POLITECNICO DI TORINO
Repository ISTITUZIONALE

BISeizuRe: BERT-Inspired Seizure Data Representation to Improve Epilepsy Monitoring

Original
BISeizuRe: BERT-Inspired Seizure Data Representation to Improve Epilepsy Monitoring / Benfenati, L., Mar Ingolfsson, T., Cossettini, A., Jahier Pagliari, D., Burrello, A., Benini, L.. - ELETTRONICO. - (2024). (2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) Orlando (USA) 15-19 July 2024) [10.1109/EMBC53108.2024.10782289].

Availability:
This version is available at: 11583/2992625 since: 2025-04-10T08:37:46Z

Publisher:
IEEE

Published
DOI:10.1109/EMBC53108.2024.10782289

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2024 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

01 November 2026

This paper has been accepted for publication in the EMBC 2024 conference
©2024 IEEE.

BISeizuRe: BERT-Inspired Seizure Data Representation to Improve Epilepsy Monitoring

Luca Benfenati[‡], Thorir Mar Ingolfsson[*], Andrea Cossettini[*],
Daniele Jahier Pagliari[‡], Alessio Burrello[†‡], Luca Benini[*†]
[*]Integrated Systems Laboratory, ETH Zürich, Zürich, Switzerland, [†]DEI, University of Bologna, Bologna, Italy
[‡]DAUIN, Politecnico di Torino, Italy

arXiv:2406.19189v1 [cs.LG] 27 Jun 2024

***Abstract*—This study presents a novel approach for EEG-based seizure detection leveraging a BERT-based model. The model, BENDR, undergoes a two-phase training process. Initially, it is pre-trained on the extensive Temple University Hospital EEG Corpus (TUEG), a 1.5 TB dataset comprising over 10,000 subjects, to extract common EEG data patterns. Subsequently, the model is fine-tuned on the CHB-MIT Scalp EEG Database, consisting of 664 EEG recordings from 24 pediatric patients, of which 198 contain seizure events. Key contributions include optimizing fine-tuning on the CHB-MIT dataset, where the impact of model architecture, pre-processing, and post-processing techniques are thoroughly examined to enhance sensitivity and reduce false positives per hour (FP/h). We also explored custom training strategies to ascertain the most effective setup. The model undergoes a novel *second pre-training* phase before subject-specific fine-tuning, enhancing its generalization capabilities. The optimized model demonstrates substantial performance enhancements, achieving as low as 0.23 FP/h, 2.5× lower than the baseline model, with a lower but still acceptable sensitivity rate, showcasing the effectiveness of applying a BERT-based approach on EEG-based seizure detection.**

***Clinical relevance*— The model enhances clinical seizure detection, offering personalized treatments and better generalization to new patients, akin to successes with transformer-based models, thus significantly improving patient safety and care.**

***Index Terms*—Healthcare, EEG, Time Series Classification, Deep learning**

I. Introduction & Related Works

Epilepsy is a neurological disorder that affects over 50 million individuals in the world [1] and is characterized by abnormal electrical activity of the brain that causes recurrent seizures. While medication management is the cornerstone of treatment, drug-resistant cases often require more advanced interventions, including surgery or neurostimulation. In this context, noninvasive Electroencephalography (EEG) data plays a crucial role in seizure detection and monitoring to trigger closed-loop actions such as neurostimulation.

Current methods of classifying raw EEG mostly rely on deep neural networks (DNNs), which demonstrated higher accuracy than classical machine learning algorithms [2]. On the other hand, DNNss often face challenges when extracting features from relatively long time windows because of the lack of exploring the global correlation of all the input samples [3], [4]. Recent approaches [5], [6], [7] showed promising results when working on a reduced number of EEG channels for the deployment of seizure detection approaches on wearable low-power devices. However, most of the works mentioned above fail to reach a False Positive per hour (FP/h) ratio that can enable the clinical implementation of such methods. Although they can correctly detect almost all seizure events, they show a FP/h ratio that jeopardizes their real-world deployment, where wrongly reported seizures will alert the patient and reduce the detection device reliability. Furthermore, another key open challenge in all DNN models is the unavailability of very large high-quality labelled datasets [8], as collecting and labelling seizure data is expensive and requires complex, human labour-intensive protocols, as well as patient hospitalization.

