ARCHIVIO ISTITUZIONALE
DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Low-latency detection of epileptic seizures from IEEG with temporal convolutional networks on a low-power parallel MCU

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Zanghieri M., Burrello A., Benatti S., Schindler K., Benini L. (2021). Low-latency detection of epileptic seizures from IEEG with temporal convolutional networks on a low-power parallel MCU. Institute of Electrical and Electronics Engineers Inc. [10.1109/SAS51076.2021.9530181].

Availability:
This version is available at: https://hdl.handle.net/11585/851533 since: 2022-02-24

Published:

DOI: http://doi.org/10.1109/SAS51076.2021.9530181

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

06 October 2026

This is the final peer-reviewed accepted manuscript of:

Zanghieri, M., Burrello, A., Benatti, S., Schindler, & K., Benini, L., Low-latency detection of epileptic seizures from IEEG with temporal convolutional networks on a low-power parallel MCU, SAS 2021.
https://doi.org/10.1109/SAS51076.2021.9530181.

The final published version is available online at: https://doi.org/10.1109/SAS51076.2021.9530181

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

Low-Latency Detection of Epileptic Seizures from iEEG with Temporal Convolutional Networks on a Low-Power Parallel MCU

Marcello Zanghieri[1], Alessio Burrello[1], Simone Benatti[1], Kaspar Schindler[2], Luca Benini[1,3]

***Abstract*— Epilepsy is a severe neurological disorder that affects about 1% of the world population, and one-third of cases are drug-resistant. Apart from surgery, drug-resistant patients can benefit from closed-loop brain stimulation, eliminating or mitigating the epileptic symptoms. For the closed-loop to be accurate and safe, it is paramount to couple stimulation with a detection system able to recognize seizure onset with high sensitivity and specificity and short latency, while meeting the strict computation and energy constraints of always-on real-time monitoring platforms. We propose a novel setup for iEEG-based epilepsy detection, exploiting a Temporal Convolutional Network (TCN) optimized for deployability on low-power edge devices for real-time monitoring. We test our approach on the Short-Term SWEC-ETHZ iEEG Database, containing a total of 100 epileptic seizures from 16 patients (from 2 to 14 per patient) comparing it with the state-of-the-art (SoA) approach, represented by Hyperdimensional Computing (HD). Our TCN attains a detection delay which is 10 s better than SoA, without performance drop in sensitivity and specificity. Contrary to previous literature, we also enforce a time-consistent setup, where training seizures always precede testing seizures chronologically. When deployed on a commercial low-power parallel microcontroller unit (MCU), each inference with our model has a latency of only 5.68 ms and an energy cost of only 124.5 µJ if executed on 1 core, and latency 1.46 ms and an energy cost 51.2 µJ if parallelized on 8 cores. These latency and energy consumption, lower than the current SoA, demonstrates the suitability of our solution for real-time long-term embedded epilepsy monitoring.**

***Index Terms* — iEEG, seizure detection, long-term monitoring, low latency, seizure detection, deep learning, Temporal Convolutional Networks, real-time, embedded platforms, edge, Tiny Machine Learning.**

I. Introduction

Although the main treatment for epilepsy is pharmacological, approximately one-third of patients are affected by drug-resistant forms of epilepsy [1]. These cases can either require surgical treatment [2], or can benefit from closed-loop brain stimulation [3]. The latter can eliminate or mitigate the seizure symptoms, and relies on the coupling of a neuromodulator with a real-time detection system that recognizes the onset of seizures based on the analysis of the brain signals. Closed-loop neuromodulators are implantable devices that read intracranial Electro-Encephalographic (iEEG) signals and stimulate the brain tissue, and implantability imposes very strict computational resources and energy budget.

Currently, the iEEG signal allows the best spatial resolution and provides the highest signal-to-noise ratio compared to other neural recording techniques [4]. With this biosignal, many attempts have been done to develop frameworks able to detect seizures.

Recently, several works have proposed methods based on Machine Learning [5], [6], [7] and Deep Learning [8], [9] to successfully detect the *ictal* (i.e. during seizure) and the *inter-ictal* (i.e. between seizures) states from the iEEG signal.

High *sensitivity* and *specificity*, and short *delay* (i.e. time between the onset and the recognition of a seizure) are fundamental parameters for evaluating the quality of an epilepsy detection system. Above all, *specificity* is critical, because studies have shown that false positives can generate high levels of anxiety and stress in patients [3], hence they must be minimized.

