EEGformer: Transformer-Based Epilepsy Detection on Raw EEG Traces for Low-Channel-Count Wearable Continuous Monitoring Devices

EEGformer: Transformer-Based Epilepsy Detection on Raw EEG Traces for Low-Channel-Count Wearable Continuous Monitoring Devices

Paola Busia*†, Andrea Cossettini†, Thorir Mar Ingolfsson†, Simone Benatti‡§, Alessio Burrello‡, Moritz Scherer†, Matteo Antonio Scrugli*, Paolo Meloni*, Luca Benini†‡

*DIEE, University of Cagliari, Cagliari, Italy
†Integrated Systems Laboratory, ETH Zürich, Zürich, Switzerland
‡DEI, University of Bologna, Bologna, Italy
§DISMI, University of Modena and Reggio Emilia, Reggio Emilia, Italy

***Abstract*—The development of a device for long-term and continuous monitoring of epilepsy is a very challenging objective, due to the high accuracy standards and nearly zero false alarms required by clinical practices. To comply with such requirements, most of the approaches in the literature rely on a high number of acquisition channels and exploit classifiers operating on pre-processed features, hand-crafted considering the available data, currently fairly limited. Thus, they lack comfort, portability, and adaptability to future use-cases and datasets. A step forward is needed towards the implementation of unobtrusive, wearable systems, with a reduced number of channels, implementable on ultra-low-power computing platforms. Leveraging the promising ability of transformers in capturing long-term raw data dependencies in time series, we present in this work EEGformer, a compact transformer model for more adaptable seizure detection, that can be executed in real-time on tiny MicroController Units (MCUs) and operates on just the raw electroencephalography (EEG) signal acquired by the 4 temporal channels. Our proposed model is able to detect 73% of the examined seizure events (100% when considering 6 out of 8 patients), with an average onset detection latency of 15.2s. The False Positive/hour (FP/h) rate is equal to 0.8 FP/h, although 100% specificity is obtained in most tests, with 5/40 outliers that are mostly caused by EEG artifacts. We deployed our model on the Ambiq Apollo4 MCU platform, where inference run requires 405 ms and 1.79 mJ at 96 MHz operating frequency, demonstrating the feasibility of epilepsy detection on raw EEG traces for low-power wearable systems. Considering the CHB-MIT Scalp EEG dataset as a reference, we compare with a state-of-the-art classifier, acting on hand-crafted features designed on the target dataset, reaching well-aligned accuracy results and reducing the onset detection latency by over 20%. Moreover, we compare with two adequately optimized Convolutional Neural Networks-based approaches, outperforming both alternatives on all the accuracy metrics.**

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

I. Introduction

Epilepsy is a chronic neurological disease affecting more than 50 million people in the world [1]. It manifests itself as recurring seizures, interfering with the normal brain electrical activity and potentially leading to loss of movement control, or even loss of consciousness. Antiepileptic drugs are the most common treatment, however alternative options for drug-resistant cases include brain surgery, and implantable neurostimulators [2]. As seizure occurrence significantly degrades the quality of life, electroencephalography (EEG) monitoring is a topic of great interest. Furthermore, since long-term monitoring can only be achieved with compact, wearable devices, there is a need for solutions based on a reduced number of EEG channels for integration in low-power and unobtrusive embedded systems, to avoid stigma [3], [4].

