	1.79s	N.A.	2s	0.8s
Process	65 nm	130 nm	180 nm	180 nm
Energy Efficiency	41.2 nJ/class	0.7 μJ/class	1.31 μJ/class	77.91 μJ/class
Area (mm^2)	1	6.386	7 *	6.5 *

*Area for the seizure detection block was not reported. The numbers have been conservatively estimated from total area breakdown

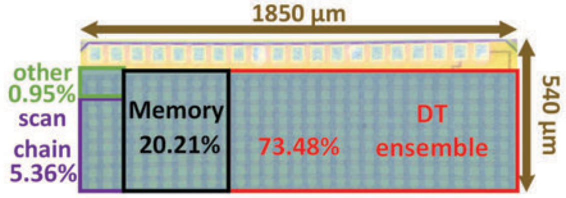

Figure 8: Chip micrograph including the area breakdown

VI. Conclusion

A low-power, hardware efficient, on-chip machine learning classifier for epileptic seizure detection is proposed. Hardware architecture, design optimization and trade-offs were discussed. The proposed classifier achieves energy efficiency of 41.2 nJ/class and can process up to 32 channels. The SoC is fabricated in TSMC 65nm mixed-signal low-power CMOS process with dimensions of 540 × 1850 um^2. The SoC breaks the strict energy-area-latency trade-off. For a fair comparison with the state of the art, power (energy) and area numbers of [1] normalized to 65 nm technology node. The proposed architecture achieves 27× improvement in energy-area-latency. This classifier can potentially enable full integration of diagnosis and termination of epileptic seizure in closed-loop therapeutic and prosthetic devices.

References

[1] M. Altaf, "A 1.83 μJ/classification Non-linear Support-Vector Machine-based Patient-specific Seizure Classification SoC," in ISSCC Dig. Tech. Papers. pp. 100-101, Feb. 2013.

[2] W. M. Chen, "A fully-integrated 8-Channel Closed-loop Neural-Prosthetic CMOS SoC for Real-Time Epileptic Seizure Control," IEEE JSSC, vol. 49, no. 1, pp. 232-247, Jan. 2014.

[3] A. Shoeb, "Application of Machine Learning to Epileptic Seizure Onset Detection and Treatment," PhD Thesis, Massachusetts Institute of Technology, Sept. 2009.

[4] S. A. Imtiaz, " An Ultra-low Power system on chip for automatic sleep staging," IEEE Journal of Solid-State Circuits, vol. 52, no. 3, pp 822-833 Mar. 2017

[5] S. Iranmanesh, "An Ultralow-power Sleep Spindle Detection System on Chip", IEEE Transactions on Biomedical Circuits and Systems, vol. 11, issue 4, Aug. 2017

[6] M. Shoaran, "A 16-Channel 1.1mm² Implantable Seizure Control SoC with Sub-μWChannel Consumption and Closed-Loop Stimulation in 0.18μm CMOS," *IEEE Symposium on VLSI Circuits*, HI, Jun. 2016.

[7] L. Logesparan, "Optimal Features for Online Seizure Detection," E. Med Biol Eng Comput (2012) 50:659.

[8] L. Boubchir, "A Review of Feature Extraction for EEG Epileptic Seizure Detection and Classification," TSP, July 2017.

[9] F. Bonini, "Frontal Lobe Seizure: From Clinical Semiology to Localization," Epilepsia, vol. 55, no. 2, pp. 264-277, Dec. 2013.

[10] M. Mirzaei, "A fully-asynchronous low-power implantable seizure detector for self-triggering treatment," IEEE Transactions on Biomedical Circuits and Systems, vol. 7, issue 5, Oct. 2013

[11] M. Shoaran, "Hardware-Friendly Seizure Detection with a Boosted Ensemble of Shallow Decision Trees," EMBC, Aug. 2016.

[12] M. Shoaran, "Energy-Efficient Classification for Resource-Constrained Biomedical Applications," IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS), submitted.

[13] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of Statistics, pp.1189-1232, 2001.

[14] www.ieeg.org

[15] M. Shoaran, "Efficient Feature Extraction and Classification Methods in Neural Interfaces" The bridge, National Academy of Engineering, Washington DC, vol. 47, no. 4, pp. 31-35, Winter 2017.