For the automated learning approach, an invaluable source of brain activity data is the iEEG recorded in Epilepsy Monitoring Units (EMU), where it is possible to perform pre-surgical long-term observations. Typically, EMU patients are monitored for only 1 to 3 weeks, to minimize the discomfort and the risk of adverse effects (e.g., infection and inflammation deriving from the iEEG electrodes implanted through the skull) [10]. The collected data are not only essential for preliminary monitoring to plan personalized surgical treatment, but they are also used as a base for training algorithms for real-time seizure recognition [11].

Given the highly patient-specific nature of seizure dynamics, seizure detection frameworks require tuning to each patient [12]. This patient-dependent approach poses significant challenges because of the highly imbalanced nature of the data, where inter-ictal states are much longer than ictal states (class-imbalance problem).

A major challenge in real-time seizure detection is to design computationally efficient frameworks, able to provide a reliable recognition while at the same time meeting the strict computation, memory, and power constraints of embedded platforms working in real-time.

In this work, we address the problem of iEEG-based detection of epileptic seizures in real-time, targeting the Short-Term SWEC-ETHZ iEEG Database [11]. We propose a solution based on a Temporal Convolutional Network

[1] M. Zanghieri, A. Burrello, S. Benatti and L. Benini are with the Department of Electrical, Electronic and Information Engineering, University of Bologna, 40136 Bologna, Italy. name.surname@unibo.it

[2] K. Schindler is with the Sleep-Wake-Epilepsy-Center, Department of Neurology, Inselspital, Bern University Hospital, University Bern, 3010 Bern, Switzerland. kaspar.schindler@insel.ch

[3] L. Benini is also with the Department of Information Technology and Electrical Engineering at ETH Zurich, 8092 Zurich, Switzerland. lbenini@iis.ee.ethz.ch

* This work was supported in part by the European H2020 FET Project OPRECOMP under Grant 732631. We thank the CINECA computing centre for granting compute time availability on the Marconi100 supercomputer through the ISCRA-C project NAS4NPC.

(TCN) designed for low-power edge monitoring platforms. We present the following contributions:

- We present a novel TCN network, with 1D dilated convolutional layers, enabling a more efficient pattern extraction from input time windows; this yields a compact model requiring just $2.52\,\mathrm{kB}$ of memory footprint and $164\,\mathrm{kMAC}$ of computation, working entirely at `int8` bitwidth; we obtain a detection delay which is up to $10\,\mathrm{s}$ shorter than the state-of-the-art setup based on Hyper-Dimensional Computing; at the same time, we satisfy the same sensitivity and specificity constraints as the SoA; furthermore, our setup is time-consistent: training seizures always precede testing ones temporally, a constraint which would be present in the clinical practice, but that is unfortunately not taken into account in the SoA work [11].
- We deploy our model on the low-power edge microcontroller GAP8 [13], [14], attaining a computation latency of just $5.68\,\mathrm{ms}$ and an energy cost of just $124.5\,\mu\mathrm{J}$ when executed on 1 core, and latency $1.46\,\mathrm{ms}$ and an energy cost $51.2\,\mu\mathrm{J}$ when distributed on 8 cores. These values are better compared to the HDC SoA [11] and are a perfect fit for long-term monitoring by an embedded SoC working in real-time.

II. Materials & Methods

A. Short-Term SWEC-ETHZ iEEG Database

Intracranial Electroencephalography (iEEG) is an invasive technique to acquire brain signals via electrodes implanted surgically directly onto the surface (strip-, grid electrodes) or even into the brain (depth electrodes) [17]. Compared to extracranial EEG, the iEEG provides better spatial and temporal resolution (mm-scale and ms-scale, respectively [18]), higher bandwidth, less noise, and fewer artifacts, though with the drawback of requiring surgery with a higher risk of infection [19].

The dataset we address in this work is the Short-Term SWEC-ETHZ iEEG Database [11], a publicly available[1] iEEG dataset containing epileptic seizure recordings from 16 patients of the epilepsy surgery program of the Inselspital Bern, for a total of 100 seizures. The number of seizures varies from 2 to 14 across patients.