Ref. [5] and the more recent [6] and [7] present an in-depth survey of the approaches to EEG processing for seizure detection explored in the literature. The approaches exploited in the last few years were multi-faceted, analyzing the EEG signal in the time [8], frequency [9], and wavelet [10] domains, or by empirical mode decomposition (EMD) [11]. Extensive research has also been done on the classifier design, either exploiting traditional machine learning methods like Support Vector Machines (SVMs), and Random Forests (RF) [3], [12], or relying on Convolutional Neural Networks (CNNs) [13]. Most recently, the use of transformers for the EEG signal processing and classification has been explored [14]–[16], targeting also the epilepsy monitoring field [17]–[19]. However, the proposed transformer models in [18], [19] work on pre-processed features extracted through short-time Fourier transforms, and their memory footprint and the number of operations (OPs) are not affordable for embedded deployment. Although the work of [17] eliminates the feature extraction step, it does not explicitly target embedded deployment, and a clear indication of the model's memory footprint and required OPs is missing. Furthermore, all referenced works rely on a complete 18/19 electrode set for data acquisition, which is not compatible with an unobtrusive wearable setup.

In this work, we take inspiration from the main findings of [3], where the possibility of non-invasive patient monitoring through a reduced implant containing only the temporal electrodes is verified, as well as the need for a subject-specific training approach, providing up to 100% sensitivity and specificity within a memory footprint suitable for a microcontroller,

This project was supported by the Swiss National Science Foundation (Project PEDESITE) under grant agreement 193813.

thus embeddable in a device causing minimal discomfort to the patient [20], [21]. The classification required hand-crafted feature extraction, namely the energy of the signal after 4-level wavelet decomposition, and achieved an average onset detection latency of 19s, whose improvement would allow a more prompt alert to the patient/caregiver. Moving from such premises, this paper provides the following contributions:

- the definition of EEGformer, a small-scale (50.6K parameters and 14.7 MOPs) transformer model for online epilepsy monitoring, targeting the wearable domain and performing seizure detection on just four raw EEG input channels, without handcrafted feature extractors, achieving results comparable with state-of-art (SOA) 4-channel feature-based classifiers with a 20% lower detection latency;
- the evaluation of an effective training strategy, demonstrating a specificity improvement up to 100% thanks to global pre-training and subject-specific fine-tuning;
- comparison of our EEGformer with two CNN models, representative of alternative deep neural network-based (DNN) detectors for low channel count;
- the implementation on a MicroController Unit (MCU), where inference execution requires 405ms and 1.79mJ, showing its feasibility for real-time monitoring.

II. Classification Model

We address the seizure detection task as a binary classification problem between a non-seizure and a seizure class. We describe in the following the architecture of our proposed EEGformer. To enable a comparison with alternative DNN approaches, we also describe two CNN-based architectures we considered as possible alternatives. To the best of our knowledge, there is no literature exploiting CNNs to perform seizure detection from a reduced number of input channels. We thus performed a design exploration to optimize their architectures, considering an implementation working on raw EEG signal and one working on features extracted through wavelet-based processing. Since we target wearable monitoring devices, we only consider data acquisition from the temporal channels: F7-T7, T7-P7, F8-T8, T8-P8, according to the 10-20 international system. We chose these channels because they are well-suited for embedding in common long-term wearable devices such as over-the-ear headphones or eyeglasses, thus avoiding stigma [4], [20].

A. EEGformer.

The EEGformer architecture is inspired by the Vision Transformer (ViT) [22], whose key feature is the attention mechanism [23], allowing us to evaluate the mutual relation between any pair of points in a time series. The attention layer evaluates three different projections of the input, called *query* q, *key* k and *value* v, and computes the attention scores as:

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

where d is the size of each projection. Different projections of the input can be examined with Multi-Head-Attention (MHA),

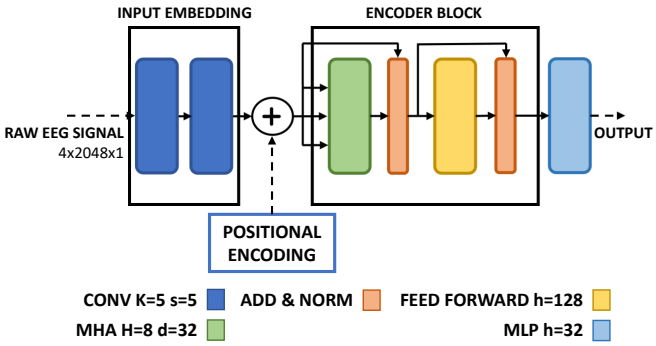


Fig. 1: EEGformer architecture.

where each head represents a parallel execution of the attention mechanism, and the output is finally linearly projected to the original input dimension. The core of the ViT architecture is represented by the encoder structure, whose simplest block consists of an attention layer, followed by a feed-forward network.

