




# SzCORE: Seizure Community Open-Source Research Evaluation framework for the validation of electroencephalography-based automated seizure detection algorithms

## Journal Article

### Author(s):

Dan, Jonathan; Pale, Una; Amirshahi, Alireza; Cappelletti, William; Ingolfsson, Thorir Mar; [Wang, Xiaying](#) ; [Cossetini, Andrea](#) ; Bernini, Adriano; [Benini, Luca](#) ; Beniczky, Sándor; Atienza, David; Ryvlin, Philippe

### Publication date:

2024

### Permanent link:

<https://doi.org/10.3929/ethz-b-000695780>

### Rights / license:

[Creative Commons Attribution-NonCommercial 4.0 International](#)

### Originally published in:




Epilepsia, <https://doi.org/10.1111/epi.18113>

### Funding acknowledgement:

193813 - PEDESITE: Personalized Detection of Epileptic Seizure in the Internet of Things (IoT) Era (SNF)

## SPECIAL ISSUE ARTICLE

# SzCORE: Seizure Community Open-Source Research Evaluation framework for the validation of electroencephalography-based automated seizure detection algorithms

Jonathan Dan<sup>1</sup>  | Una Pale<sup>1</sup> | Alireza Amirshahi<sup>1</sup> | William Cappelletti<sup>2</sup> | Thorir Mar Ingolfsson<sup>3</sup> | Xiaying Wang<sup>3,4</sup> | Andrea Cossetti<sup>3</sup> | Adriano Bernini<sup>5</sup> | Luca Benini<sup>3,6</sup> | Sándor Beniczky<sup>7</sup>  | David Atienza<sup>1</sup> | Philippe Ryvlin<sup>5</sup> 

<sup>1</sup>Embedded Systems Laboratory, EPFL, Lausanne, Switzerland

<sup>2</sup>LTS4, EPFL, Lausanne, Switzerland

<sup>3</sup>Integrated Systems Laboratory, ETH Zürich, Zürich, Switzerland

<sup>4</sup>Research Department, Swiss University of Traditional Chinese Medicine, Zurzach, Switzerland

<sup>5</sup>Service of Neurology, Centre Hospitalier Universitaire Vaudois, Lausanne, Switzerland

<sup>6</sup>Department of Electrical, Electronic, and Information Engineering, University of Bologna, Bologna, Italy

<sup>7</sup>Aarhus University Hospital and Danish Epilepsy Center, Aarhus University, Dianalund, Denmark

## Correspondence

Jonathan Dan, Embedded Systems Laboratory, EPFL, Lausanne, Switzerland.

Email: [jonathan.dan@epfl.ch](mailto:jonathan.dan@epfl.ch)

## Funding information

Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: 193813

## Abstract

The need for high-quality automated seizure detection algorithms based on electroencephalography (EEG) becomes ever more pressing with the increasing use of ambulatory and long-term EEG monitoring. Heterogeneity in validation methods of these algorithms influences the reported results and makes comprehensive evaluation and comparison challenging. This heterogeneity concerns in particular the choice of datasets, evaluation methodologies, and performance metrics. In this paper, we propose a unified framework designed to establish standardization in the validation of EEG-based seizure detection algorithms. Based on existing guidelines and recommendations, the framework introduces a set of recommendations and standards related to datasets, file formats, EEG data input content, seizure annotation input and output, cross-validation strategies, and performance metrics. We also propose the EEG 10–20 seizure detection benchmark, a machine-learning benchmark based on public datasets converted to a standardized format. This benchmark defines the machine-learning task as well as reporting metrics. We illustrate the use of the benchmark by evaluating a set of existing seizure detection algorithms. The SzCORE (Seizure Community Open-Source Research Evaluation) framework and benchmark are made publicly available along with an open-source software library to facilitate research use, while enabling rigorous evaluation of the clinical significance of the algorithms, fostering a collective effort to more optimally detect seizures to improve the lives of people with epilepsy.

## KEYWORDS

brain imaging data structure, electroencephalography, machine-learning benchmark, seizure detection algorithms

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Author(s). *Epilepsia* published by Wiley Periodicals LLC on behalf of International League Against Epilepsy.