The iEEG signals were acquired by either implanted strip, grid, and depth electrodes, or by a mixed configuration of these electrode types. Electrode numbers (varying from 36 to 100 across subjects) and implantation schemes were established based on clinical needs. An extracranial electrode localized between the `Fz` and `Cz` positions (*10-20 system*) was used as reference. The sampling rate was either $512\,\mathrm{Hz}$ or $1024\,\mathrm{Hz}$, depending on whether each patient had more or less than 64 electrodes implanted. Prior to further analyses, the signals recorded with less than 64 electrodes were downsampled to $512\,\mathrm{Hz}$. All signals were re-referenced against the median of all electrodes free of permanent artifacts (e.g., $50\,\mathrm{Hz}$ PLI), as judged by visual inspection [20], [21]. The signals were digitalised to $16\,\mathrm{bit}$ and band-passed with a 4^{th}-order Butterworth filter with band $0.5 \div 150\,\mathrm{Hz}$.

For seizure onset marking, which constitutes the dataset's ground truth, the iEEG traces were visually inspected by an experienced board-certified epileptologist (K.S.) [11]. Electrodes permanently corrupted by artifacts were excluded by the same procedure. The dataset's ictal segments range from $10\,\mathrm{s}$ to $1002\,\mathrm{s}$. In addition, each recording includes $180\,\mathrm{s}$ of inter-ictal state preceding the seizure and $180\,\mathrm{s}$ of post-ictal state.

[1] http://ieeg-swez.ethz.ch

B. Temporal Convolutional Networks Framework

In this work, we address epileptic seizure detection treating the iEEG signal as a time series, applying a Temporal Convolutional Network (TCN) based on the state-of-the-art EEGNet [15].

TCNs are a recent class of deep neural networks that have surged to the SoA in numerous tasks of time series modeling, surpassing Recurrent Neural Networks (RNNs) for accuracy and ease of training [22], [23], [24]. TCNs have two distinctive properties, characterizing their 1D convolutions over the time dimension: (*i*) *causality*: filters only include the past neighborhood of each sample, thus getting no information from the future; (*ii*) *dilation*: a fixed distance d is inserted between the filter inputs, so as to expand the receptive field while keeping the model size constant. Hence, a causal dilated convolutional layer computes its output as follows:

$$\mathbf{y}^{c_{\text{out}}}(t) = \text{Conv}(\mathbf{x}) = \sum_{c_{\text{in}}=1}^{C_{\text{in}}} \sum_{k=0}^{K-1} \mathbf{W}_k^{c_{\text{in}},c_{\text{out}}} \mathbf{x}^{c_{\text{in}}}(t - d \cdot k) \quad (1)$$

with t time index, $\mathbf{x}$ and $\mathbf{y}$ input and output activations, $c_{\text{in}} = 1, \cdots, C_{\text{in}}$ and $c_{\text{out}} = 1, \cdots, C_{\text{out}}$ input and output channel respectively, K filter size $\mathbf{W} \in \mathbb{R}^{K \times C_{\text{in}} \times C_{\text{out}}}$ tensor of filters, and d dilation factor.

The TCN we use in this work is inspired by EEGNet [15], a CNN specialized for EEG. EEGNet has proven powerful on several tasks ranging from the classification of steady-state visual evoked potentials [16] to motion imagery recognition [25]. On top of these results, we exploit the EEGNet topology as a base to design a TCN that matches the memory and computation constraints of low-power edge microcontrollers. In particular, we preserve the block structure, whilst applying a reduction of parameter number.

Our EEGNet-inspired TCN is shown in Figure 1. We use 3 Convolutional Blocks each composed of 4 filters with batch-normalization [26]:

- Convolutional Block I has unit kernel $k = 1$ since it is in charge of the *spatial filtering* in the EEG sense: it mixes the input iEEG electrodes (spatially distributed) into new network channels, with a time-independent linear combination;
- Convolutional Blocks II and III extract the temporal information performing dilated causal convolutions with $k_{\text{II}} = k_{\text{III}} = 3$, and dilation $d_{\text{II}} = 2$ and $d_{\text{III}} = 4$.

