Reducing False Alarms in Wearable Seizure Detection with EEGformer: A Compact Transformer Model for MCUs

Paola Busia, Andrea Cossettini, Thorir M. Ingolfsson, Simone Benatti, Alessio Burrello, Victor J. B. Jung, Moritz Scherer, Matteo A. Scrugli, Adriano Bernini, Pauline Ducouret, Philippe Ryvlin, Paolo Meloni, Luca Benini

***Abstract*—The long-term, continuous analysis of electroencephalography (EEG) signals on wearable devices to automatically detect seizures in epileptic patients is a high-potential application field for deep neural networks, and specifically for transformers, which are highly suited for end-to-end time series processing without handcrafted feature extraction. In this work, we propose a small-scale transformer detector, the EEGformer, compatible with unobtrusive acquisition setups that use only the temporal channels. EEGformer is the result of a hardware-oriented design exploration, aiming for efficient execution on tiny low-power micro-controller units (MCUs) and low latency and false alarm rate to increase patient and caregiver acceptance.**

Tests conducted on the CHB-MIT dataset show a 20% reduction of the onset detection latency with respect to the state-of-the-art model for temporal acquisition, with a competitive 73% seizure detection probability and 0.15 false-positive-per-hour (FP/h). Further investigations on a novel and challenging scalp EEG dataset result in the successful detection of 88% of the annotated seizure events, with 0.45 FP/h.

We evaluate the deployment of the EEGformer on three commercial low-power computing platforms: the single-core Apollo4 MCU and the GAP8 and GAP9 parallel MCUs. The most efficient implementation (on GAP9) results in as low as 13.7 ms and 0.31 mJ per inference, demonstrating the feasibility of deploying the EEGformer on wearable seizure detection systems with reduced channel count and multi-day battery duration.

***Index Terms*—deep learning, electroencephalography, time traces, transformer, wearable**

I. Introduction

Epilepsy is a common neurological disorder causing the recurrence of seizures temporarily compromising brain function. Wearable seizure-detecting solutions could enable prompt interventions from caregivers during or immediately after the seizures to reduce their impact and provide physicians with more reliable information for optimizing therapy. Currently, seizure detectors approved by health authorities only detect generalized convulsive seizures, which account for less than 20% of all seizures, relying on other signals than electroencephalography (EEG) despite the fact that the latter provides the hallmark of the brain's epileptic activity [1]. This is due to the lack of unobtrusive and non-stigmatizing EEG systems suitable for very long-term monitoring [2], which could potentially leverage our capacity to detect all seizure types. Moreover, despite analyses in epilepsy monitoring units being essential for identifying seizure types, long-term recording in ambulatory patients also appears as a significant challenge that needs to be addressed to improve the estimation of seizure occurrence and the notification of seizure events.

While several artificial intelligence models have been tested for the detection of seizures based on complete, high-electrode count EEG acquisition setups, there is still a need for efficient solutions targeting the wearable domain and managing to reach the required accuracy standards based on low-channel count disguisable acquisition devices. In particular, minimizing false alarms appears as the major goal to enable long-term monitoring [3]. In this work, we assess the performance as a seizure detector of the EEGformer model, first presented in [4]. The EEGformer is a compact transformer model for online epilepsy monitoring, designed to target low-power devices causing minimal discomfort to the patient [5, 6], thanks to a small acquisition setup limited to the temporal channels, a memory footprint of 50.6K parameters and a complexity of 14.7MOPS. EEGformer operates on the raw EEG signal, and it represents an adaptable solution combining data-driven feature extraction and classification. EEGformer targets wearable epileptic seizure devices for everyday life use. As such, it focuses on minimizing the false alarm rate, as requested by patients and caregivers [3].

A brief outline of the contents of the paper is given in the following. After revising in Section II the state of the art of EEG processing and epilepsy monitoring approaches, Section III presents the architectural description of EEGformer. In Section IV, we assess its performance on the state-of-art CHB-MIT dataset [7, 8], considering the most typical performance metrics. Particularly, we target minimizing the false alarm rate and the onset detection latency.

Given the encouraging results of this first assessment, Section V presents a novel scalp-EEG epilepsy dataset recorded at Epilepsy Monitoring Units (EMUs), for which no signal cutting and reduction has been made in order to preserve

This project was supported by the Swiss National Science Foundation (Project PEDESITE) [grant number 193813], and by the SEVERITY Swiss NSF/Div3 project [grant number 320030-179240].

Paola Busia, Matteo A. Scrugli, and Paolo Meloni are with the DIEE, University of Cagliari, Cagliari, Italy (e-mail: paola.busia@unica.it , matteoa.scrugli@unica.it , paolo.meloni@unica.it).

Andrea Cossettini, Thorir M. Ingolfsson, Victor J. B. Jung, Moritz Scherer, and Luca Benini are with the Integrated Systems Laboratory, ETH Zürich, Zürich, Switzerland.

Simone Benatti, Alessio Burrello, and Luca Benini are with the DEI, University of Bologna, Bologna Italy.

Simone Benatti is with the DISMI, University of Modena and Reggio Emilia, Reggio Emilia, Italy

Adriano Bernini, Pauline Ducouret, and Philippe Ryvlin are with the Lausanne University Hospital (CHUV), Lausanne, Switzerland.

the natural unbalance of seizure events vs normal state. We evaluate the EEGformer on this new dataset, discussing the challenges of training a classifier on the data typically available in clinical practice. The comparison with the state of the art of seizure detectors is discussed in Section VI. Finally, in Sections VII-A and VII-B we deploy the EEGformer on three resource-constrained platforms suitable for low-power continuous health monitoring, exploiting parallel execution to speed up the computations and reduce the energy consumption of the monitoring device.

This work significantly extends the preliminary results presented in [4]. The main novel contributions of this paper are:

- presentation of a novel scalp-EEG dataset for epilepsy monitoring, providing a test benchmark close to the clinical practice;
- first-time seizure-detection assessment on the novel dataset, achieving detection of 88% of the annotated seizure events and reaching 0.45 FP/h with the EEGformer;
- demonstration of state-of-the-art performance on the CHB-MIT dataset for systems with a small number of electrodes (4), detecting 73% of the seizure events while guaranteeing an FP/h rate as low as 0.15 (with 5 out of 8 tested patients exhibiting zero false alarms);
- first-time implementation of EEGformer on two parallel ultra-low-power architectures of the GAP family of processors: GAP8 and GAP9;
- new state-of-the-art energy efficiency for a transformer embedded implementation on a parallel RISC-V architecture, reaching 13.7ms inference time and 0.31 mJ/inference energy consumption, approx. 5× lower than the implementation on the Apollo4 platform [4].

II. Related Work

A rich literature describes the EEG processing solutions targeting the epilepsy monitoring task [9, 10, 11], and the best-performing approaches reach up to perfect sensitivity, with a false alarm rate limited to 0.04 FP/h, considering subject-specific models, and acquisition from a large number of channels spread over the whole surface of the head [12]. However, long-term monitoring in normal life conditions requires non-stigmatizing wearable devices, where the acquisition is limited to minimal recording setups with a reduced number of channels. As such, approaches based on full electrode coverage are impractical. In fact, for wearable long-term monitoring scenarios (which is the main target of this paper), the following constraints apply:

- compact detection model size, to fit on an embedded platform, and low energy consumption suitable for long-term monitoring on low-power wearable devices;
- low number of acquisition channels, placed on the temporal region, as required for compact and easily concealable wearable solutions [5, 6, 13];
- minimized false alarms, even at the expense of lower sensitivity, to guarantee that the final user will be able and willing to use the device [2, 3, 14];
- minimized detection latency to promptly issue alarms.

Accurate seizure detection becomes even more challenging when this set of constraints is applied. In the following, we limit the state-of-the-art (SoA) discussion to works contributing to this challenge, by presenting solutions based on a reduced acquisition setup, exploiting less than 8 channels localized in the temporal region.

Several works explored seizure detection based on reduced channel count EEG acquisition during at-home real-time monitoring [15, 16] or in-hospital recording [17]. The authors of [17] presented a behind-the-ear recording system, feeding a subject-specific Support Vector Machine (SVM), and tested it on the data collected from 54 patients by the University Hospital Leuven. Long-term monitoring through 2-channels subcutaneous acquisition was presented in [15], where a residual Convolutional Neural Network (CNN) classifier was trained and tested on 490 days of EEG recordings from 9 patients with epilepsy and 12 control healthy patients. Similarly, the authors of [16] examined the EEG recordings of 102 patients, acquired with a single-channel headband and including 364 seizure events, exploiting a CNN classifier for seizure detection. These relevant contributions have the merit of assessing their proposed systems on real-life monitoring scenarios, showing how a generally limited performance is achieved, especially in terms of event-level seizure detection sensitivity. However, a direct comparison with these results is not feasible, due to the private nature of the referenced datasets.