Our work investigates whether a model based on the Transformer architecture, which has recently gained widespread adoption in numerous deep learning applications, coupled with self-supervision and pre-training on large unlabeled datasets, can overcome the data labelling bottleneck and improve the state-of-the-art EEG-based seizure detection for long-term epilepsy monitoring. Our work focuses on the reduction of the FP/h ratio, even at the expense of a reduced sensitivity, working towards the objective of clinical deployment of seizure detection methods: we want a model that predicts a sufficient number of seizure events, with the least number of false alarms, thus reducing stress and anxiety of the monitored patients. The state-of-the-art in this domain is dominated by CNN-based methods [9], [10], [11], [12], with the best approach achieving 100% sensitivity and 0.58 FP/h on CHB-MIT dataset [13]. We take inspiration from a recent study [14] that applied Transformers and self-supervised sequence learning in Brain–Computer Interface (BCI) with EEG data. Our contributions are:

- The adaptation for a seizure detection task on CHB-MIT of BERT-inspired Neural Data Representations (BENDR) [14], an approach inspired by wav2vec 2.0 [15] and Bidirectional Encoder Representations from Transformers (BERT) [16] which allows to exploit massive amounts of unlabeled EEG data.
- An extensive task-specific optimization of the base BENDR model, including i) custom training strategies to improve seizure-detection fine-tuning on CHB-MIT; ii) the evaluation of the impact of architectural choices; iii) the application of pre- and post-processing techniques.
- A promising application of a transformer-based approach,

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

©2024 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

pre-trained on a large amount of unlabelled data, for the seizure detection task.

Thanks to our optimizations, our best model can reduce the FP/h to as low as 0.23, outperforming the best SoA model by 2.5×, with a lower yet acceptable sensitivity of 72.58% [17].

II. Materials & Methods

A. Model training and base architectures

1) Self-supervised pre-training: The first model training step considered in this work consists in the same self-supervised pre-training scheme of BENDR, which in turn closely mirrors wav2vec 2.0 [15]. Namely, the model is pre-trained on a large unlabelled EEG dataset (TUEG) [18] to learn intrinsic patterns in EEG data. This initial phase then lays the groundwork for subsequent fine-tuning on our smaller and task-specific labelled dataset for seizure detection (CHB-MIT). Fig. 1a illustrates the model architecture used in the self-supervised pre-training phase. The encoder's objective is to reconstruct the original sequence elements. This is achieved thanks to a contrastive loss function, which aims to align each element of the transformer's output sequence c_i with the corresponding input b_i, notwithstanding the masking. In our work, we do not re-implement this phase, and directly leverage the pre-trained weights of the original BENDR paper [14].

2) Fine-tuning: The masking component is omitted in the fine-tuning phase, and BENDR vectors are fed directly to the Transformer. A classification block composed of one or more linear layers with a final Softmax activation is then employed to predict seizure occurrences based on the first element of the transformer output sequence. Notably, during pre-training, this first input token was assigned a special value of -5 (selected to be out-of-distribution), and excluded from the contrastive loss. This step is crucial to ease downstream specialization to identify seizures accurately. Fig. 1b shows the architecture used in this phase.

A significant modification of the fine-tuning task with respect to BENDR is the use of the Sensitivity-Specificity Weighted Cross-Entropy (SSWCE) loss, introduced by [7].

Building upon the insights from prior research [19], this study employs a subject-specific approach for fine-tuning the seizure detection model. Such a tailored approach is pivotal in capturing the unique EEG patterns and seizure characteristics inherent to each individual, enhancing sensitivity and specificity. Precisely, we employ a Leave-One-Out Cross-Validation (LOOCV) strategy, which involves training the model on all seizure-containing records of a patient except one. The excluded record is then used as test set, to assess the model's efficacy. A further 20% of the training records is extracted to obtain a validation set, used for early stopping, learning rate scheduling, etc. The whole procedure is repeated cyclically, considering different records as test set.