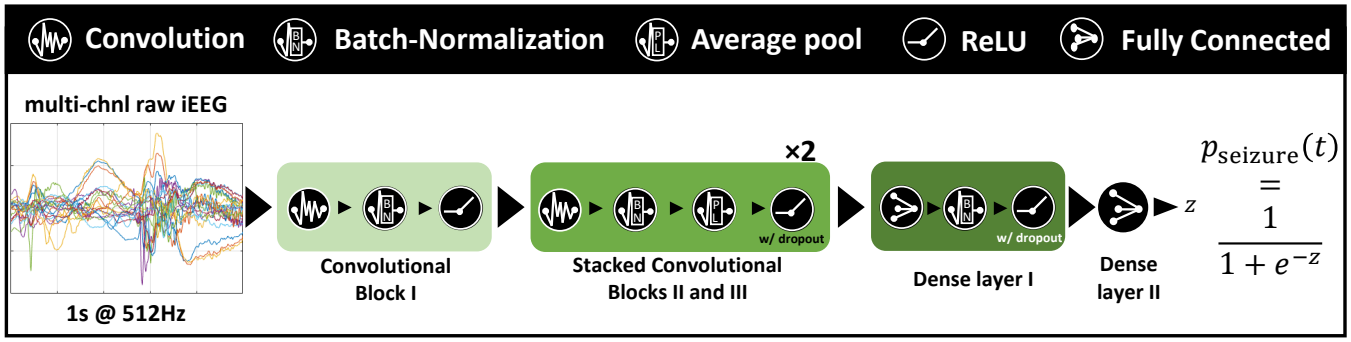

Fig. 1. The Temporal Convolutional Network topology used in this work, inspired by EEGNet [15], [16].

Finally, two stacked 32-unit dense layers compute the probability of the input window belonging to the ictal state, which is returned by a sigmoid activation. Overall, our TCN has 2520 parameters and executes 164 kMAC.

C. Baseline and Time-Consistent Setup

Inspiring works have addressed the Short-Term SWEC-ETHZ iEEG Database with automated learning [11], using models both from classical Machine Learning (Random Forest, Support Vector Machine, Multi-Layer Perceptron, and Hyper-Dimensional Computing) and from Deep Learning (2d-CNN and LSTM).

The current SoA algorithm on the targeted dataset is a Hyper-Dimensional Computing (HDC) Ensemble applied on 3 signal features, namely Local Binary Pattern, Line Length, and Amplitude, which provides a seizure detection with a specificity of 97.3 % and a detection delay of 8.81 s, while missing only 3.6 % of the dataset's seizures [11]. In particular, Table I details how the HDC Ensemble stands out as the SoA baseline by attaining better specificity and detection delay compared to deep models such as 2d-CNN and LSTM, at a comparable miss rate.

A limitation of all the aforementioned works on the Short-Term SWEC-ETHZ iEEG dataset is that training and test seizures are not temporally consistent: training seizures do not always precede testing ones. This is because the focus of all the cited approaches is the determination of the minimum number of training seizures required to have a good recognition on unseen seizures (*few-shot learning*), regardless of chronological order. In contrast, we are interested in time consistency in this work. Hence, we performed training on the first half of each patient's seizures, and test on the second half. Note in the clinical practice the training would indeed be performed on EEG traces and seizures that happened in the past, with the goal of detecting future ones. Though epileptic seizures of an individual patient are generally considered to be very similar, in a recent landmark study Schroeder et al. have reported that they found significant variability in seizure evolutions, with more similar seizures occurring closer together in time [27]. Furthermore, others have observed that seizure patterns and severity may change under conditions such as pre-surgical evaluation, when anti-epileptic drugs are often rapidly tapered to provoke seizures and thus shorten the time needed to obtain enough information for a decision about the feasibility of surgically removing the epileptogenic brain regions [28]. To perform a fair comparison, we apply the time-consistent training setup to both the SoA HDC approach and our EEGNet-inspired TCN.

TABLE I
SUMMARY OF THE RESULTS OF [11], INDICATING THE HDC ENSEMBLE AS SOA BASELINE ON THE SHORT-TERM SWEC-ETHZ IEEG DATABASE, AGAINST OTHER DEEP MODELS.

Model	Missed seizures	Specificity	Detection delay
HDC Ensemble	3.6%	**0.973**	**8.81s**
STFT + 2d-CNN	**2.3%**	0.836	17.9s
raw iEEG + LSTM	4.7%	0.948	14.7s

D. Details on the Machine Learning Setup

1) Timing: Both the HDC Ensemble and the TCN are fed with 1 s-windows of the multichannel iEEG signal at 512 Hz. The HDC performs feature extractions as described in II-C, whereas the TCN directly executes convolutions on the raw signal. For training, windows are taken with a slide of 0.5 s for HDC and 32 ms for the TCN; at inference time, the slide is 0.5 s for both algorithms, thus delivering 2 inferences per second.

2) TCN Training: The TCN (implemented in PyTorch 1.6) was trained with binary cross-entropy loss, AdaM optimizer, initial learning rate 0.001, and minibatch size 64, for 15 epochs in `float32` plus 1 epoch in `int8` (quantization details in II-E).