Limited channel count acquisition was also emulated on the open-source CHB-MIT dataset. An example is represented by the work of [18], where a K-nearest-neighbor (KNN) classifier working only on the data acquired from 5 selected channels was considered. However, this solution is not unobtrusive, as the data of the complete acquisition setup was considered in order to guide the channel selection. Furthermore, a common issue is represented by a low specificity value, incompatible with a comfortable user experience [2, 3].

Suboptimal specificity is also reported in the works of [19, 20] and [21]. In particular, the authors of [19] exploited discrete wavelet transform pre-processing on only 2 acquisition channels and classification based on the Random-Forest (RF) model, whereas [20] presented an energy-efficient wearable seizure detection system, based on 8-channel frontal lobe acquisition and an SVM classifier, and [21] presented a CNN model with 4 convolutional layers, applied to statistical and power features extracted after tunable Q-wavelet.

Our main reference for the detection task based on low-channel count acquisition is provided by the work of [22], which was able to reach perfect sensitivity and specificity for some patients on the CHB-MIT dataset, with subject-specific training and careful exploration and tuning of the signal windowing. The best results were obtained with the RF and AdaBoost (AB) classifiers, applied on features representing the energy after 4-level Haar-wavelet decomposition of the signal acquired only by the temporal channels. However, these promising numbers were obtained on a very small subset of the CHB-MIT data.

The main limitations of the listed works will be quantitatively discussed in Section VI. In general, although well-suited for truly wearable long-term monitoring as relying on a reduced set of acquisition channels, these solutions still report a too-high number of false positives [19, 20, 21] and long detection latency [22] compared to the approach presented in this paper. Furthermore, removing the complexity of the feature extraction step and relying on data-driven features would be of interest to enhance the adaptability to new subjects and datasets.

To address these challenges, in [4] we presented EEGformer, a transformer model for subject-specific seizure detection based on raw EEG signals from temporal channels, assessing its performance on the CHB-MIT benchmark dataset and demonstrating its deployment on a single-core low-power microcontroller (Apollo4). In comparison with the other transformer-based seizure detectors from the literature, it does not rely on a full acquisition setup [23, 24, 25]. Furthermore, these solutions exploit complex architectures, stacking 4 or 6 encoder blocks [24, 25], or three parallel encoder towers [23], and thus are not suitable for efficient wearable deployment.

The EEGformer satisfies the efficiency constraints of compact and wearable monitoring solutions, with a complexity and memory footprint suitable for efficient execution on tiny microcontroller units, and an energy consumption compatible with low-power long-term monitoring (see Section VII). The proposed model aims at reducing the false-alarm rate and the onset detection latency of existing solutions [19, 20, 21, 22].

In this work, we examine in more depth the preliminary results of [4], by extending the assessment of the EEGformer to a novel dataset, demonstrating performance aligned with similar solutions at the state-of-the-art tested on private clinical data [15, 16], especially in terms of event-level sensitivity, and with a comparable tree-based approach presented in Section VI-B. Moreover, to the best of our knowledge, this paper presents the first deployment of a transformer-based algorithm for non-obstructive seizure monitoring on parallel microcontrollers, targeting two RISC-V multi-processor platforms, where state-of-the-art time for inference and energy efficiency were achieved.

III. EEGformer

In the following, we describe our proposed seizure detector model, EEGformer. Since we target an unobtrusive acquisition setup, we only consider data acquisition from the temporal channels (F7-T7, T7-P7, F8-T8, T8-P8, according to the 10-20 international system), which is compatible with using non-stigmatizing wearable EEG devices such as over-ear headphones, headbands, or e-glasses [13, 22]. We translated the epilepsy monitoring problem into the periodic classification of the raw EEG signal, avoiding the need for handcrafted feature extraction, into a non-seizure or a seizure class.

TABLE I: Parameters and topology of the EEGformer.

Stage	Layer	Parameters[1]
Embedding stage	Convolution0	Input: 4×W× 1, Kernel: K×1 Output: E×S'×1, Stride: K
	Convolution1	Input: E×S'×1, Kernel: K×1 Output: E×S×1 Stride: K
	Add Pos. Encoding	Input/Output: E×S×1
Encoder stage	Layer Norm	Input/Output: E×S×1
	Multi-Head-Attention	Embedding: E, Sequence: S q,k,v projections: d, Heads: H
	Layer Norm	Input/Output: E×S×1
	Dense0	Input: E×S×1 Output: h×S×1
	Dense1	Input: h×S×1 Output: E×S×1
	Layer Norm	Input/Output: E×S×1
Classification stage	Reduce Mean	Input: E×S×1, Output: E
	Dense2	Input: E, Output: 2

[1] In the EEGformer W=2048, K=5, E=32, S'=405, S=81, d=32, H=8, and h=128.

TABLE II: Classification performance and complexity of the design points considered for the EEGformer parameter exploration.

Parameter	Model	Accuracy	Footprint	MOPS
Window size	w2_K5_H8_h128	99.33	66 kB	2.4
	w4_K5_H8_h128	99.44	87 kB	5.6
	w8_K5_H8_h128	**99.5**	**150 kB**	**14.7**
	w16_K5_H8_h128	99.27	355 kB	43.3
Kernel size	w8_K10_H8_h128	99.39	61 kB	2.1
	w8_K3_H8_h128	99.55	**585 kB**	73.3
# Heads	w8_k5_H4_128	99.39	87 kB	8.7
	w8_K5_H16_128	99.44	275 kB	26.7
Hidden size	w8_K5_H8_64	99.44	146 kB	14
	w8_K5_H8_h256	**99.5**	**158 kB**	**16**

The EEGformer architecture is based on a Vision Transformer ([26]), where the input image is replaced by a 4-row matrix of consecutive samples, acquired with a 256Hz sampling rate. Each row in the input matrix corresponds to one of the channels of interest. Figure 1 provides a general overview of our classification system, whereas Table I describes the general network topology considered for the design of the EEGformer, inspired to [27]. The network architecture is composed of three main processing stages. The first *embedding stage* performs data preparation to adapt the matrix of EEG samples for transformer-based processing. According to the experience of Bioformers [27], the embedding stage is configured as a sequence of two 1D-convolutional layers, applying non-overlapped filtering kernels of size K to the input signal. The output of this convolution block has size $E \times S$, where E is the number of output channels and S is the resulting data length, which is reduced based on the size K of the filtering kernels. This mechanism thus resembles the image decomposition into a sequence of embedded patches, which is performed in [26], where S is interpreted as the sequence length and E as the size of the embedded patches. A set of learned positional weights is also added to the flattened sequence of patches to encode ordering information.

The core of the algorithm is the *encoder stage*, introducing the most relevant transformer mechanism: the attention layer [28], indicated as MHA (Multi-Head-Attention) in Figure 1. First, three linear projections of the data are computed, called *query* q, *key* k and *value* v, each of size d. The attention matrix is obtained as the dot-product between these projections, according to Equation 1, resulting in a set of attention scores reflecting the mutual relevance between two points in the examined window.

$$\text{Attention}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \text{Softmax}(\frac{\mathbf{qk}^T}{\sqrt{\text{d}}})\mathbf{v} \tag{1}$$

MHA enhances the detecting power of the attention mechanism with multiple parallel threads, indicated as heads, performing independent projections of the input, and recovering

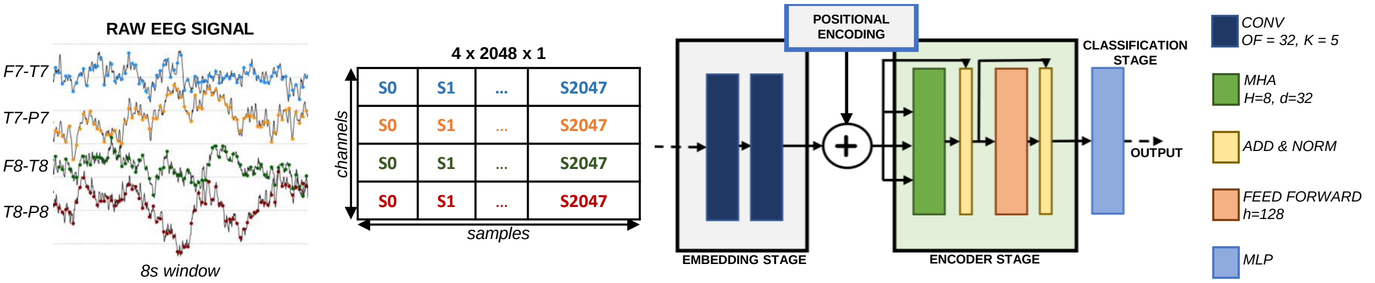


Fig. 1: Architecture of the EEGformer. The Multi-Head-Attention layer is indicated as MHA.