## 1 | INTRODUCTION

Scalp electroencephalography (EEG)-based seizure detection algorithms can optimize and facilitate the diagnostic workup performed in persons with epilepsy (PWE) to improve patients' care and quality of life.<sup>1</sup> Currently, such algorithms are primarily used during in-hospital long-term video-EEG monitoring (LTM) performed in epilepsy monitoring units (EMUs) over periods of a few days to several weeks. Recordings can be processed online (i.e., in real time) or offline. Real-time detection helps inform the EMU staff about an ongoing seizure, thus promoting prompt intervention,<sup>2</sup> whereas offline detection can reduce the physician's EEG reading workload and help detect subtle seizures.

In the past decade, home-based video-EEG has been gradually developed as an alternative to EMU LTM, which enables the prospect of very long-term ambulatory EEG.<sup>3</sup> Home-based video-EEG has similar diagnostic objectives to EMU LTM but can last longer, thanks to lower daily cost and patient and health care system burden.<sup>4</sup> It also benefits from automatic seizure detection, because it is performed without the permanent supervision of health care professionals, and generates large volumes of data.

Ultra long-term ambulatory monitoring, from months to years, has a different scope from LTM and home-based video-EEG recording.<sup>5-7</sup> It can be used to inform PWE and their caregivers of an ongoing seizure to enable protective interventions, provide physicians with more precise seizure counts than what is recalled by PWE and their caregivers to optimize therapy, and document eventual recurrence patterns, which may allow seizure forecasting.<sup>8</sup>

The field of EEG-based seizure detection has benefited from advances in machine learning and the provision of EEG datasets from PWE to train such models. Yet, such datasets with annotated seizures remain rare and often kept private as they must comply with strict legal requirements for personal health data. In contrast, datasets with permissive licenses are recognized as catalysts for developing machine-learning algorithms.<sup>9</sup> The machine-learning task can be formulated as a segmentation problem that aims at identifying the start and end of each seizure event. However, current automated scalp EEG-based seizure detection solutions do not meet the level of performance of human experts.<sup>10</sup>

A key obstacle hindering progress in the field is the lack of standardized protocols for the training and evaluation of seizure detection algorithms. When developing a novel algorithm, researchers can opt to reimplement selected algorithms for comparison within their own evaluation framework. Such a process is highly time-consuming, and often prone to error. Therefore, it is rarely done in practice, resulting in analyses relying on reported metrics

### Key points

- Heterogeneity in the validation of seizure detection algorithms poses challenges for a comprehensive evaluation of these algorithms.
- Free datasets with permissive licenses and algorithms with open-source code allow accessibility, transparency, and reproducibility of results.
- The proposed framework offers an impartial and standardized way to assess the performance of EEG-based seizure detection algorithms.

that are not necessarily comparable.<sup>1</sup> This issue has been tackled in other research fields by providing a standard machine-learning task definition and benchmark, effectively leading to dramatic improvements in fields such as image classification,<sup>11</sup> conversational agents,<sup>12</sup> and computational models of brain function.<sup>13</sup>

The validation of seizure detection algorithms lacks standardization in EEG datasets, evaluation methodology, and performance metrics, as discussed in detail below.

### 1.1 | EEG datasets

EEG datasets collected for the purpose of individual studies are common in the field.<sup>1,14,15</sup> Such private datasets prohibit direct comparison with studies on other datasets, as algorithm performance is highly data-dependent.<sup>16</sup> To date, several datasets have been made publicly available, including Physionet CHB-MIT Scalp EEG Database,<sup>17,18</sup> TUH EEG Seizure Corpus,<sup>19</sup> Physionet Siena Scalp EEG Database,<sup>17,20,21</sup> and SeizeIT1.<sup>36</sup> Working with multiple datasets is challenging owing to various data formats and standards, with disparities in EEG electrodes, reference electrodes, montage, channel nomenclature, channel sequence, sampling frequencies, and annotation formats. A previous community effort attempted to standardize reporting of EEG features for computer-based systems, suggesting the Standardized Computer-Based Organized Reporting of EEG (SCORE) nomenclature, which has been endorsed by the International League Against Epilepsy (ILAE) and International Federation of Clinical Neurophysiology (IFCN).<sup>22</sup> Others have worked on a unified organization of brain imaging files and metadata, suggesting the Brain Imaging Data Structure (BIDS), which is increasingly used in research<sup>23</sup> and which was then extended to organize EEG data.<sup>24</sup> Recent work has made SCORE machine readable and compatible with BIDS through the Hierarchical Event Descriptor (HED)-SCORE schema specification.<sup>25</sup> In Section 2.1 of our framework,