3) Post-processing and Delay-Specificity Curves: We post-process both the HDC's and TCN's outputs by a n-sample checker, as per [11]: after each inference, a window of n model outputs (0.5 s apart, as detailed in II-D.1) in the past is considered, and a positive label is returned only if all n are positive. By using different values of n (starting from $n = 1$, i.e. no post-processing), we explore different trade-offs between specificity and detection delay (which will be defined in Subsection III-A). A smaller n means a shorter window, hence a milder attenuation of positives, thus prioritizing high sensitivity and short detection delay

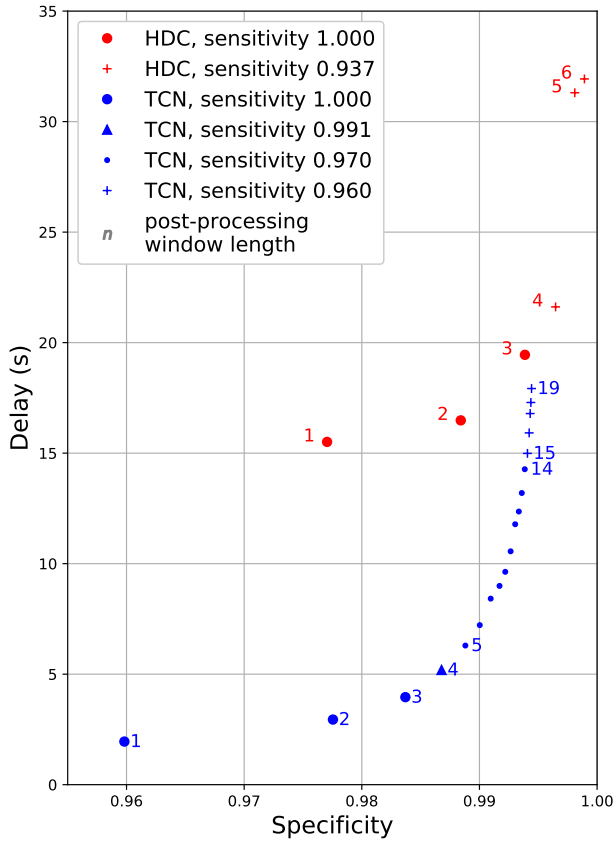

Fig. 2. Delay-specificity curves of HDC and TCN, obtained by applying a different amount of post-processing that inhibits positives.

over specificity. Conversely, a larger n takes a longer window, leading to a stronger attenuation of positives, yielding higher specificity at the cost of lower sensitivity and longer detection delay. In this way, we are able to characterize the HDC and TCN in terms of specificity-delay curves.

E. Network Quantization and Deployment

After training the TCN in `float32` format for 15 epochs, we applied 8-bit Post-Training Quantization and ran 1 further epoch of quantization-aware training to recover accuracy. The quantization method applied was the PArameterized Clipping acTivation (PACT) [29], as implemented in the open-source library NEMO (NEural Minimizer for tOrch, [30], [31]), developed to minimize CNN network memory footprint and latency to enable implementation on resource-constrained ultra-low-power platforms. In particular, quantizing our 2520-parameter model from `float32` to `int8` cuts its memory footprint by $4\times$, from 10.08 kB to 2.52 kB.

We deployed our `int8`-TCN on the Parallel Ultra-Low Power (PULP) microcontroller GAP8 [13], [14], to measure the inference latency and the energy cost per inference. To do so, we exploited the open-source tool DORY (Deployment Oriented to memoRY, [32]), using an extension to the backend to enable the support of dilated convolutional layers.

TABLE II
DEPLOYMENT METRICS OF OUR PROPOSED TCN COMPARED AGAINST THE SOA HDC ALGORITHM.

Model:	**HDC Ensemble** [11]	**EEGNet-inspired TCN (this work)**	
Memory	17.8 kB	2.52 kB (0.14×)	
Arithm. op.	32.8 M	328 k (0.01×)[1]	
Platform:	**Quentin**	**GAP8** [13], [14] **1-core**	**8-core**
V_{DD}	0.52 V	2.8 V	
f_{CLK}	187 MHz	100 MHz	
Cycles (k)	33100	568.1 ± 0.6 (0.017×)	146.4 ± 0.2 (0.005×)
Latency (ms)	177.0	5.681 ± 0.006 (0.032×)	1.464 ± 0.002 (0.009×)
Energy (µJ)	287.9	124.52 ± 0.27 (0.43×)	51.19 ± 0.13 (0.18×)

[1] 1 MAC = 2 arithmetic operations

III. Experimental Results

A. Evaluation Metrics

We evaluate the detection of epileptic seizures using three metrics, following the standard of the previous works on the dataset [11]:

- *sensitivity*: the fraction of detected seizures for each patient; seizures are detected when the classifier returns at least 1 True Positive inference, i.e. at least 1 positive inference over the ictal segment ($t > 180.0$ s in all recordings); note that this sensitivity is defined per-patient as a count over seizures, not over single inferences within a single seizure;
- *specificity*, defined as the fraction of True Negative inferences over the inter-ictal segment ($t < 180.0$ s in all recordings);
- *detection delay*, measured as the time distance between the ground-truth seizure onset (at instant $t_0 = 180.0$ s in all recordings) and the first True Positive inference[2]; following the definition in the baseline work [11], undetected seizures are discarded from the calculation of the average delay; doing so is fair as long as sensitivity is high, i.e. very few of the 100 seizures included in the dataset are missed.

B. Delay-specificity Pareto Frontier

The results of recognition of our TCN and of the baseline HDC Ensemble are shown in Figure 2. The plot displays the specificity-delay curve obtained for each model by varying n for the n-sample checker, i.e. the length of the post-processing window, as explained in Subsection II-D.3. With no post-processing (i.e., $n = 1$), all the positive outputs are retained, leading to the configuration that most favors a short detection delay over higher specificity. Increasing the number of samples used for post-processing removes

[2] Note that this *detection delay* is distinct from the *computation latency* to execute the model inference. In our solution, the latter is negligible compared to the former, as explained in the Results in Subsection III-C.

a higher fraction of positives, shifting the tradeoff toward higher specificity, at the cost of an increased detection delay.

Since the average detection delay is well-defined only at high sensitivity, i.e. when few seizures are missed (as discussed in Subsection III-A), we consider the curve points valid only for sensitivity > 0.93, and we stop the upper-right end of the curves at this threshold.

Remarkably, our TCN is able to provide a shorter detection delay, when specificity is in the interval $[0.975, 0.995]$. Thus, our TCN constitutes the Pareto frontier in this region. Furthermore, if the sensitivity requirement is raised above 0.95, the Pareto frontier is entirely represented by our TCN only. This is because the high-specificity points of the HDC curve have sensitivity 0.937, whereas the high-specificity points of the TCN curve always have sensitivity ≥ 0.96. The TCN's sensitivity-specificity tradeoff is thus more robust, because working points with sensitivity below the set threshold (application-dependent) are not valid, leaving only TCN points on the Pareto frontier[3].

In general, the sensitivity threshold is dictated by the desiderata of each particular scenario, and is application-dependent just like the preferred delay-specificity tradeoff point chosen on the Pareto curve for a specific use-case. Our results show that, depending on the desired sensitivity requirement, our TCN improves the Pareto frontier compared to the current SoA approach.

C. Deployment on Parallel MCU

Finally, we deployed our 8-bit TCN on the multi-core MCU GAP8 [13], [14], specialized for deep learning applications at the edge. Table II reports the deployment figures of merit, compared with the SoA HDC Ensemble. Our model requires just $2.52\,\mathrm{kB}$ of model parameters storage and $164\,\mathrm{kMAC} = 328\,\mathrm{k}$ arithmetic operations, as detailed in II-B and II-E. These values are $7.1\times$ and $100\times$ lower than the requirements of the HDC Ensemble SoA, respectively. The experimental values of computation latency and energy consumption were measured running the model on GAP8 at $V_{\mathrm{DD}} = 2.8\,\mathrm{V}$ and $f_{\mathrm{CLK}} = 100\,\mathrm{MHz}$. Averages and standard deviations were taken over 20 repetitions of the model execution. The energy cost E was determined experimentally by measuring the consumed current $i(t)$ and integrating it over the model execution time: $E = V_{\mathrm{DD}} \int_0^T i(t)\mathrm{d}t$, where $[0, T]$ is the time interval required of the execution, which was identified experimentally. These measurements yielded relative uncertainties of the order of 10^{-3}; this variability across repetitions is due to unpredictable cache effects of the GAP8 processor. This amount of variability is negligible for end-to-end use. Using 1 core, each inference requires on average just $5.68\,\mathrm{ms}$ of computation latency and $124.5\,\mu\mathrm{J}$ of energy cost. Both these values are better than the SoA. Moreover, parallelizing the inference on all the 8 cores of GAP8 shortens the latency to $1.46\,\mathrm{ms}$ and decreases the energy consumption to $51.2\,\mu\mathrm{J}$. It is to remark that this $5.68\,\mathrm{ms}$ computation latency is negligible compared to the detection delay, which is of the order of seconds, as shown in Figure 2.