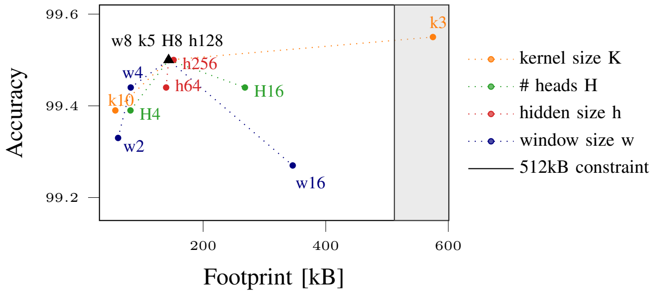


Fig. 2: Design points explored for EEGformer architecture definition. The selected point representing EEGformer is highlighted in black. The different curves represent the performance resulting from the selection of the parameters explored: window size, kernel size, number of heads, and hidden layer size. Shaded grey area: outside of the 512 kB feasibility region.

the input dimensionality with an additional final linear projection. The attention layer is typically followed by a feed-forward network and their sequence can be replicated multiple times, into a stack of encoder blocks. As it can be noticed from Table I, we considered a single encoder block, where the two main sublayers, the MHA and the feed-forward network, are combined with Layer Normalization and residual connections to their respective input. Each Dense layer in the feed-forward network is combined with GELU activation.

Finally, the *classification stage* evaluates the mean of the sequence resulting from the encoder processing and applies a Multi-Layer Perceptron (MLP) to compute the output probabilities. In the general topology considered, the MLP is collapsed into a single Dense layer.

The EEGformer model was implemented in TensorFlow Keras [29] and designed based on the general architecture in Table I, where the main parameters were selected with an exploration considering the working memory (SRAM) typically available on tiny MCUs (we set a maximum footprint of 512KB), and the seizure detection performance, evaluated on the CHB-MIT dataset, referencing subject 1 (further details on the assessment are given in Section IV). The accuracy is evaluated, based on its standard definition, as the ratio of correctly classified windows over the overall number of processed windows. The outcome of the exploration is summarized in the plot of Figure 2, where the evaluated design points are placed based on their accuracy and footprint.

We started from the evaluation of the input window size, defining a set of possible values $w = \{2s, 4s, 8s, 16s\}$. The remaining parameters were kept fixed, and equal to $K = 5$, $H = 8$, and $h = 128$. A window of $w = 8s$ resulted in the best accuracy and was considered for the rest of the exploration. As a second parameter, we tuned the kernel size, considering values of $K = 3$ (i.e., increasing the sequence length and the number of operations performed in each attention head) and $K = 10$. The first choice resulted in storage requirements non-compatible with the footprint constraint, whereas an accuracy drop was observed when increasing the kernel size. Hence, $K = 5$ was kept for the following explorations. Similarly, we evaluated the impact of reducing or increasing the number of parallel heads $H = \{4, 16\}$, and the size of the hidden layer in the feed-forward network $h = \{64, 256\}$.

The two best-performing combinations obtained are highlighted in Table II, which also reports a detailed analysis of the memory footprint and computational complexity of the evaluated models (for each of the explored parameter configurations). Table II and Fig. 2 also reveal that no further performance improvement is obtained by increasing the complexity through the choice of H and h, which have a limited impact on the network complexity, especially compared to w and K. The parameters selected for the EEGformer are detailed in Table I and result in a topology having 50.6K parameters, with a memory footprint of 150 kB and requiring the execution of 14.7 MOPS.

IV. Assessment on CHB-MIT Dataset

We consider in this section the CHB-MIT Scalp EEG dataset for seizure detection [7, 8], which has been an important reference for researchers over the years and provides a meaningful common benchmark to set the state-of-the-art context. As a reminder, we target to maximize specificity, since having nearly zero false alarms is a strict requirement for the acceptance of continuous monitoring devices by patients and caregivers [2, 3, 14].

Dataset description. The CHB-MIT dataset is an open-source collection of scalp EEG recordings from 23 pediatric patients, curated by the Children's Hospital Boston and the Massachusetts Institute of Technology. It provides a list of records of different duration, and a summary file reporting the expert annotations about the time of occurrence of seizure events. The data is collected with a 256Hz sampling frequency, including 18 to 23 channels.

Training strategy. As a first step, we have compared two training strategies, respectively consisting of a single-phase subject-specific training or of a two-phase approach, with a global subject-independent pre-training and subject-specific fine-tuning. In the single-phase solution, we perform 100 epochs of training, whereas, in the two-phase one, we dedicate 100 epochs to the first phase and 50 additional epochs to the second. To evaluate the detection performance, we define a test set obtained with the leave-one-out strategy, consisting of one record among those available for the test patient. Then we randomly split the remaining data between a training set and a validation set, with an 8:2 ratio. We alternatively test all the seizure records of the considered subject. The training and validation data are represented by non-overlapped windows of signal, whereas for a more detailed performance evaluation, we consider test data obtained as sliding windows of 8s length, overlapped with 2s intervals.

The comparison between the two strategies has taken patient CHB 1 as a test subject and has included the records provided for patients CHB 2 to 8 in the pre-training data.

In Table III we compare the two approaches. Observing the exploration results, we highlight the positive effect of the pre-training phase in improving the specificity of the detection, reflected in a 0 FP/h rate. This advantage does not compromise the percentage of seizure episodes detected, which is still 100%.

Detection performance assessment. We evaluate at this point, referring to the two-phase strategy, the detection performance of the EEGformer. The assessment covers a subset of 8 patients, whose seizure records are tested with the leave-one-out approach, after a pre-training phase, including the data of the other patients who are not the test subject. Exploiting a typical approach to filter out isolated errors in the classification, we post-process the inference output with a majority voting method, considering a buffer of 3 classified windows for the EEGformer. Moreover, considering that after a seizure episode the EEG signals appear as altered for several minutes, we neglect any FP registered within 15 minutes after the annotated end of the event.

We summarize in Table IV the performance metrics obtained with a comprehensive evaluation of 40 records including a seizure. EEGformer allows detecting 73% of the examined seizure events (32/44), with 65.5% average segment-level sensitivity (570 False Negatives (FNs) registered over 1652 seizure windows tested). This detection rate was obtained while preserving high specificity values, with 5/8 patients exhibiting 100% specificity.

Figure 3 shows a boxplot of the distribution of the event-level FP/h trend across the 40 tests performed. We achieved a median value of 0 FP/h, which is the false alarm rate registered in over 80% of the left-out records tested, with 5/40 outliers. Considering consecutive false positives as single events, we obtained a state-of-the-art false alarm rate of 0.15 FP/h, with the great majority of subjects exhibiting no false alarms at all. The maximum duration is reported in the corresponding column in Table IV and is mostly affected by two long-lasting EEG artifacts.

Among the performance metrics in Table IV, we report the detail of the event-level sensitivity on single patients: on 6/8 subjects, 100% of the annotated events are successfully detected, while the overall sensitivity is compromised by the poor performance obtained on patient CHB 6, presenting seizures lasting on average less than 16s. Finally, we also report the average onset detection latency: for every test, the latency value is evaluated based on the number of FNs registered at the beginning of the seizure, multiplied by the time interval between successive windows (which is 2s in our sliding test approach) and incremented to account for the majority voting delay.

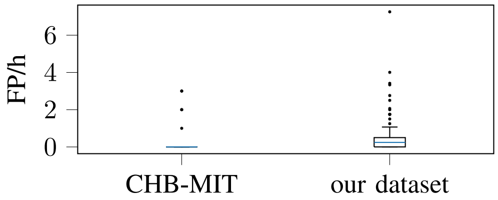


Fig. 3: Record-wise FP/h on the CHB-MIT dataset and our dataset with the EEGformer.

TABLE III: EEGformer training strategy evaluation on the CHB-MIT dataset, with 2s test evaluation period.

	Event-level Sensitivity	**Segment-level Sensitivity**	**Segment-level Specificity**	**FP/h**
w/o Pre-Training	100%	95.4%	99.9%	0.9
Pre-Training	100%	86.6%	100%	0

TABLE IV: Seizure detection on the CHB-MIT dataset with EEGformer.

Patient	**Segment-level**		**Event-level**		**Average Detection Latency**
	Sens	**Spec**	**Sens**	**FP/h; FP Event [s]**	
CHB1	80.5%	100%	100% ; 7/7	0 ; 0s	9.4s
CHB2	63.5%	99.3%	100% ; 3/3	1.8 ; 38s	11.3s
CHB3	75.5%	100%	100% ; 7/7	0 ; 0s	16.9s
CHB4	26.2%	99.9%	50% ; 2/4	0.3 ; 28s	21s
CHB5	79.8%	100%	100% ; 5/5	0 ; 0s	21.2s
CHB6	0%	100%	0% ; 0/10	0 ; 0s	-
CHB7	72%	100%	100% ; 3/3	0 ; 0s	16.7
CHB8	70%	99.9%	100% ; 5/5	0.8 ; 6s	14s
overall	**65.5%**	**99.9%**	**73% ; 32/44**	**0.15 ; 38s**	**15.2s**

V. Assessment on a private dataset

The results of the assessment conducted on the CHB-MIT dataset are very encouraging, showing the EEGformer can provide a good trade-off between sensitivity and specificity. Nevertheless, this dataset depicts a scenario that deviates considerably from a typical monitoring use case. To further assess the performance of EEGformer, we refer in this section to a novel dataset curated by the Lausanne University Hospital (CHUV), as a detection scenario that aligns more closely with practical clinical settings.