3) Second Supervised Pre-training: A novel aspect of this work is the implementation of a *second pre-training* phase on CHB-MIT, prior to the subject-specific LOOCV. In this phase, the model is trained in a supervised way (as described in Sec. II-A2 and with the architecture of Fig. 1b), to predict

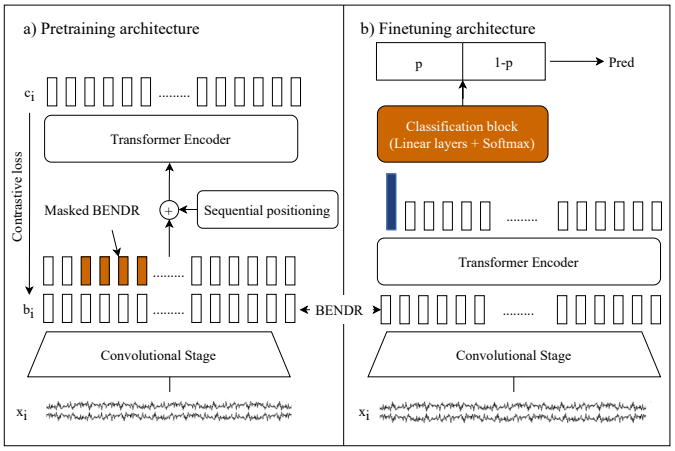


Fig. 1. Model architecture used during (a) self-supervised pretraining (b) the second supervised pre-training and subject-specific fine-tuning.

seizures on *all subjects except the target* one. This step aims to imbue the model with a broader understanding of EEG patterns associated with seizures across different subjects, before honing in on the specific characteristics of a single patient. Figure 2 illustrates the comprehensive training approach employed in the study.

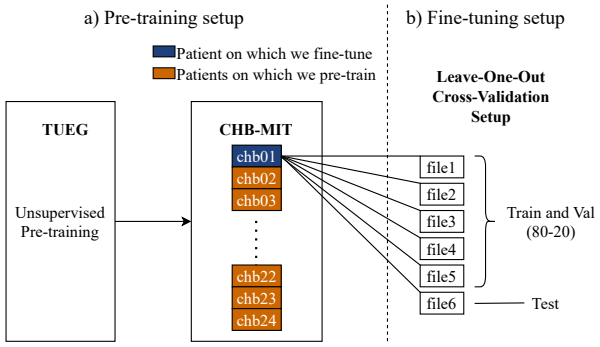


Fig. 2. The model training scheme considering the pre-training, the second pre-training and the subject-specific LOOCV fine-tuning step

B. Task-specific optimizations

In order to maximize seizure detection accuracy, we explore with varying the model architecture, and apply several pre- and post-processing optimizations.

Keeping the fundamental architecture of Fig. 1, we explored different **classification block architectures**, finally converging to a sequence of 4 fully connected layers with decreasing dimensions, progressively reducing the feature space {(512, 256), (256, 128), (128, 64), (64, 2)} and improving the discriminative power of the model. We then **vary the number of blocks** in the Convolutional stage in {3, 6}, and the number of layers and heads in the Transformer Encoder in {2, 4, 6, 8, 12} to maximize seizure detection accuracy. Additionally, we evaluate different **weights initialization policies** [20], in order to apply weights pre-trained on TUEG to models of different sizes (e.g., with a smaller/larger number of layers). Namely, to initialize the additional layers of a bigger model with respect

to the original one, we tested both the duplication of the pre-trained weights and random initialization. On the other hand, when testing a smaller model, we initialize its weights only with the layers shared with the original model.