The obtained memory footprint, latency, and energy consumption prove that our solution successfully meets the requirements for implementation on resource-constrained devices.

[3]Even if HDC can maximize specificity, reaching higher values compared to TCN, the HDC's maximum-specificity points have lower sensitivity and higher delay.

IV. Conclusions

We have presented a deep learning technique to address the iEEG-based detection of epileptic seizures in real-time. We applied a Temporal Convolutional Network (TCN) directly on raw iEEG, imposing a time-consistent setup, which ensures that training seizures always precede testing ones. Maintaining temporal consistency is a realistic constraint, which must be taken into account for deployment of detection algorithms, and which was not yet considered in previous works on the Short-Term SWEC-ETHZ iEEG dataset. Our TCN attained a shorter detection delay compared to the SoA algorithm, at constant sensitivity and specificity, thus improving the Pareto curve of these metrics. Furthermore, we deployed and profiled our solution on the commercial Parallel Ultra Low Power (PULP) microcontroller GAP8, demonstrating a shorter computation latency and better energy consumption.

References

[1] K. M. Fiest, K. M. Sauro, S. Wiebe, S. B. Patten, C.-S. Kwon, J. Dykeman, T. Pringsheim, D. L. Lorenzetti, and N. Jetté, "Prevalence and incidence of epilepsy: a systematic review and meta-analysis of international studies," *Neurology*, vol. 88, no. 3, pp. 296–303, 2017.

[2] L. Kalilani, X. Sun, B. Pelgrims, M. Noack-Rink, and V. Villanueva, "The epidemiology of drug-resistant epilepsy: a systematic review and meta-analysis," *Epilepsia*, vol. 59, no. 12, pp. 2179–2193, 2018.

[3] M. Hirsch, D.-M. Altenmüller, and A. Schulze-Bonhage, "Latencies from intracranial seizure onset to ictal tachycardia: a comparison to surface EEG patterns and other clinical signs," *Epilepsia*, vol. 56, no. 10, pp. 1639–1647, 2015.

[4] C. Rummel, E. Abela, R. G. Andrzejak, M. Hauf, C. Pollo, M. Müller, C. Weisstanner, R. Wiest, and K. Schindler, "Resected brain tissue, seizure onset zone and quantitative EEG measures: towards prediction of post-surgical seizure control," *PLoS One*, vol. 10, no. 10, p. e0141023, 2015.

[5] B. C. Munsell, C.-Y. Wee, S. S. Keller, B. Weber, C. Elger, L. A. T. da Silva, T. Nesland, M. Styner, D. Shen, and L. Bonilha, "Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data," *Neuroimage*, vol. 118, pp. 219–230, 2015.

[6] A. K. Jaiswal and H. Banka, "Local pattern transformation based feature extraction techniques for classification of epileptic EEG signals," *Biomedical Signal Processing and Control*, vol. 34, pp. 81–92, 2017.

[7] S. N. Baldassano, B. H. Brinkmann, H. Ung, T. Blevins, E. C. Conrad, K. Leyde, M. J. Cook, A. N. Khambhati, J. B. Wagenaar, G. A. Worrell *et al.*, "Crowdsourcing seizure detection: algorithm development and validation on human implanted device recordings," *Brain*, vol. 140, no. 6, pp. 1680–1691, 2017.

[8] Y. Yuan, G. Xun, K. Jia, and A. Zhang, "A multi-view deep learning framework for EEG seizure detection," *IEEE journal of biomedical and health informatics*, vol. 23, no. 1, pp. 83–94, 2018.

[9] R. Hussein, H. Palangi, Z. J. Wang, and R. Ward, "Robust detection of epileptic seizures using deep neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 2546–2550.

[10] Y. Nagahama, A. J. Schmitt, D. Nakagawa, A. S. Vesole, J. Kamm, C. K. Kovach, D. Hasan, M. Granner, B. J. Dlouhy, M. A. Howard *et al.*, "Intracranial EEG for seizure focus localization: evolving techniques, outcomes, complications, and utility of combining surface and depth electrodes," *Journal of neurosurgery*, vol. 130, no. 4, pp. 1180–1192, 2018.

[11] A. Burrello, S. Benatti, K. A. Schindler, L. Benini, and A. Rahimi, "An ensemble of hyperdimensional classifiers: Hardware-friendly short-latency seizure detection with automatic iEEG electrode selection," *IEEE journal of biomedical and health informatics*, 2020.