we propose refinements of existing data formats for storing EEG and associated seizure annotations that are based on the EEG-BIDS standard and the HED-SCORE nomenclature. The data format provides standardized inputs and outputs for seizure detection algorithms, allowing any seizure detection algorithm to be operated on any thus standardized dataset. Furthermore, this allows visualization and processing of output seizure annotations irrespective of the algorithm that produces them.

## 1.2 | Evaluation methodology

The methodology used to evaluate seizure detection algorithms has a large influence on reported results. Cross-validation is a statistical method used in machine learning to estimate the performance of an algorithm on independent data.<sup>26</sup> To perform cross-validation, the data are split into multiple folds, which are combined to yield pairs of training set test sets (in this paper, we do not cover the notion of a validation set that can be used to determine hyperparameters of a model). The performance of an algorithm is reported as the average performance on all test sets after generating multiple models using different splits of training and test data. Many methods exist to split the data, but they do not necessarily meet the requirement of independence between the training and test sets, which could lead to overestimation of the performance of an algorithm. Overestimation of the accuracy of patient-independent models can occur if some of the same subjects are present in both the training and test sets or when datasets are too small.<sup>27</sup> Moreover, the chronology of recordings should be respected by only using data in the training set that was acquired prior to the acquisition of the data in the test set for personalized models.<sup>28</sup> In Section 2.2, we propose recommendations for cross-validation of subject-independent and personalized models.

## 1.3 | Performance metrics

The choice of metrics is critical to estimate the performance of automatic seizure detection. The current use of different metrics makes it difficult to perform comparisons between studies. Reported results use different combinations of general performance metrics, such as sensitivity, specificity, precision, accuracy, area under the receiver operating characteristic curve, f1-score, and false alarm rate. These metrics are computed by comparing ground-truth reference annotations provided by a human expert with hypothesis annotations provided by an algorithm. This comparison allows counting of “true positives” (TP; i.e., seizures correctly detected by the algorithm), “false

positives” (FP; i.e., incorrectly labeled as seizures by the algorithm), and “false negatives” (FN; i.e., seizures missed by the algorithm).

However, TP, FP, and FN can be counted using either sample-based scoring or event-based scoring, which can result in very different interpretations of the performance metrics. Sample-based scoring computes performance metrics on a sample-by-sample basis and is sometimes referred to as epoch-based scoring<sup>29</sup> or window-based scoring. Sample-based scoring is widely adopted by the machine-learning community, and it integrates tightly with standard training schemes. Although sample-based scoring captures the fine detail agreement between the reference and hypothesis annotations at the timescale of labels, it does not provide answers to the clinically relevant questions “How many seizures did the patient have?” or “How many seizures were missed by the seizure detection algorithm?” or “How many false alarms were triggered by the system?” Answering these questions requires a scoring method that operates at the granularity level of events (or epileptic seizures), that is, event-based scoring. This can be computed in different ways, such as “any-overlap” or “time-aligned event scoring.”<sup>29</sup> In Section 2.3, we propose metrics for the evaluation of seizure detection algorithms that are designed to address questions of the clinical community and requirements of the machine-learning community.

In summary, the lack of common research practices regarding datasets, cross-validation methodologies, and performance metrics when validating seizure detection algorithms is a limiting factor for fair evaluation of algorithms. In this paper, we propose an open framework for the validation of EEG-based seizure detection algorithms: Seizure Community Open-Source Research Evaluation (SzCORE). This framework is the result of discussions with stakeholders in the field, including PWE, physicians and other health care providers, engineers, computer scientists, and other scientists working on the development of seizure detection algorithms. It aims to lift the technical barriers that slow down the development of new algorithms, allowing them to operate on multiple datasets and to be assessed using a fair and objective methodology. Based on the framework, we propose the 10–20 EEG seizure detection benchmark (Section 3) that defines the datasets, tasks, and performance evaluation of seizure detection algorithms. In the future, other benchmarks that target other datasets and tasks can be constructed based on this work, for example, focusing on wearable sensors or intracranial EEG. Additionally, we provide an open-source code library available on GitHub (<https://github.com/esl-epfl/sz-validation-framework>). The library is designed to allow continuous improvement by the community. The framework, benchmark, and supporting code

library are described on an online platform (<https://esl-web.epfl.ch/epilepsybenchmarks>), which also serves as the central hub for a community-built benchmark, where new seizure detection algorithms can be fairly compared.