Dataset description. This dataset is acquired as a part of the currently running Pedesite study[1], during routine clinical evaluations at the in-hospital EMU, where patients are investigated in order to record and characterize their epileptic seizures. Patient monitoring lasts from 2 consecutive days up to two weeks. All the recording period is available. Approval for

[1] a project funded by the Swiss National Foundation that aims at developing innovative wearable solutions for seizure detection

TABLE V: Composition of our dataset from Lausanne CHUV.

Pat.	**Recording duration**	**# Seizures**	**Seizure *ID* and duration**
0	4d 10h 59min	3	***1***: 5min 2s; ***2***: 3min 4s; ***3***: 16min 19s
1	3d 20h 32min	7	***4***: 1min 15s; ***5***: 1min 54s; ***6***: 54s ***7***: 1min 26s; ***8***: 1min 11s; ***9***: 16s; ***10***: 29s
2	2d 21h 58min	5	***11***: 2min 14s; ***12***: 2min 54s; ***13***: 3min 15s ***14***: 2min 18s; ***15***: 2min 7s
3	3d 20h 13min	3	***16***: 1min 49s; ***17***: 2min 44s; ***18***: 3min 56s
4	3d 18h 23min	4	***19***: 1min 10s; ***20***: 1min 2s ***21***: 1min 12s; ***22***: 1min 5s
5	5d 19h 23min	3	***23***: 2min 17s; ***24***: 3min 2s; ***25***: 2min 39s

TABLE VI: Seizure detection on our dataset with EEGformer.

Patient	**Segment-level**		**Event-level**		**Det Latency; Seizure Duration**
	Sens	**Spec**	**Sens**	**FP/h ; FP event [s]**	
0	43.6	99.9	100 ; 3/3	0.2 ; 30s	34s ; 8min 8s
1	42.1	99.9	57 ; 4/7	0.1 ; 44s	43.5s ; 1min 3s
2	62.6	99.6	100 ; 5/5	0.9 ; 55s	57.8s ; 2min 34s
3	66	99.3	100 ; 3/3	0.6 ; 14min 4s	1min 6s;2min 50s
4	79	99.5	100 ; 4/4	0.5 ; 3min 50s	18.5s ; 1min 7s
5	29.1	99.8	100 ; 3/3	0.4 ; 42s	44.7s ; 7min 58s
overall	**50.2**	**99.7**	**88 ; 22/25**	**0.45 ; 14min 52s**	**38.1s ; 2min 38s**
AB [22]	71.5	95.7	92 ; 23/25	9.4 ; 4s	50.9s ; 2min 38s

retrospective data analysis with a waiver of informed consent due to the retrospective nature of the study was obtained from the local Ethical Committee of the University of Lausanne (study nr 2021-01419). The study report conforms to the STROBE statement for the report of observational cohort studies.

The dataset used in the present analysis is a subset of 6 patients from the overall study that has been curated by the CHUV. Table V summarizes the data provided for the examined patients, identified in column 1 with progressive ID numbers: for each one, we list the duration of the available EEG recordings and the number of annotated seizure events, as well as the duration of the seizures recorded. The data was acquired with an SD LTM PLUS 64[2] at 1024 Hz sampling frequency, with a setup of up to 24 Compumedics disposable Ag/AgCl sintered electrodes. A team of expert neurologists annotated the onset and end of the seizure events. However, it is important to note that the exact onset is at times uncertain due to several factors, including: 1) the epileptic discharge can occur in deep brain regions several seconds before it is detectable on scalp EEG, 2) similarly, clinical manifestations might not be present at seizure onset, and 3) EEG artifacts might obscure the seizure onset. It can also be difficult to precisely identify when the seizure ends. Due to these issues, we consider an uncertainty of 20s in the following.

Seizure detection with EEGformer. In the following, we analyze the performance obtained with EEGformer. As we are targeting unobtrusive devices, we select only the data acquired from the 4 temporal channels, down-sampled with a ratio of 4:1 to adapt it to the expected input dimensions reported in Figure 1, corresponding to an 8s window of the signal. As we did for the CHB-MIT dataset, we performed leave-one-out tests for cross-validation, including in the test set a significant number of non-seizure records. For each test record, we repeated the same train-test schedule:

- global pre-training on the seizure records of all subjects, excluding the test patient;
- subject-specific fine-tuning on the test patient, including all the training seizure records and at least three non-seizure records;
- test on the left-out record.

In this case, both training phases were conducted for 100 epochs, exploiting a weighted loss function to remedy the imbalance between seizure and non-seizure samples affecting the dataset. We also removed from the training and testing data the 15 minutes following the seizure occurrence, as during this interval, the signal is often unstable and affected by artifacts.

Table VI summarizes the detection performance for each of the patients after post-processing based on majority voting was applied. As can be derived from a general comparison with Table IV, this dataset represents a harder detection challenge than the CHB-MIT one. We will discuss in the following the possible reasons for this increased complexity, which suggested the selection of a different and slightly more complex averaging approach: we extended the averaging period to 15s around the examined window of signal and evaluated inference more frequently (one inference per second), to have more information available during the voting.

Furthermore, to recover the best detection sensitivity while still being able to filter out random FPs, we rely on an asymmetric voting criterium, defining a *non-seizure* and a *seizure state*. In the *non-seizure state*, an output seizure classification requires half of the windows in the voting buffer to be classified as seizure; then, once the voted output results in a seizure, a *seizure state* is entered, where only 1 over 4 of the buffered windows are required to be seizures for an output seizure classification. Finally, considering the uncertainty interval declared for the dataset annotations (labels have an uncertainty of ± 20 seconds), we excluded from the reported results the FPs occurring within 20s before the annotated onset, or alternatively the FNs occurring within 20s after the annotation. We finally report the event-level FP/h rate, where consecutive false positives are considered as a single event.

While the segment-level sensitivity value exceeds 60% only for 3/6 subjects, 88% of the examined seizure events were detected (22/25), with the exception of seizures 6, 9, and 10 of patient 1, having the shortest duration in the dataset. The segment-level sensitivity value results from 1774 FNs registered over 3560 seizure windows tested. Table VI also reveals that most of the seizures are detected with some delay, resulting in an increased average onset detection latency compared to the results reported in Table IV. Figure 3 reports on the right the record-wise event-level FP/h rate, in the 134 tests performed (23 seizure records and 111 non-seizure records), whose average duration is 3h and 45 min. The median value is 0.2 FP/h, compared to the 0 FP/h obtained on the CHB-MIT.

Discussion. We identify two reasons for the degradation of the expected detection performance: the uncertainty of the annotations and the presence of multiple unlabelled EEG artifacts. Both non-ideal characteristics are most likely present in a practical clinical scenario. While the presence of artifacts mostly affects the performance assessment during the test

[2] https://micromedgroup.com/products/brainquick/brainquick-ltm/

TABLE VII: Summary of scalp EEG-based seizure detection processing SoA, based on signal acquisition from a reduced number of channels (< 8).

Work	Subjects	Channels	Pre-processing	Detector	Event-level	Segment-level		FP/h
					Sensitivity	Sensitivity	Specificity	
CHB-MIT dataset								
Zeng et al. [18]	23	5 [a]	Kurtosis channel selection wavelet	KNN	-	**99.77%**	99.88%	-
Zanetti et al. [19]	23	2	Discrete wavelet transform	RF	-	96.6%	92.5%	0.7 [b]
Zhan et al. [20]	23	8	spectral energy	SVM	-	92.5%	80.1%	-
Ingolfsson et al. [22]	8	4	wavelet energy	AdaBoost	86% (38/44)	72%	**99.9%**	0.5 (4s) [c]
Mingkan et al. [21]	16	8	Tunable Q wavelet	CNN	98.90% (90/91)	-	97.87%	-
this work	**8**	**4**	**-**	**Transformer**	**73% (32/44)**	**65.5%**	**99.9%**	**0.15** (38s) [c]
proprietary clinical datasets								
Vandecasteele et al. [17]	54	4	Kurtosis, Entropy, Power,...[d]	SVM	-	63.4%	-	0.88 (10s) [c]
Remvig et al. [15]	21	2	-	CNN	86% (81/94)	-	-	0.08 - 0.5 [e]
Japaridze et al. [16]	102	1	-	CNN	79%	-	-	0.59 (2s) [c]
this work	6	4	-	Transformer	**88% (22/25)**	50.2%	**99.7%**	0.45 (844s)[3]

[a] Relies on 5 channels after a selection among all channels.
[b] Evaluated as FP/(FP+TP)×3600, where TP represents the number of True Positives.
[c] Maximum duration of the False Positive event. When not reported, it is assumed equal to the input window size.
[d] Zero-crossing, Skewness, Kurtosis, RMS amplitude, Total Power, Peak frequency, Power in frequency bands, Sample entropy, Shannon entropy, Spectral entropy, Power asymmetry in frequency bands.
[e] The paper reports a range of FP/h values evaluated on the data of different patients, with a minimum of 0.08 FP/h, and a maximum of 0.5 FP/h.

phase, causing a higher false-alarm rate which could be recovered and limited with the use of an artifact detector, the uncertainty of the labeling of the signal has also a significant impact on the training of the classifier. If some of the samples listed for one of the classes belonged to the other, the learning process would also be affected.