[12] W. Stacey, M. Le Van Quyen, F. Mormann, and A. Schulze-Bonhage, "What is the present-day EEG evidence for a preictal state?" *Epilepsy research*, vol. 97, no. 3, pp. 243–251, 2011.

[13] E. Flamand, D. Rossi, F. Conti, I. Loi, A. Pullini, F. Rotenberg, and L. Benini, "GAP-8: A RISC-V SoC for AI at the Edge of the IoT," in *2018 IEEE 29th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*. IEEE, 2018, pp. 1–4.

[14] https://greenwaves-technologies.com/gap8_gap9/.

[15] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of neural engineering*, vol. 15, no. 5, p. 056013, 2018.

[16] N. Waytowich, V. J. Lawhern, J. O. Garcia, J. Cummings, J. Faller, P. Sajda, and J. M. Vettel, "Compact convolutional neural networks for classification of asynchronous steady-state visual evoked potentials," *Journal of neural engineering*, vol. 15, no. 6, p. 066031, 2018.

[17] A. Bablani, D. R. Edla, D. Tripathi, and R. Cheruku, "Survey on brain-computer interface: An emerging computational intelligence paradigm," *ACM Computing Surveys (CSUR)*, vol. 52, no. 1, pp. 1–32, 2019.

[18] J. Parvizi and S. Kastner, "Promises and limitations of human intracranial electroencephalography," *Nature neuroscience*, vol. 21, no. 4, pp. 474–483, 2018.

[19] J. Becedas, "Brain–machine interfaces: basis and advances," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 825–836, 2012.

[20] A. Nirkko, J. Slotboom, K. Schindler, and C. Rummel, "Die gemeinsame referenz bei der elektroencephalografie: Medianwerte im vergleich zu mittelwerten," *Klinische Neurophysiologie*, vol. 40, no. 01, p. P318, 2009.

[21] W. A. Ríos-Herrera, P. V. Olguín-Rodríguez, J. D. Arzate-Mena, M. Corsi-Cabrera, J. Escalona, A. Marín-García, J. Ramos-Loyo, A. L. Rivera, D. Rivera-López, J. F. Zapata-Berruecos *et al.*, "The influence of EEG references on the analysis of spatio-temporal interrelation patterns," *Frontiers in neuroscience*, vol. 13, p. 941, 2019.

[22] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks for action segmentation and detection," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 156–165.

[23] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[24] M. Zanghieri, S. Benatti, A. Burrello, V. Kartsch, F. Conti, and L. Benini, "Robust real-time embedded emg recognition framework using temporal convolutional networks on a multicore iot processor," *IEEE transactions on biomedical circuits and systems*, vol. 14, no. 2, pp. 244–256, 2019.

[25] T. Schneider, X. Wang, M. Hersche, L. Cavigelli, and L. Benini, "Q-EEGNet: An energy-efficient 8-bit quantized parallel EEGNet implementation for edge motor-imagery brain-machine interfaces," in *2020 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, 2020, pp. 284–289.

[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[27] G. M. Schroeder, B. Diehl, F. A. Chowdhury, J. S. Duncan, J. de Tisi, A. J. Trevelyan, R. Forsyth, A. Jackson, P. N. Taylor, and Y. Wang, "Seizure pathways change on circadian and slower timescales in individual patients with focal epilepsy," *Proceedings of the National Academy of Sciences*, vol. 117, no. 20, pp. 11 048–11 058, May 2020. [Online]. Available: https://doi.org/10.1073/pnas.1922084117

[28] P. Q. Duy, G. L. Krauss, N. E. Crone, M. Ma, and E. L. Johnson, "Antiepileptic drug withdrawal and seizure severity in the epilepsy monitoring unit," *Epilepsy & Behavior*, vol. 109, p. 107128, Aug. 2020. [Online]. Available: https://doi.org/10.1016/j.yebeh.2020.107128

[29] J. Choi, Z. Wang, S. Venkataramani, P. I.-J. Chuang, V. Srinivasan, and K. Gopalakrishnan, "Pact: Parameterized clipping activation for quantized neural networks," *arXiv preprint arXiv:1805.06085*, 2018.

[30] https://github.com/pulp-platform/nemo.

[31] F. Conti, "Technical report: Nemo dnn quantization for deployment model," 2020.

[32] A. Burrello, A. Garofalo, N. Bruschi, G. Tagliavini, D. Rossi, and F. Conti, "Dory: Automatic end-to-end deployment of real-world dnns on low-cost iot mcus," 2020.