## 2 | SzCORE FRAMEWORK

### 2.1 | EEG datasets and data format

Datasets should include raw EEG signals, recording specifics, seizure annotations, and patient details, for example, according to EEG-BIDS specifications.<sup>23,24</sup> They should be organized to allow computer systems to process them efficiently. An example of EEG-BIDS data file-structure organization for a dataset of PWE is provided in Data S1.

To allow algorithms to operate seamlessly on any dataset, we propose standardization of EEG data that is at least consistent with the IFCN and ILAE minimum recording standards that are recommended for EEG.<sup>30</sup> Recordings should be stored in .edf files. They should contain the 19 electrodes of the international 10–20 system in a unipolar common average montage. The recording should be resampled to 256 Hz for storage, and source data should be acquired with a sampling frequency of at least 256 Hz. We recommend providing the channels in the following order: Fp1-Avg, F3-Avg, C3-Avg, P3-Avg, O1-Avg, F7-Avg, T3-Avg, T5-Avg, Fz-Avg, Cz-Avg, Pz-Avg, Fp2-Avg, F4-Avg, C4-Avg, P4-Avg, O2-Avg, F8-Avg, T4-Avg, T6-Avg.

The annotation format should be constructed so that it can be used for both source annotations (ground truth) and the output of seizure detection algorithms. The format we propose is a tab-separated values (.tsv) file that is human-readable. It is a text file that uses a tab as a delimiter to separate the different columns of information, with each row representing one event. Each annotation file is associated with a single EEG recording. A detailed description

and an example of the information contained in annotation files is provided in [Supplementary material S1](#). These files adhere to the EEG-BIDS guidelines and use the hierarchical ILAE-based classification of seizures defined by HED-SCORE.<sup>23,25,31</sup> The seizure nomenclature is presented in [Figure S3](#) in [Supplementary material S1](#).

### 2.2 | Evaluation methodology

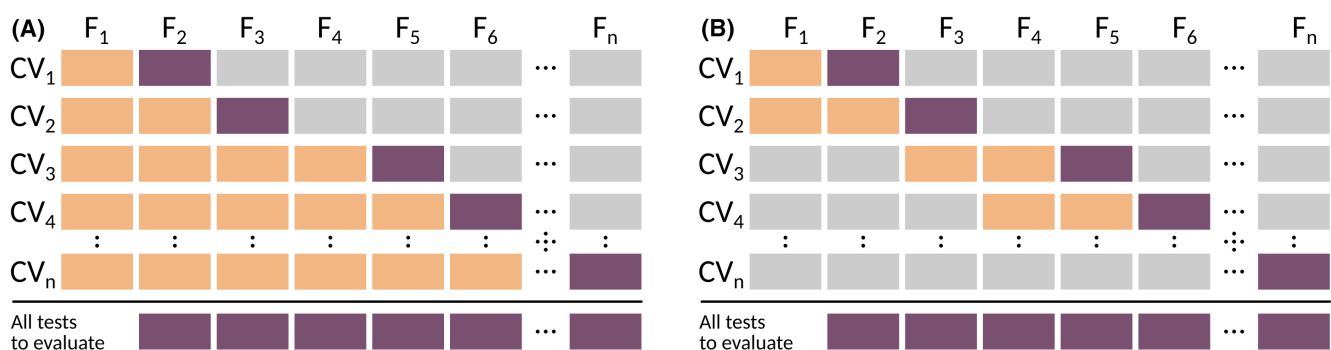
To evaluate seizure detection algorithms, a training set is used to determine the parameters of the machine learning algorithm, and an independent test set is used to estimate its performance. These sets should be independent to guarantee that results can be generalized to other data. If data are only available from a single setting, the dataset can be split into a training set and a test set. This process is repeated multiple times (i.e., folds) to obtain robust estimates of performance by rotating data between the training set and the test set, that is, cross-validation.<sup>26</sup>

#### 2.2.1 | Personalized models

Personalized models are trained for a specific patient. These models should successfully detect seizures in unknown recording sessions that took place after the model was initially trained. To evaluate these models, at each fold, the training set should include only data that were acquired prior to the acquisition of the test set; this is referred to as time-series cross-validation (TSCV).