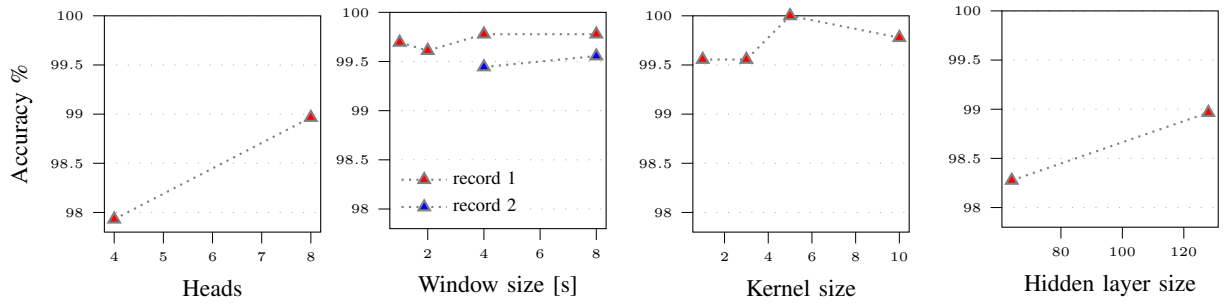
Figure 1 shows the architecture considered in this work, which was selected through an exploration process exploiting the BioFormer structure [24] as a starting point, and further optimized for the EEG seizure detection task. Figure 2 summarizes the outcome of the exploration, aiming at accuracy optimization with a higher focus on specificity. The selected design embeds $N = 1$ encoder blocks, with $H = 8$ heads in the MHA layer (Figure 2a). Based on the exploration outcome, it operates on signal windows of 8s (Figure 2b), processed through an embedding layer including two 1D convolutional layers, with kernel size $K = 5$ and stride $s = 5$ (Figure 2c), reducing the size of the input to the encoder layer. The size of keys, queries, and values was selected to be $d = 32$, whereas the hidden layer size of the encoder feed-forward network is $h = 128$ (Figure 2d). The token provided to the Multi-Layer-Perceptron (MLP) for classification is obtained as the mean across the sequence of tokens produced by the encoder block. The examined architecture exploits 50.6 K parameters and 14.7 MOPs.

B. CNN on raw EEG signal.

As an example of CNN working on raw EEG signal, we considered the model described in I, where all the convolutional layers (Conv#) are followed by Rectified Linear Unit (ReLU) activation functions, and Fully Connected layers are reported as FC#. It was obtained through an analogous exploration process, and exploits some architectural features which resulted successful for the EEGformer: it works on 8s input windows and the first feature-extracting layer reproduces the one exploited in the input embedding layer. The network processes raw EEG input signal, exploiting 325 KB of parameters with 8-bit representation, and 2.22 MOPs.

C. CNN on preprocessed input features.

Table II summarizes the parameters of the CNN model considered to assess the impact of pre-calculating input features.



(a) # heads: 4, **8** (b) window size: 1s, 2s, 4s, **8s** (c) kernel size: 1, 3, **5**, 10 (d) hidden size: 64, **128**

Fig. 2: EEGformer architecture parameter exploration. The selected solution is highlighted in bold: tests on two records were considered to select between 4s and 8s window size.