Based on the hypothesis that early layers may capture the underlying, task-independent structure of EEG data, we consider **freezing the weights** of convolutional layers and fine-tuning only the Transformer encoder. For completeness, we also test how the model behaves when the Transformer encoder is frozen as well.

As **pre-processing**, we consider applying a 5^{th}-order Butterworth bandpass filter within a 0.5–50 Hz frequency range to the input, to smooth the signal and reduce ripple effects, followed by either MinMax or MeanStd normalization. Moreover, to address the challenge posed by the highly unbalanced nature of EEG data, we tested two **oversampling** methods during supervised training phases, Synthetic Minority Oversampling TEchnique (SMOTE) and Weighted Random Sampler.

Lastly, we apply **post-processing** on the model's output to further reduce FP/h. Specifically, we go through the predicted labels with a sliding window and replace the central element with an aggregate prediction over the window. We consider two aggregation criteria: majority voting and minPooling. In the latter, we select the smallest predicted value in the window as the output. Windows of lengths 3, 5, 7 were considered.

C. Training Protocol

We train our models using an Adam optimizer with a learning rate of $1e^{-4}$ and 0.01 weight decay, reducing the learning rate of factor 0.1 when the validation loss stops improving for 5 consecutive epochs. We apply early stopping on the validation loss with a patience of 15 epochs. Dropout layers with 50% probability are added both in the Convolutional stage (between 1D convolutions and Group Normalization layers) and in the Transformer encoder to reduce overfitting.

D. Datasets

1) Self-supervised Pre-training Data: For the self-supervised pre-training phase, the Temple University Hospital EEG Corpus (TUEG) [18] was utilized, encompassing 1.5 TB of clinical EEG recordings. This dataset includes over 10,000 individuals, with diverse demographics, including 51% females and a wide age range. Pre-existing pre-training weights were employed as the starting point for model initialization.

2) Supervised Fine-tuning Data: The supervised fine-tuning phase utilized the CHB-MIT Scalp EEG Database [13], comprising EEG recordings from 24 pediatric subjects with intractable seizures at Boston Children's Hospital. The dataset, collected using the International 10-20 system, includes 664 EDF files, with 198 containing seizure events. For this study, only files with seizure occurrences were considered. Additionally, only 20 out of 23 channels were considered to match the channels used in the pre-training task on TUEG. All signals were sampled at 256 samples per second with 16-bit resolution and a window length of 8s was considered, without overlap.

III. Experimental Results

All results that are presented were obtained considering the LOOCV setup described in II-A3, oversampling training data with a Weighted Random sampler. In terms of metrics, we focus on sensitivity and FP/h: sensitivity represents the ratio of correctly identified seizure *events*, i.e., an event for which at least one of the segments that compose it is assigned a positive label; FP/h are the number of false alarms in an hour, which is inversely proportional to the specificity. After an extensive hyperparameters search for the SSWCE loss, we found $\alpha = 0.8$ and $\beta = 0.2$ to be the best trade-off to maximize sensitivity and reduce FP/h.

A. Model Performance

Table I details our seizure detection results on the CHB-MIT dataset. First, we report the baseline models, i.e., the ones obtained fine-tuning on CHB-MIT the original BENDR architecture [14] without any of our optimizations. This serves as a starting reference to demonstrate the improvements achieved by more refined solutions. Baseline results are reported both with pre-trained weights from TUEG and without any pre-training. Without pre-training, the baseline achieves a 50.15% sensitivity and an extremely high FP/h of 132.24. This is reduced to 10.38 FP/h when using pre-trained weights, demonstrating the crucial role of self-supervised pre-training for this kind of model. In subsequent table rows, we report our successive refinements of the model obtained with optimizations described in the previous section, applied incrementally to the baseline. Results in bold represent the configurations that are used after each optimization step as a starting point for the next one.