At the same time, removing altogether from the training material the windows falling within the uncertainty interval around the onset would aggravate the class imbalance, and, what matters most, remove the earliest stages of the seizure from the learning process, thus possibly compromising the possibility of early detection.

VI. Comparison with the State of the Art

We discuss in this section how the EEGformer compares to the state-of-the-art low-channel count seizure detectors described in the literature. To this end, we consider both the CHB-MIT and our novel datasets. Table VII summarizes the most relevant works presenting epilepsy monitoring based on reduced acquisition setups.

A. CHB-MIT dataset

In the first section of the Table, we list the most relevant works based on the CHB-MIT dataset (see Sect. II). Despite the remarkable sensitivity values achieved with the seizure detectors of [19] and [20] (96.6% and 92.5%, respectively), they also result in a limited specificity (92.5% and 80.1%, respectively). A similar trade-off is presented in the work of [21], where the successful detection of 98.9% of the occurred seizure events was obtained with a specificity value limited to 97.87% and a non-negligible 2.13% FP percentage. On the other hand, despite providing a better balance between a 99.77% sensitivity and 99.88% specificity, the work of [18] relies on a complete acquisition setup (large number of channels) and therefore its performance cannot be directly compared to small-channel-count solutions.

As frequent false alarms would compromise user experience and the practical use of the device [2, 3], we consider as a main reference [22], which reported up to 100% sensitivity and specificity on some of the examined patients, and demonstrated the highest specificity when evaluated on a larger scale, with an average 99.9% specificity and 72% sensitivity, a false alarm rate of only 0.5 FP/h, and an average onset detection latency of 19.2s. Moreover, to enrich the comparison, we also implemented two CNN models specifically designed for the seizure detection task based on low-channel count acquisition setups and optimized on the CHB-MIT dataset (see below). We designed these custom models in order to enable a direct comparison with the performance of the EEGformer, considering the same test and post-processing strategy.

Specifically, we first designed a CNN model (CNN B) to directly process the raw EEG signal, providing performance comparable with the topology proposed in [16] for the classification of the data from patients CHB 1 and CHB 6 of the CHB-MIT dataset (not shown). Unlike the model in [16], our custom CNN works on EEG signal segments of 8s length, in alignment with the test scenario considered for EEGformer. Furthermore, we considered a CNN performing classification on energy features obtained with wavelet decomposition (CNN C), replicating the pre-processing strategy exploited in [22] and providing comparable accuracy in the classification of the data from patient CHB 1. Both models were implemented in PyTorch [30] and trained with the two-step strategy evaluated for the EEGformer, thus benefiting from the global pre-training phase. The training process exploited the Adam optimizer, 5e-5 learning rate, and batch size 16. The first pre-training phase was conducted for 100 epochs, while 50 epochs were exploited

TABLE VIII: CNN detectors considered for the comparison.

Layer	CNN B			CNN C		
	Op	Input	Kernel	Op	Input	Kernel
0	Conv0	4x2048x1	32x5x1	Conv0	4x8x8	16x3x1
1	MaxPool	32x405x1	-	Conv1	16x8x8	32x3x1
2	Conv1	32x203x1	32x5x1	Conv2	32x8x8	64x3x1
3	MaxPool	32x102x1	-	Conv3	64x8x8	64x3x1
4	FC0	32x51x1	200x1x1	Conv4	64x8x8	64x3x1
5	FC1	200x1x1	2x1x1	MaxPool	64x8x8	-
6	-	-	-	FC0	64x4x4	2x1x1

TABLE IX: Performance comparison on CHB-MIT dataset considering acquisition from temporal channels.

Model	Segment-level		Event-level		Average Detection Latency
	Sens	Spec	Sens	FP/h ; FP event [s]	
EEGformer	**65.5**	**99.9**	**73 ; 32/44**	**0.15 ; 38s**	**15.2s**
CNN B	65.4	99.9	73 ; 32/44	0.34 ; 42s	18.2s
CNN C	53.5	99.7	68 ; 30/44	0.53 ; 80s	22.6s
AB ([22])	72	99.9	86 ; 38/44	0.5 ; 4s	19s

for the subject-specific fine-tuning step.

1) CNN on raw EEG signal: This CNN operates on raw EEG signals. The model design inherits some of the take-outs of the exploration conducted for the EEGformer: the input is arranged into windows of 8s length and the first convolutional layer replicates the one successfully exploited in the EEGformer embedding stage for the input dimensionality reduction. The architectural details are described in the left half of Table VIII (CNN B): the model consists of two convolutional layers (Conv#), followed by Rectified Linear Units (ReLU) activation functions, a MaxPooling and two Fully Connected (FC#) layers. Its complexity is lower than the EEGformer's (2.22 MOPs), while its storage requirements are higher (325 KB of parameters with 8-bit representation), although still suitable for our target platforms.

2) CNN on pre-processed input features: Table VIII shows a second CNN model, CNN C, whose input is obtained from the windowed signal with Haar-wavelet decomposition. We will consider it in the following as a reference to define the impact of feature extraction on detection performance. The architecture exploits a sequence of 5 convolutional layers, followed by ReLu activation, and finally a MaxPooling layer and an FC module. This sequence of operations is applied to a 3D tensor, of shape (*channels, height, width*) (C,H,W). Each input item is obtained evaluating the energy of 8 wavelet levels on 8 successive window frames (each of 8s duration), partially overlapped with 1s step size. The computational complexity (12.5 MOPs) is comparable to the EEGformer's and should be considered in addition to the online pre-processing for feature extraction. On the other hand, the storage requirements (105.3 KB) are lower than the first CNN example.

3) Discussion: We provide in Table IX the comparison of the EEGformer performance with the CNN-based detectors described in the previous sections and with the SoA AB model for seizure detection based on low-channel count acquisition [22]. The reported results refer to the same subset of the CHB-MIT dataset and to a comparable testing and post-processing strategy: the output of the AB model is filtered with 3 windows majority voting, while averaging over 5 successive windows was considered for the CNN models, to obtain a higher specificity.

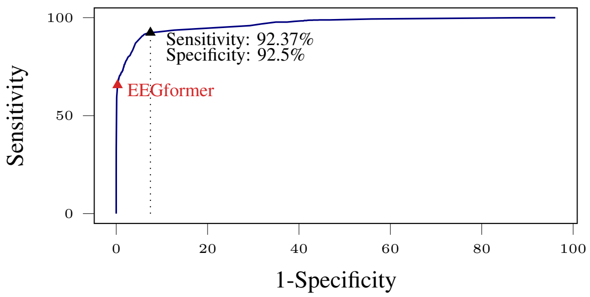


Fig. 4: Receiver Operating Characteristic for the EEGformer, considering classification performance of the set of considered patients on the CHB-MIT dataset. The red point highlights the EEGformer working point designed to minimize the False Positive Rate, whereas we report in black the sensitivity-specificity trade-off at a lower specificity level, i.e. 92.5%, as in [19], bringing to 79% event-level sensitivity, with the detection of all the seizure events in all patients except 6.

No significant performance degradation resulted from the elimination of the feature-extraction step, which is required both by the CNN C detector and by the SoA AB detector. Overall, the EEGformer reaches quality metrics comparable to the state-of-the-art reference for unobtrusive detection restricted to the temporal channels, introducing a 20% reduction in the average onset detection latency. Other works, like [31], report a lower detection latency, however at the price of higher false alarms, thereby making the solution not acceptable for practical settings. Considering the more general scenario reported in Table VII, the low sensitivity value is mostly impacted by the performance obtained on patient CHB 6 (see also column 4 of Table IV), which is not included in the set of tested patients in [21]. Excluding the results obtained on this patient, the event-level sensitivity reaches 94%, thereby approaching the performance of alternative models with lower specificity, providing a reliable detection rate. To complete the assessment of EEGformer detection performance, we provide in Figure 4 the Receiver Operating Characteristic curve, obtained evaluating different thresholds in the range [0,1] to discriminate a seizure prediction from a non-seizure prediction. The plot refers to a cumulative analysis of the segment-level sensitivity and specificity over all the leave-one-record-out tests performed on the eight patients considered from the CHB-MIT dataset. Figure 4 reveals that the results reported in Table VII for the EEGformer correspond to a region of the plot aimed at the minimization of the False Positive Rate (evaluated as $1 - Specificity = FP/(FP + TN)$). Higher sensitivity reaching over 90%, can be obtained at the price of a reduced specificity. However, as this work targets long-monitoring devices, where minimal false alarm rates have to be preserved, a similar trade-off would not encourage the practical use of the device. Due to this reason, we also did not make extensive use of approaches to balance the number of training instances from the two classes, to avoid compromising the specificity in favor of a higher sensitivity.