As reported, the input to the network is a 3D tensor of shape (*channels, height, width*) (C,H,W), where each item of size (H,W) is obtained as 8 columns, representing the energy of 8 wavelet levels on different temporal frames. The wavelet is evaluated on 8s wide windows, overlapped with a 1s step size. The window size, the number of wavelet levels H, and stacked temporal frames W were subject to exploration. The convolutional layers are followed by ReLU activation. The memory footprint of the model, with 8-bit representation, is 105.3 KB, whereas the number of OPs required is 12.5 MOPs.

TABLE I: CNN architecture with raw signal input

Layer	Input Features	Output Features	Input Size	Kernel Size	Stride
Conv1	4	32	1x2048	1x5	5
Maxpool	32	32	1x405	1x2	2
Conv2	32	32	1x203	1x5	2
Maxpool	32	32	1x102	1x2	2
FC1	1632	200	1x1		
FC2	200	2	1x1		

TABLE II: CNN architecture with pre-processed input

Layer	Input Features	Output Features	Input Size	Kernel Size	Stride
Conv1	4	16	8x8	3x3	1
Conv2	16	32	8x8	3x3	1
Conv3	32	64	8x8	3x3	1
Conv4	64	64	8x8	3x3	1
Conv5	64	64	8x8	3x3	1
Maxpool	64	64	8x8	2x2	2
FC1	1024	2	1x1		

III. Assessment on CHB-MIT dataset

In this section, we evaluate the classification performance of the DNN models described in Section II on the CHB-MIT Scalp EEG dataset for seizure detection [25], [26].

TABLE III: EEGformer training strategy evaluation

	Evaluation Period	Sensitivity	Specificity	FP/h	Detected Seizures
w/o Pre-Training	2s	95.4	99.9	0.9	7/7
Pre-Training	2s	86.6	100	0	7/7

Training strategy. We compared a plain subject-specific training with a two-step training, consisting of a 100 epochs subject-independent pre-training and 50 epochs of subject-specific fine-tuning. A leave-one-out strategy is exploited for the definition of the test set, whereas 20% of the available training data is exploited as a validation set. Non-overlapping windows of signal are used as training data, whereas during the test phase we consider a sliding windowing of the input signal, with an evaluation period of 2s. Table III shows the performance of the EEGformer with both training approaches, with and without (w/o) pre-training, targeting the seizure detection of CHB patient 1. In this phase, the pre-training pool contains all the available data from 7 patients, subjects 2 to 8. The training strategy exploration shows that the specificity of the EEGformer, as well as the False Positive/hour (FP/h) rate, benefit from the pre-training phase. Despite the drop in sensitivity observed, 100% of the tested seizure events were detected. Furthermore, the approach with pre-training appears more suitable for the actual deployment in clinical settings, where patients and caregivers require a zero false alarm rate.

Classifier selection. We proceeded at this point with the assessment of the considered classifiers. We referenced a significant subset of the CHB-MIT dataset, consisting of subjects 1 to 8.The pre-training phase was also exploited to improve the performance of the CNN alternatives considered. The pre-training pool containing all patients, except for the one that is being tested. We repeated the leave-one-out test until all of the available seizure records were evaluated. Table IV summarizes the performance of the examined classifiers, based on the cumulative evaluations of the tests performed, considering also the direct comparison with the SOA AdaBoost (AB) model operating on the temporal channels. The reported results refer to post-processed outputs, obtained with moving averages between successive windows. We consider 3 windows for the EEGformer and for the model in [3], whereas for the CNN on the raw input signal (CNN B) and the CNN on pre-processed input features (CNN C) 5 output windows are averaged. The detection latency is defined based on the first window correctly classified as seizure, and accounting for the averaging delay. Furthermore, as the EEG signal remains unstable for several minutes after the occurrence of a seizure, we neglect in the reported results the FPs occurring within 15 minutes after the seizure ended. All the classifiers resulting from our exploration reached over 99% window-level test

accuracy, allowing us to detect up to 32 of the 44 seizures in the test set. The EEGformer provides the shortest detection latency and is well-aligned with the SOA, especially if the possibility of an artifact-removal stage is considered, demonstrating that acceptable performance can be obtained without relying on handcrafted feature extractors. Other works [27], [28] obtained a further reduction of the detection latency, but this advancement results in a lower specificity, which is crucial for the practical use of the device.