First, we explore freezing the weights of the two main parts of the architecture: as expected, freezing the encoder weights worsens the results, while freezing the initial convolutional blocks leads to good generalization capabilities. We impute this to the fact that initial convolutional layers extract generic enough (i.e., not-task-specific) features from EEG signals during the pre-training. This model reaches better sensitivity (54.93%) and FP/h (12.44). On top of this, by applying our second (supervised) pre-training and filtering pre-processing, we obtain a model with an improved sensitivity of 69.11% and a reduced FP/h of 6.95. The last steps are the exploration of the model architecture and the application of pre- and post-processing. Firstly, MeanStd normalization behaves better than the MinMax one. From what concerns model architecture, we notice that a more complex classifier, can learn more discriminative EEG patterns and make more accurate predictions, reaching 81.87% sensitivity and 2.75 FP/h. Then, we tune the number of convolutional layers and transformer block, obtaining the best generalization capabilities with 6 convolutional blocks and 4 attention layers in the encoder, with 4 heads each. Our best post-processing pipeline applies minPooling on the output, using a sliding window of length 3. This leads to 72.58% of seizures detected, with 0.23 FP/h.

TABLE I
RESULTS OBTAINED ON OPTIMIZATION DIMENSIONS ON CHB-MIT

Optimization Dimension	Setup		Detected Seizures [%]	FP/h
Baseline	w/o pre-trained weights		50.15	132.24
	w pre-trained weights		**43.24**	**10.38**
Weight Initialization	freeze conv. layers		**54.93**	**12.44**
	freeze transf. encoder		57.96	91.56
Pre-training	w second-pretraining		**69.11**	**6.95**
Pre-processing	MeanStd + Filtering		**82.44**	**4.01**
	MinMax + Filtering		79.12	11.32
Model Architecture	new classifier		**81.87**	**2.75**
	Dimension (conv., transf.)	6, 12	86.55	4.59
		6, 4	**83.57**	**2.81**
		6, 2	85.91	5.35
		3, 4	82.01	3.88
Post-processing	Majority voting		78.16	0.60
	minPooling		**72.58**	**0.23**
	Majority + minPooling		73.28	0.26

TABLE II
COMPARISON WITH SOA SEIZURE DETECTION ON CHB-MIT

Type	Work	Detected Seizures [%]	FP/h
CNN-based	Qui et al. [23]	97.1	7.51
	Wang et al. [9]	97.52	3.42
	Zhao et al. [10]	97.79	1.25
	Li et al. [11]	98.47	0.63
	Sahani et al. [12]	100	0.58
Transformer-based	Yan et al. [24]	96.01	27.14
	Hussein et al. [21]	99.8	1.08
	Our work	**72.58**	**0.23**

B. Comparison with State-of-the-Art

Table II shows a comparative analysis with the latest state-of-the-art works that address seizure detection on CHB-MIT considering all patients and using from 18 up to 23 EEG channels. We distinguish between CNN-based and Transformer-based approaches. Our approach outperforms all of them in terms of FP/h, at the cost of lower sensitivity, demonstrating the capability of modern transformer models to capture complex epileptic patterns in EEG. Our approach reduces FP/h of the best performing CNN-based model [12] by 2.5×, while simplifying the heavy pre-processing and feature extraction steps, typical of CNN-based approaches, at the expense of reduced sensitivity of the model (27.42% decrease). Our approach significantly improves upon other transformer-based models, with a 4.7× lower FP/h with respect to [21]. Noteworthy, this is crucial for a real-life closed-loop system, given that a high number of FP/h per hour would cause many warnings in patients, which in turn increase their stress, as discussed in [22]. Moreover, a sensitivity over 50% has already been showed to be an acceptable requirement for a seizure detection method [17].

IV. CONCLUSION

This work proposes a BERT-based approach for seizure detection on the CHB-MIT dataset, demonstrating the potential of transfer learning from general EEG data to seizure detection task. We also validate the effectiveness of a Transformer-based architecture which, once pre-trained on a large amount of unlabelled data, can partially overcome data labelling bottleneck and improve the state-of-the-art results. We then extensively explore hyper-parameters and pre-/post-processing techniques to improve the model performance. The best model found obtains 0.23 FP/h, detecting 72.58% of seizures. Our future work will include compressing the model and reducing the EEG channels to explore the possibility of its deployment on a wearable device, as well as implementing artefacts detection and removal algorithms [25] to further improve performances towards the clinical implementation of such methods.