Improvements are still needed to reach the performance achievable with access to complete acquisition setups (100%

sensitivity and 0.04 FP/h in [12]). Nonetheless, the performance achievable with EEGformer indicates that minimal channel-count systems are a viable solution for monitoring outside of EMUs. The EEGformer is compliant with the main constraints of long-term monitoring highlighted in Section II. It is distinguished by its minimized false alarm rate and detection latency, which guarantee timely alarm responses without compromising on accuracy. Moreover, the unobtrusiveness of the required acquisition setup, along with the compact computational workload, make EEGformer especially suitable for integration into wearable devices for long-term monitoring.

B. Comparison to a tree-based approach on clinical dataset

As previously mentioned, since our dataset represents a very novel resource, there is still no state of the art on it. Hence, to compare EEGformer to other approaches, we consider the AB, which is the current SoA reported for the CHB-MIT dataset [22]. Table VI reports in the last row the performance achieved with AB on our data. When evaluated on a dataset different from CHB-MIT, considered during its development, AB shows lower specificity and higher FP/h, indicating that further optimizations are required. The comparison to EEGformer further confirms that our model holds promise for the successful implementation of epilepsy detection on wearable devices with minimal false positives and fast detection times.

Finally, the bottom section of Table VII compares the performance of EEGformer to other models on proprietary clinical datasets. Firstly, we notice that the achieved performance is generally lower than the one reported on works based on the CHB-MIT dataset. At the same time, despite the numbers being not directly comparable (due to the different datasets considered for the evaluation), the event-level sensitivity achieved by EEGformer is higher than the values reported by [15] and [16] on their respective tasks (where 86% and 79% of the occurred seizure events were detected). In terms of false alarm rates, a comparable performance was obtained with respect to [15] (for which the false alarms vary from 2 FP/day to 13 FP/day based on the target patient) and [16], whereas nearly 2× less false positives are achieved with respect to [17].

VII. Deployment

Finally, we show how EEGformer can be efficiently exploited to provide real-time detection on low-power health monitoring devices, describing its implementation on three different resource-constrained hardware targets. The selection of the platforms considered is oriented to demonstrate that EEGformer is suitable for efficient execution on single-core and multi-core devices, with different levels of available computational resources and technological maturity. We regulate the frequency and voltage to optimize the energy efficiency on all targets. We used the Quantlab software package [32] to perform quantization up to 8-bit precision, thus reducing the memory footprint of the model and enabling efficient byte-level processing with no accuracy drop (not shown). Since the targeted platforms do not support sub-byte arithmetic, more aggressive quantization schemes appear as not beneficial, as they would not enable significant efficiency gains, while they would adversely impact accuracy [33]. To speed up the execution of the model, we use the implementation of Integer Softmax, LayerNorm, and GELU from I-BERT [34], where non-linear operands are replaced with their polynomial or iterative approximations. Hence, we avoid the expensive dequantization and exponential computation normally required to implement those non-linear layers. As reported in Table X, the final memory footprint of the model is 150 kB, considering the buffers required for the storage of the intermediate results.

A. Deployment on Apollo4

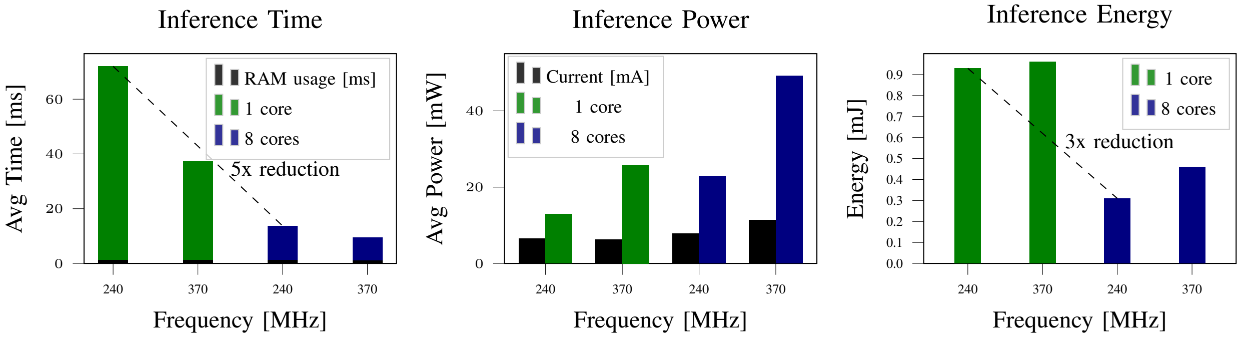
As a first deployment target, we considered the Ambiq Ultra-Low-Power Apollo4 MCU [35]. This power-efficient platform, requiring 5μA/MHz, embeds a 32-bit ARM Cortex-M4 processor accessing a 2MB MRAM and a 1.8 MB SRAM and allows for application-oriented frequency tuning. The technical details about the implementation, based on the CMSIS-NN library [36], are provided in [37]. We report the inference metrics in the first column of Table XI. Having tuned the frequency to 96MHz, we measured 405 ms inference time and 1.79mJ energy consumption, based on the average power consumption measured with the Keysight N6715C analyzer.

B. Deployment on GAP: exploit parallelism

The intrinsic parallel nature of the transformer workload offers great opportunities for efficient inference through parallel computation. The landscape of edge-processing platforms offers heterogeneous computing units, providing parallel and/or hardware-aided computational power. We evaluated the effects of exploiting parallel computations moving to the GAP family of processors by Greenwaves, representing IoT modules optimized for machine learning applications. We first deployed the EEGformer on the GAP8 processor, and finally targeted the more recent GAP9, which demonstrated the best accuracy vs energy efficiency performance in tiny-ML benchmarks [44, 45].

The GAP8 embeds 9 RISC-V parallel processors, one acting as a control processor, and 8 constituting the computing cluster, whose voltage and working frequency can be modulated based on the application requirements. The memory hierarchy includes a 512KB L2 shared memory and a small 64KB L1 memory, local to the computing cluster, with multiple DMAs allowing for autonomous and power-efficient data transfers. It is based on the TSMC 55 nm LP technology, enabling a clock frequency of up to 250 MHz. Leveraging the parallel nature of transformer-based inference, we exploited the increased computational power provided by the 8 parallel cores.

As it is shown in Figure 6, the MHA layer represents the main computational workload. It is parallelized along the $heads$ dimension, according to the implementation described in [37], allowing for an almost linear speedup with the number of cores, as reported in Figure 8. Table XI reports in columns 2 and 3 the inference performance, referring to 65MHz clock frequency, and 1V supply voltage. Even though the GAP8 platform is not more energy efficient than the Apollo4 for



(a) (b) (c)

Fig. 5: Inference performance at different frequencies for single-core and multi-core execution on GAP9.

TABLE X: State of the art for epilepsy monitoring hardware implementations on programmable cores.

	Manzouri et al. [38]	**Heller at al.** [39]	**Burrello et al.** [40]	**Lee et al.** [41]	**Ingolfsson et al.** [22]	**this work**
EEG type	intracranial	intracranial	intracranial	intracranial	surface	surface
Model	RF	CNN	HD	CNN	AB	transformer
Platform	MSP430FR5994	MSP430FR serie	Quentin [42]	custom RISC-V	BioWolf [43]	GAP9
# Channels	1-4	4	4-64	2	4	4
Footprint	>256kB	18.8kB	10.3-17.8kB	5kB[1]	4kB	150kB
Frequency	16MHz	8 MHz	187MHz	1MHz	100MHz	240MHz
Time/Inference	N.R.	0.5s	11ms[1]	12ms	0.88ms[2]	13.7ms
Total power	0.2 - 1.12mW	0.8mW	N.R.	0.1mW	27.92mW	22.9mW
Total energy/Inference	0.2 - 0.8mJ	0.8mJ	0.02-0.29mJ	0.95μJ	24.57μJ	0.31mJ

[1] Estimated based on the reported data.
[2] Numbers including feature extraction, inference execution requires 0.17-0.25mW/channel in [38], and 57μs and 1.3μJ in [22].

TABLE XI: Inference performance on hardware.

	Apollo4	**GAP8**		**GAP9**	
	1 core	*1 core*	*8 cores*	*1 core*	*8 cores*
Frequency	96 MHz	65MHz	65MHz	240MHz	240MHz
Time/inf	405ms	283.9ms	62.2ms	72ms	13.7ms
Total Power	4.4mW	10mW	18.1mW	12.9mW	22.9mW
Energy/inf	1.79mJ	2.9mJ	1.2mJ	0.93mJ	0.31mJ

single core execution, the available parallelism allows for over 3× speedup (limited by the less parallel operands in the network), resulting in an energy consumption per inference 30% lower than the one required for execution on the Apollo4.