For the sake of completeness, a more general summary of the unconstrained SOA is reported in Table V, where the size of the models in [29], [30] is not explicitly reported and KNN stands for K-nearest-neighbor. Despite the difficulties in comparing results obtained with different testing and post-processing strategies, we see that indeed low-channel count detection is more challenging, but EEGformer is not too far even from unconstrained, full-channel count detectors.

FP analysis. As it is highlighted in the boxplot in Figure 3, representing the distribution of the FP/h on the 40 records tested, the average FP rate for the EEGformer is 0 FP/h, except for 5/40 outliers. Based on visual analysis, 2/5 are due to signal artifacts mistakenly detected as seizures. Removing those two records from the pre-training and training sets, the number of tests resulting in FP is further reduced by one, revealing the importance of an artifact removal stage before the epilepsy detection [31]. Finally, as summarized in Table VI, the proposed model allows the detection of 100% of the seizures reported for 6/8 patients, whereas the average sensitivity value drops dramatically when considering the data from patient 6, which proved to be highly challenging with all of the classifiers listed in Table IV due to the short duration of the seizure events. We consider it as an indication that, although our proposed architecture and training strategy perform generally well on multiple patients, in some cases a more subject-specific approach should be adopted.

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

	Evaluation Period	Sensitivity; Specificity	FP/h;	Average Latency	Detected Seizures
EEGformer	2s	65.5; 99.9	0.8	15.2s	32/44
EEGformer*[1]	2s	66; 99.9	0.12	15.2s	32/42
CNN B	2s	65.3; 99.9	2.8	18.2s	32/44
CNN C	2s	53.5; 99.7	8.2	22.6s	30/44
AB [3]	4s-8s	72; 99.9	0.5	19s	38/44

TABLE V: SOA on the CHB-MIT dataset.

	Sensitivity; Specificity	FP/h	# acquisition channels	needs pre-processing	# params
EEGformer	**65.5; 99.9**	**0.8**	**4**	✗	**50.6 K**
EEGformer*[1]	66; 99.9	0.12	4	✗	50.6 K
AB [3]	72; 99.9	0.5	4	✓	4 K
SVM [30]	97.34; 97.5	0.63	18-23	✓	-
KNN [29]	98.4; 99.1	-	18-23	✓	-
CNN [32]	88.14; 99.62	0.2	18-23	✗	105 K

IV. Deployment

To validate the EEGformer for real-life on-edge performance, we considered its deployment on the Ambiq Ultra-Low-Power Apollo4 MCU [33]. It exploits a 32-bit ARM Cortex-M4 processor, working at up to 192 MHz clock frequency, accessing a 2MB MRAM and a 1.8 MB SRAM. Power-efficiency ensuring as low as 5μA/MHz makes this platform particularly suitable for battery-powered health monitoring devices. To reduce the memory footprint of the model we performed quantization to 8-bit precision, exploiting the Quantlab software package for quantization aware fine-tuning [34]. Finally, for efficient deployment, we exploited the CMSIS-NN library, specifically optimized for ARM Cortex-M processors [35], as it is described in [36]. As reported in Table VII, inference execution requires 405 ms and 1.79mJ, based on the average power consumption measured with the Keysight N6715C DC power analyzer, with a clock frequency of 96 MHz, sufficient to run 1 inference/s.

[1]Results excluding the two records with the recognized artifacts.

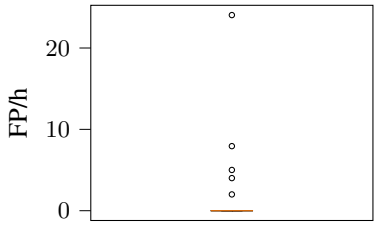


Fig. 3: Record-wise FP/h with the EEGformer.