REFERENCES

[1] World Health Organization, "Epilepsy: a public health imperative," *who.int/publications/i/item/epilepsy-a-public-health-imperative*, 2019.

[2] A. Shoeibi *et al.*, "Epileptic seizures detection using deep learning techniques: A review," *International Journal of Environmental Research and Public Health*, 2021.

[3] R. Zanetti *et al.*, "Approximate zero-crossing: A new interpretable, highly discriminative and low-complexity feature for eeg and ieeg seizure detection," *Journal of Neural Engineering*, 2022.

[4] I. Olokodana *et al.*, "Ezcap: A novel wearable for real-time automated seizure detection from eeg signals," *IEEE TCE*, 2021.

[5] P. Busia *et al.*, "Reducing false alarms in wearable seizure detection with EEGformer: A compact transformer model for MCUs," *IEEE TBioCAS*, 2024.

[6] T. M. Ingolfsson *et al.*, "Towards long-term non-invasive monitoring for epilepsy via wearable EEG devices," in *2021 IEEE BioCAS*, 2021.

[7] ——, "EpiDeNet: An energy-efficient approach to seizure detection for embedded systems," in *Proc. BioCAS*, 2023, pp. 1–5.

[8] K. Rasheed *et al.*, "A generative model to synthesize eeg data for epileptic seizure prediction," *IEEE TNSRE*, 2021.

[9] X. Wang *et al.*, "One dimensional convolutional neural networks for seizure onset detection using long-term scalp and intracranial eeg," *Neurocomputing*, 2021.

[10] S. Zhao *et al.*, "Energy-efficient neural network for epileptic seizure prediction," *IEEE TBME*, 2021.

[11] C. Li *et al.*, "Seizure onset detection using empirical mode decomposition and common spatial pattern," *IEEE TNSRE*, 2021.

[12] M. Sahani *et al.*, "Epileptic seizure recognition using reduced deep convolutional stack autoencoder and improved kernel rvfln from eeg signals," *IEEE TBioCAS*, 2021.

[13] A. Goldberger *et al.*, "Physiobank, physiotoolkit, and physionet : Components of a new research resource for complex physiologic signals," *Circulation*, 2000.

[14] D. Kostas *et al.*, "Bendr: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of eeg data," *Frontiers in Human Neuroscience*, 2021.

[15] A. Baevski *et al.*, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in NIPS*, 2020.

[16] J. Devlin *et al.*, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019.

[17] S. Mathieson *et al.*, "Validation of an automated seizure detection algorithm for term neonates," *Clinical Neurophysiology*, 2015.

[18] A. Harati *et al.*, "The temple university hospital eeg corpus," in *2013 IEEE GlobalSIP*, 2013.

[19] G. Varone *et al.*, "A comprehensive machine-learning-based software pipeline to classify eeg signals: A case study on pnes vs. control subjects," *Sensors*, 2020.

[20] M. Narkhede *et al.*, "A review on weight initialization strategies for neural networks," *Artificial Intelligence Review*, 2022.

[21] R. Hussein *et al.*, "Multi-channel vision transformer for epileptic seizure prediction," *Biomedicines*, 2022.

[22] E. Bruno *et al.*, "Wearable technology in epilepsy: The views of patients, caregivers, and healthcare professionals," *E&B*, 2018.

[23] S. Qiu *et al.*, "Lightseizurenet: A lightweight deep learning model for real-time epileptic seizure detection," *IEEE JBHI*, 2022.

[24] J. Yan *et al.*, "Seizure prediction based on transformer using scalp electroencephalogram," *Applied Sciences*, 2022.

[25] T. M. Ingolfsson *et al.*, "Energy-efficient tree-based eeg artifact detection," in *2022 IEEE EMBC*, 2022.