The GAP9 processor is fabricated as a more advanced technological node, based on the TSMC 22 nm LP technology and reaching up to 400MHz working frequency. The computing cluster includes an additional supervising core, and the memory hierarchy is based on a 1.6MB L2 memory and a 128KB L1 memory. Columns 4 and 5 of Table XI report the inference performance on GAP9 measured at 240MHz clock frequency and 1.8V supply voltage.

As shown in the plots in Figure 5b and 5c, this configuration represents the most power efficient setup for the platform: the power consumption is reduced by a factor of 2 for parallel execution on 8 cores, resulting in 1.5× energy saving. The parallel execution on the computing cluster allows us to reach an inference time equal to 22% of the one required by GAP8 and results in 82% energy savings over the first Apollo4 implementation. The measurements were performed with a Power Profiler Kit II (PPK2) connected to the GAP9 Evaluation Kit. The measurement setup is shown in Figure 7b.

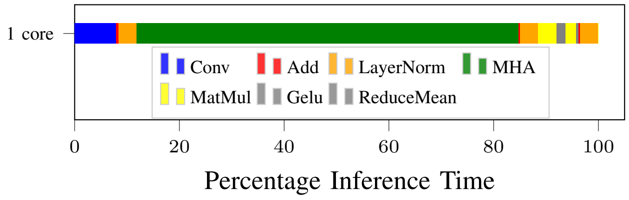
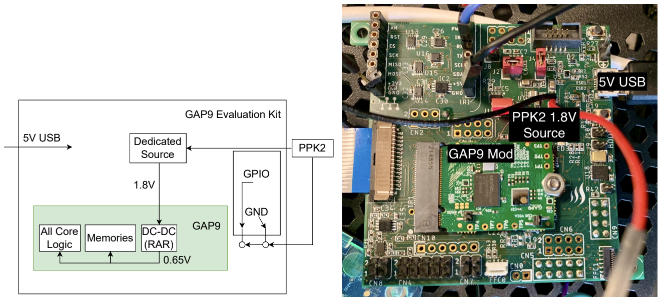


Fig. 6: Percentage inference time distribution for the different operators in EEGformer on GAP9 and 1 core execution.



(a) (b)

Fig. 7: (a) System-level overview of how the PPK2 current supply is connected to the GAP9 Evaluation Kit. (b) Photo of the GAP9 Evaluation Kit with measurement connections. The measurement is performed running the cluster domain of GAP9 at the most energy-efficient point of 240 MHz.

C. Discussion

In the following, we compare the efficiency of our deployment solution with the state of the art. We limit the comparison to programmable solutions, based on MCUs, which are

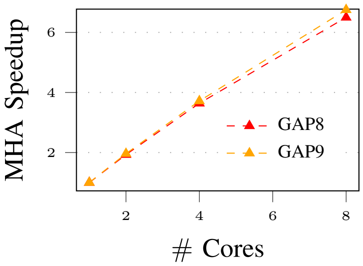


Fig. 8: Parallel speedup of the MHA execution time for execution on the GAP processors.

inherently less efficient, yet more flexible, than highly specialized ASIC systems. Table X reports recent works presenting energy-efficient systems for wearable epilepsy monitoring. The analysis of the results reported in the Table highlights how our proposed implementation is aligned with the performance achieved in the works of [38, 39, 40] in terms of energy consumption per inference, with the work of [40] reporting the best numbers for a reduced acquisition setup.

We note that the systems proposed in [38, 39, 40, 41] work with intracranial EEG, which is inherently less noisy. The detector exploited in [38] is based on a simple RF model, where most of the memory requirements are represented by the need to store the training data for inference execution. On the other hand, [40] and [39] exploit detectors of lower complexity compared to the EEGformer, based on Hyperdimensional computing (HD) and CNN.

Even higher efficiency is achieved in the work of [41], which presents a tiny CNN for the classification of the EEG signal of six rats. This interesting work achieves as low as 0.95μJ per inference, with a custom specialized co-processor, based on RISC-V and exploiting a custom instruction set. Nonetheless, the device embeds a tiny 6kB SRAM memory, which limits the range of possible deployable detectors, and the targeted tasks cannot be directly compared.

The solution presented in reference [22] stands out as the most energy-efficient general-purpose approach in the comparison. It relies on a very lean AB model for detection, which requires minimal inference execution time. In contrast to [22], our solution aims to achieve a lower false positive/healthy rate and faster onset detection latency through a more complex algorithm. Consequently, despite the system's power consumption being over 1.2× lower, the EEGformer model dissipates significantly more energy per inference, exceeding the energy dissipation of reference [22] by one order of magnitude.

Nevertheless, our proposed solution is still executable in real-time and is comparable with some existing ASIC implementations, whose energy consumption per classification ranges from as low as 2.73μJ per classification [46] up to 0.17mJ per classification [47].

VIII. Conclusion

We presented EEGformer, a transformer-based seizure detector designed to enable efficient long-term monitoring of the raw EEG signal with unobtrusive devices, recording only from the temporal channels. We first assessed its detection performance through cross-validation and leave-one-out tests on the open-source CHB-MIT dataset, and compared its performance to state of the art of seizure detectors based on low-channel count acquisition. EEGformer sets a new state-of-the-art 15.2s average onset detection latency for temporal-channels-only detection and detects 73% of the considered seizure events while achieving 0 FP/h in 35/40 of the tests performed (more than 60% of the considered patients do not experience false alarms).

Furthermore, we presented a novel dataset, containing continuous recordings lasting on average 3 days. We evaluated EEGformer on this new realistic benchmark for clinical practice, showing it is able to detect 88% of the annotated seizure events with only 0.45 FP/h.

We finally considered the deployment on three edge-processing platforms, the Apollo4 MCU and two GAP processors. With the first implementation, we show how the memory footprint and the computational complexity of the EEGformer can be accommodated on a tiny MCU, and executed efficiently with 405ms and 1.79mJ per inference, at 96MHz operating frequency. Parallelizing the inference workload on multiple cores on the GAP8 and GAP9 processors allows further benefit from parallel execution, reaching up to 82% energy savings. To conclude, the EEGformer represents an efficient transformer-based detector targeting low-power and low-channel count continuous monitoring systems.

Future work will focus on further improving the event-level sensitivity and detection latency while preserving or further improving specificity. In addition, further studies should include a more general exploration of the transformer topology, considering the effects of exploiting overlapping filtering kernels in the first embedding stage, and especially of including the decoder structure, to consider previous history when elaborating on each EEG window. Finally, EEGformer represents a promising seizure model to be tested on wearable devices for real-time seizure detection in hospital or ambulatory settings.

References

[1] K. Kuhlmann, LevinAU Lehnertz, M. P. Richardson, B. Schelter, H. P. Zaveri, Seizure prediction — ready for a new era, Nature Reviews Neurology 14 (2018) 618 – 630.

[2] C. Baumgartner, J. P. Koren, Seizure detection using scalp-eeg, Epilepsia 59 (2018) 14–22.

[3] E. Bruno, P. F. Viana, M. R. Sperling, M. P. Richardson, Seizure detection at home: Do devices on the market match the needs of people living with epilepsy and their caregivers?, Epilepsia 61 (2020) S11–S24.

[4] P. Busia, A. Cossettini, T. M. Ingolfsson, S. Benatti, A. Burrello, M. Scherer, M. A. Scrugli, P. Meloni, L. Benini, Eegformer: Transformer-based epilepsy detection on raw eeg traces for low-channel-count wearable continuous monitoring devices, in: 2022 IEEE Biomedical Circuits and Systems Conference (BioCAS), 2022, pp. 640–644. doi:10.1109/BioCAS54905.2022.9948637.

[5] N. Pham, T. Dinh, Z. Raghebi, T. Kim, N. Bui, P. Nguyen, H. Truong, F. Banaei-Kashani, A. Halbower, T. Dinh, T. Vu, Wake: A behind-the-ear wearable system for microsleep detection, Association for Computing Machinery (2020) 404–418.

[6] M. Guermandi, S. Benatti, V. J. Kartsch Morinigo, L. Bertini, A wearable device for minimally-invasive behind-the-ear eeg and evoked potentials, 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS) (2018) 1–4.

[7] A. L. Goldberger, et al., Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals, circulation 101 (2000) e215–e220.

[8] A. H. Shoeb, Application of machine learning to epileptic seizure onset detection and treatment, Ph.D. dissertation, MIT (2009).

[9] T. N. Alotaiby, S. A. Alshebeili, A. Tariq, A. Ishtiaq, E. A. E.-S. Fathi, Eeg seizure detection and prediction algorithms: a survey, EURASIP Journal on Advances in Signal Processing (2014).