TABLE VI: Percentage of detected seizures per patient.

Patient	1	2	3	4	5	6	7	8
Detected %	100	100	100	50	100	0	100	100

TABLE VII: Inference execution on Apollo4.

	EEGformer
Time/inference	405 ms
Power	4.4 mW
Energy/inference	1.79 mJ

V. Conclusions

We presented a transformer model for non-invasive epilepsy monitoring, performing seizure detection on raw EEG signals collected from temporal electrodes. We explored the training strategy, showing the advantages of a two-step approach, with a global pre-training phase and a subject-specific fine-tuning. The EEGformer reaches a performance comparable with the SOA, exhibiting a 65.5% sensitivity, with a new SOA 15.2s average onset detection latency. Excluding 5/40 outliers, most of which are caused by EEG artifacts, 0 FP can be achieved. The EEGformer is suitable for deployment on unobtrusive devices: inference run on the Apollo4 MCU, requires 405ms and 1.79mJ at 96MHz operating frequency. This work demonstrates the feasibility of employing transformers on raw EEG traces for seizure detection with reduced latency, in the context of resource-constrained wearable devices for long-term continuous monitoring at low power consumption. The results obtained on this limited exploration pave the way for the exploitation of their robustness and flexibility in future, more complex, and variable data environments.

REFERENCES

[1] W. H. Organization, "Epilepsy." [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/epilepsy

[2] "Epilepsy: Treatment options." *American family physician*, pp. 87–96, July 2017.

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

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

[5] T. N. Alotaiby, S. A. Alshebeili, A. Tariq, A. Ishtiaq, and E. A. E.-S. Fathi, "Eeg seizure detection and prediction algorithms: a survey," *EURASIP Journal on Advances in Signal Processing*, 2014. [Online]. Available: https://doi.org/10.1186/1687-6180-2014-183

[6] P. Yash, "Various epileptic seizure detection techniques using biomedical signals: a review," *Brain Informatics*, 2018. [Online]. Available: https://doi.org/10.1186/s40708-018-0084-z

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

[8] T. Runarsson and S. Sigurdsson, "On-line detection of patient specific neonatal seizures using support vector machines and half-wave attribute histograms," *International Conference on Computational Intelligence for Modelling, Control and Automation and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'06)*, vol. 2, pp. 673–677, 2005.

[9] A. Van Esbroeck, L. Smith, Z. Syed, S. Singh, and Z. Karam, "Multi-task seizure detection: addressing intra-patient variation in seizure morphologies," *Machine Learning*, vol. 102, no. 3, pp. 309–321, 2016. [Online]. Available: https://doi.org/10.1007/s10994-015-5519-7

[10] C. Guangyi, X. Wenfang, B. T. D., and K. Adam, "Automatic epileptic seizure detection in eeg using nonsubsampled wavelet–fourier features," *Journal of Medical and Biological Engineering*, pp. 123–131, 2017. [Online]. Available: https://doi.org/10.1007/s40846-016-0214-0

[11] V. Desai, "Eeg signal classification into seizure and non-seizure class using empirical mode decomposition and artificial neural network," *"Imperial Journal of Interdisciplinary Research (IJIR)*, vol. 3, 01 2017.

[12] S. Pattnaik, N. Rout, and S. Sabut, "Machine learning approach for epileptic seizure detection using the tunable-q wavelet transform based time–frequency features." *International Journal of Information Technology*, 2022. [Online]. Available: https://doi.org/10.1007/s41870-022-00877-1

[13] K. Singh and J. Malhotra, "Smart neurocare approach for detection of epileptic seizures using deep learning based temporal analysis of eeg patterns." *Multimedia Tools and Applications*, 2022. [Online]. Available: https://doi.org/10.1007/s11042-022-12512-z

[14] J. Sun, J. Xie, and H. Zhou, "Eeg classification with transformer-based models," *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*, pp. 92–93, 2021.