TSCV can be performed in two ways:

- Training data increase as the model is evaluated on future test folds (variable number of data; [Figure 1A](#)).
- Training data keep a fixed size, with oldest folds removed from the training data as the model is evaluated



**FIGURE 1** Time series cross-validation for personalized models. Each box represents an epoch of data. Orange boxes are used for training; purple boxes are used for testing. Each row represents a cross-validation (CV) fold. The final results are calculated by appending all cross-validation folds (shown in the last row). (A) Cross-validation scheme with variable number of training data. (B) Cross-validation scheme with fixed number of training data.

on future folds (fixed number of data; Figure 1B). This approach ensures models are more sensitive to new data while keeping training complexity and time fixed.

## 2.2.2 | Subject-independent models

Subject-independent models are designed to operate on data from any patient and seizure type. These models should successfully detect seizures in subjects whose data were not used to train the model. Several methods can be used to validate subject-independent models, provided that independence of subjects between training and test sets is maintained:

- Leave one subject out (LOO) is a technique in which many different models are trained.<sup>32</sup> Each model is trained using all the data except those from one subject. The data from that subject are used for testing. This allows maximization of the number of training data provided to the model. Final performance is reported by averaging the testing results of all subjects (each using their subject-independent model). This strategy also allows assessment of the performance of each subject, which can then be compared between different algorithms. However, the technique is not appropriate for large datasets with many subjects, as training models can be computationally expensive and need to be re-trained for every subject.
- K-fold cross-validation uses a similar strategy to LOO.<sup>32</sup> The dataset is split into a training and testing subset with a ratio of subjects of  $(K - 1) / K$  for the training set and  $1 / K$  for the test set. This split is repeated  $K$  times until all subjects are included once in the test set. For each split, a model is trained and performance is reported as an average of each model. This is faster to train and test and thus more appropriate for larger datasets, as the

number of splits is determined by  $K$ , irrespective of the number of subjects. However, this method uses fewer data in the training set than LOO, which can lead to suboptimal models with larger variability in estimated performance. LOO is a special case of  $K$ -fold, where  $K$  is equal to the number of subjects.

- Fixed training and test sets with predetermined subjects in each set are appropriate for large datasets (e.g., TUH EEG Seizure Corpus). However, they can lead to more variability in estimated performance in small datasets. Although cross-validation allows a fair assessment of algorithms during development, the performance of algorithms for real-world use should be evaluated on large independent datasets, which are currently missing in our community.

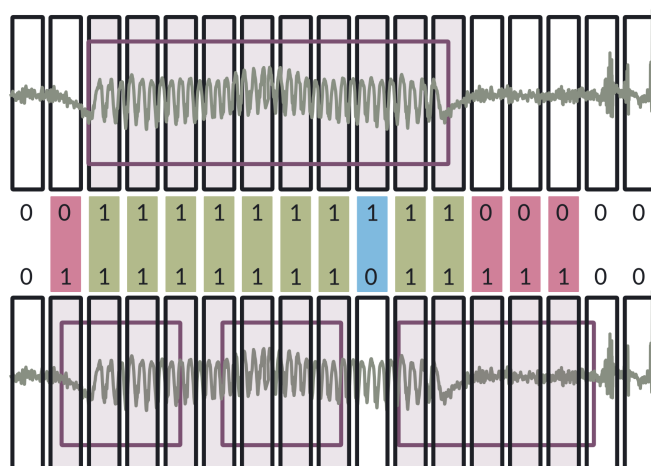
## 2.3 | Performance metrics

To assess the performance of seizure detection algorithms, we propose two complementary scoring methodologies, sample-based and event-based scoring. Both of these scoring metrics should be reported when communicating results of algorithms, as sampled-based metrics provide a high granularity to machine-learning experts and event-based metrics provide clinically relevant results.