[10] P. Yash, Various epileptic seizure detection techniques using biomedical signals: a review, Brain Informatics (2018).

[11] J. Prasanna, M. S. P. Subathra, M. Mohammed, R. D. R, N. J. Sairamya, S. George, Automated epileptic seizure detection in pediatric subjects of chb-mit eeg database-a survey., J Pers Med. (2021).

[12] M. Sahani, S. K. Rout, P. K. Dash, Epileptic seizure recognition using reduced deep convolutional stack autoencoder and improved kernel rvfln from eeg signals, IEEE Transactions on Biomedical Circuits and Systems 15 (2021) 595–605.

[13] D. Sopic, A. Aminifar, D. Atienza, e-glass: A wearable system for real-time detection of epileptic seizures, 2018 IEEE International Symposium on Circuits and Systems (ISCAS) (2018) 1–5.

[14] A. Van de Vel, K. Smets, K. Wouters, B. Ceulemans, Automated non-eeg based seizure detection: Do users have a say?, Epilepsy & Behavior 62 (2016) 121–128.

[15] L. S. Remvig, J. Duun-Henriksen, F. Fürbass, M. Hartmann, P. F. Viana, A. M. Kappel Overby, S. Weisdorf, M. P. Richardson, S. Beniczky, T. W. Kjaer, Detecting temporal lobe seizures in ultra long-term subcutaneous eeg using algorithm-based data reduction, Clinical Neurophysiology 142 (2022) 86–93.

[16] G. Japaridze, D. Loeckx, T. Buckinx, S. Armand Larsen, R. Proost, K. Jansen, P. MacMullin, N. Paiva, S. Kasradze, A. Rotenberg, L. Lagae, S. Beniczky, Automated detection of absence seizures using a wearable electroencephalographic device: a phase 3 validation study and feasibility of automated behavioral testing, Epilepsia (2022).

[17] K. Vandecasteele, T. De Cooman, J. Dan, E. Cleeren, S. Van Huffel, B. Hunyadi, W. Van Paesschen, Visual seizure annotation and automated seizure detection using behind-the-ear electroencephalographic channels, Epilepsia 61 (2020) 766–775.

[18] J. Zeng, X. dan Tan, C. A. Zhan, Automatic detection of epileptic seizure events using the time-frequency features and machine learning, Biomedical Signal Processing and Control 69 (2021) 102916.

[19] R. Zanetti, A. Aminifar, D. Atienza, Robust epileptic seizure detection on wearable systems with reduced false-alarm rate, in: 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2020, pp. 4248–4251. doi:10.1109/EMBC44109.2020.9175339.

[20] T. Zhan, S. Z. Fatmi, S. Guraya, H. Kassiri, A resource-optimized vlsi implementation of a patient-specific seizure detection algorithm on a custom-made 2.2 cm^2 wireless device for ambulatory epilepsy diagnostics, IEEE Transactions on Biomedical Circuits and Systems 13 (2019) 1175–1185.

[21] M. Shen, P. Wen, B. Song, Y. Li, Real-time epilepsy seizure detection based on eeg using tunable-q wavelet transform and convolutional neural network, Biomedical Signal Processing and Control 82 (2023) 104566.

[22] T. M. Ingolfsson, A. Cossettini, X. Wang, E. Tabanelli, G. Tagliavini, P. Ryvlin, L. Benini, S. Benatti, Towards long-term non-invasive monitoring for epilepsy via wearable eeg devices, BioCAS 2021 - IEEE Biomedical Circuits and Systems Conference, Proceedings (2021).

[23] J. Yan, J. Li, H. Xu, Y. Yu, T. Xu, Seizure prediction based on transformer using scalp electroencephalogram, Applied Sciences 12 (2022).

[24] A. Bhattacharya, T. Baweja, S. P. K. Karri, Epileptic seizure prediction using deep transformer model, International Journal of Neural Systems 32 (2022).

[25] Y. Ma, C. Liu, M. S. Ma, Y. Yang, N. D. Truong, K. Kothur, A. Nikpour, O. Kavehei, Tsd: Transformers for seizure detection, bioRxiv (2023).

[26] D. Alexey, B. Lucas, K. Alexander, W. Dirk, Z. Xiaohua, U. Thomas, D. Mostafa, M. Matthias, H. Georg, G. Sylvain, U. Jakob, H. Neil, An image is worth 16x16 words: Transformers for image recognition at scale, International Conference on Learning Representations (2021).

[27] A. Burrello, F. B. Morghet, M. Scherer, S. Benatti, L. Benini, E. Macii, M. Poncino, D. J. Pagliari, Bioformers: Embedding transformers for ultra-low power semg-based gesture recognition, IEEE 2022 DATE (2022).

[28] A. Vaswani, et al., Attention is all you need, Advances in Neural Information Processing Systems 2017-December (2017) 5999 – 6009.

[29] TensorFlow, Keras: the high-level api for tensorflow, 2023. URL: https://www.tensorflow.org/guide/keras?hl=en.

[30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[31] M. A. Bin Altaf, J. Yoo, A 1.83 μj/classification, 8-channel, patient-specific epileptic seizure classification soc using a non-linear support vector machine, IEEE Transactions on Biomedical Circuits and Systems 10 (2016) 49–60.

[32] M. Spallanzani, G. Rutishauser, M. Scherer, A. Burrello, F. Conti, L. Benini, QuantLab: a Modular Framework for Training and Deploying Mixed-Precision NNs, https://cms.tinyml.org/wp-content/uploads/talks2022/Spallanzani-Matteo-Hardware.pdf, 2022.

[33] G. Rutishauser, F. Conti, L. Benini, Free bits: Latency optimization of mixed-precision quantized neural networks on the edge, in: 2023 IEEE 5th International Conference on Artificial Intelligence Circuits and Systems (AICAS), 2023, pp. 1–5. doi:10.1109/AICAS57966.2023.10168577.

[34] S. Kim, A. Gholami, Z. Yao, M. W. Mahoney, K. Keutzer, I-bert: Integer-only bert quantization, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 5506–5518. URL: https://proceedings.mlr.press/v139/kim21d.html.

[35] Ambiq, Apollo4 ambiq, 2022. URL: https://ambiq.com/apollo4/.

[36] L. Lai, N. Suda, V. Chandra, Cmsis-nn: Efficient neural network kernels for arm cortex-m cpus, CoRR abs/1908.09791 (2019).

[37] A. Burrello, M. Scherer, M. Zanghieri, F. Conti, L. Benini, A microcontroller is all you need: Enabling transformer execution on low-power iot endnodes, 2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS) (2021) 1–6.

[38] F. Manzouri, S. Heller, M. Dümpelmann, P. Woias, A. Schulze-Bonhage, A comparison of machine learning classifiers for energy-efficient implementation of seizure detection, Frontiers in Systems Neuroscience 12 (2018).

[39] S. Heller, M. Hügle, I. Nematollahi, F. Manzouri, M. Dümpelmann, A. Schulze-Bonhage, J. Boedecker, P. Woias, Hardware implementation of a performance and energy-optimized convolutional neural network for seizure detection, in: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018, pp. 2268–2271. doi:10.1109/EMBC.2018.8512735.

[40] A. Burrello, S. Benatti, K. Schindler, L. Benini, 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 25 (2021) 935–946.

[41] S.-Y. Lee, Y.-W. Hung, Y.-T. Chang, C.-C. Lin, G.-S. Shieh, Risc-v cnn coprocessor for real-time epilepsy detection in wearable application, IEEE Transactions on Biomedical Circuits and Systems 15 (2021) 679–691.

[42] P. D. Schiavone, D. Rossi, A. Pullini, A. Di Mauro, F. Conti, L. Benini, Quentin: an ultra-low-power pulpissimo soc in 22nm fdx, in: 2018 IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), 2018, pp. 1–3. doi:10.1109/S3S.2018.8640145.

[43] V. Kartsch, G. Tagliavini, M. Guermandi, S. Benatti, D. Rossi, L. Benini, Biowolf: A sub-10-mw 8-channel advanced brain–computer interface platform with a nine-core processor and ble connectivity, IEEE Transactions on Biomedical Circuits and Systems 13 (2019) 893–906.

[44] Greenwaves, Ultra low power gap processors, 2022. URL: https://greenwaves-technologies.com/low-power-processor/.

[45] ML Commons, Inference: tiny. v1.0 Results, 2022. URL: https://mlcommons.org/en/inference-tiny-10/, Accessed: 15-11-2022.

[46] M. A. Bin Altaf, C. Zhang, J. Yoo, A 16-channel patient-specific seizure onset and termination detection soc with impedance-adaptive transcranial electrical stimulator, IEEE Journal of Solid-State Circuits 50 (2015) 2728–2740.

[47] G. O’Leary, D. M. Groppe, T. A. Valiante, N. Verma, R. Genov, Nurip: Neural interface processor for brain-state classification and programmable-waveform neurostimulation, IEEE Journal of Solid-State Circuits 53 (2018) 3150–3162.