[15] Z. Wang, Y. Wang, C. Hu, Z. Yin, and Y. Song, "Transformers for eeg-based emotion recognition: A hierarchical spatial information learning model," *IEEE Sensors Journal*, vol. 22, no. 5, pp. 4359–4368, 2022.

[16] A. Arjun, A. S. Rajpoot, and M. Raveendranatha Panicker, "Introducing attention mechanism for eeg signals: Emotion recognition with vision transformers," *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pp. 5723–5726, 2021.

[17] J. Pedoeem, S. Abittan, G. B. Yosef, and S. Keene, "Tabs: Transformer based seizure detection," *2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–6, 2020.

[18] A. Bhattacharya, T. Baweja, and S. P. K. Karri, "Epileptic seizure prediction using deep transformer model," *International Journal of Neural Systems*, vol. 32, no. 02, 2022.

[19] J. Yan, J. Li, H. Xu, Y. Yu, and T. Xu, "Seizure prediction based on transformer using scalp electroencephalogram," *Applied Sciences*, vol. 12, no. 9, 2022. [Online]. Available: https://www.mdpi.com/2076-3417/12/9/4158

[20] N. Pham, T. Dinh, Z. Raghebi, T. Kim, N. Bui, P. Nguyen, H. Truong, F. Banaei-Kashani, A. Halbower, T. Dinh, and T. Vu, "Wake: A behind-the-ear wearable system for microsleep detection," *Association for Computing Machinery*, p. 404–418, 2020. [Online]. Available: https://doi.org/10.1145/3386901.3389032

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

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

[23] A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 2017-December, p. 5999 – 6009, 2017.

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

[25] A. H. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment," *Ph.D. dissertation, MIT*, 2009.

[26] A. L. Goldberger *et al.*, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," *circulation*, vol. 101, pp. e215–e220, 2000.

[27] M. A. Bin Altaf and 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*, vol. 10, no. 1, pp. 49–60, 2016.

[28] M. A. Bin Altaf, C. Zhang, and 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*, vol. 50, no. 11, pp. 2728–2740, 2015.

[29] M. Li, X. Sun, and W. Chen, "Patient-specific seizure detection method using nonlinear mode decomposition for long-term eeg signals," *Medical & Biological Engineering & Computing*, vol. 58, 12 2020. [Online]. Available: https://doi.org/10.1007/s11517-020-02279-6

[30] C. Li, W. Zhou, G. Liu, Y. Zhang, M. Geng, Z. Liu, S. Wang, and W. Shang, "Seizure onset detection using empirical mode decomposition and common spatial pattern," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 458–467, 2021.

[31] T. M. Ingolfsson, A. Cossettini, S. Benatti, and L. Benini, "Energy-efficient tree-based eeg artifact detection," 2022. [Online]. Available: https://arxiv.org/abs/2204.09577

[32] X. Wang, X. Wang, W. Liu, Z. Chang, T. Kärkkäinen, and F. Cong, "One dimensional convolutional neural networks for seizure onset detection using long-term scalp and intracranial eeg," *Neurocomputing*, vol. 459, pp. 212–222, 2021.

[33] "Apollo4 ambiq." [Online]. Available: https://ambiq.com/apollo4/

[34] M. Spallanzani, G. Rutishauser, M. Scherer, A. Burrello, F. Conti, and 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, March 2022.

[35] L. Lai, N. Suda, and V. Chandra, "Cmsis-nn: Efficient neural network kernels for arm cortex-m cpus," *CoRR*, vol. abs/1908.09791, 2019. [Online]. Available: https://arxiv.org/abs/1908.09791

[36] A. Burrello, M. Scherer, M. Zanghieri, F. Conti, and 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)*, pp. 1–6, 2021.