Sample-based scoring compares annotation labels, which are provided at a fixed frequency (we propose 1 Hz), sample by sample, to detect TP, FP, and FN, as shown in Figure 2. We propose a frequency of labels of 1 Hz, as it corresponds to the resolution expected by a human annotator. It should be noted that this frequency does not dictate the duration of data windows used to generate machine-learning predictions. These can use an arbitrary duration and overlap as long as they provide predictions at 1 Hz. For annotation labels that overlap only partially with epileptic seizures, we propose assigning a “seizure” label to a sample if the overlap exceeds 50%.



**FIGURE 2** Sample-based scoring compares annotation labels sample by sample. Correct detections (green), false detections (red), and missed detections (blue) are shown. Seizure annotations are indicated in purple.



Event-based scoring, in which events are seizures, relies on overlap between reference and hypothesis annotations (Figure 3). Overlap is considered as correct detection, that is, TP. Hypothesis events that do not overlap with a reference event are counted as FP.

Accurate annotations of epileptic seizures marking a clear start and end are notoriously difficult. This may be complicated by gradual changes in EEG at the beginning and end of seizures or by other factors, for example, muscle or movement artifacts. Subtle EEG changes prior to the marked seizure onset or following marked offset are often detected by various algorithms.<sup>33,34</sup> Some tolerance is therefore required with regard to the onset and duration of seizure to match annotations between two reviewers (e.g., computer algorithm and human expert). From a practical perspective, slight misalignment in seizure annotation onset and duration should not negatively impact the estimated performance metrics. On the contrary, early detection could be beneficial to the patient when the detection algorithm serves as an alarm.

Another issue concerns seizure duration. As most seizures do not occur in rapid succession, it is reasonable to merge annotations separated by only a few seconds. Finally, because seizures are only exceptionally longer than 5 min and longer events are defined as status epilepticus,<sup>35</sup> long events are split into multiple events of a maximum of 5 min.

These considerations are encoded into the following additional rules and parameters to count seizures:

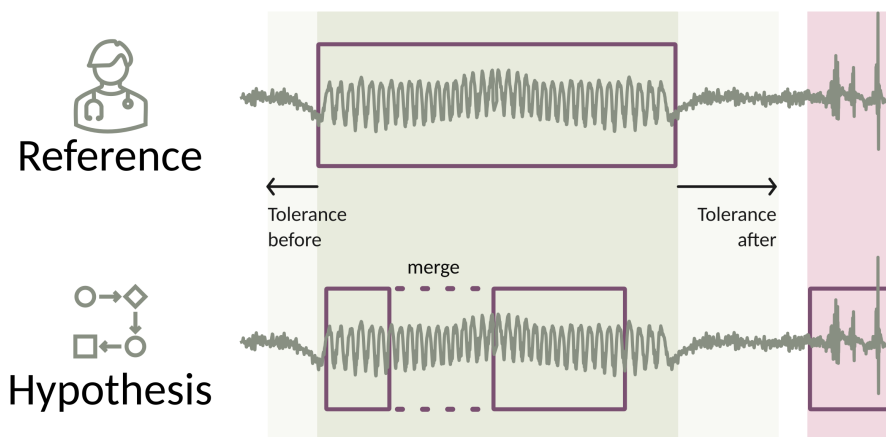
- Minimum overlap between the reference and hypothesis for a detection. We use any overlap, however short, to enhance sensitivity.
- Preictal tolerance, namely, tolerance with respect to the start time of an event that would count as a detection. We advise a 30-s preictal tolerance.
- Postictal tolerance, namely, tolerance with respect to the end time of an event that would still count as a detection. We advise a 60-s postictal tolerance.
- Minimum duration between events resulting in merging events that are separated by less than the given duration. We advise merging events separated by <90s, which corresponds to the combined pre- and postictal tolerance.
- Maximum event duration resulting in splitting events longer than the given duration into multiple events. We advise splitting events longer than 5 min.

### 2.3.1 | Performance metrics

Both the sample-based scoring and event-based scoring produce a count of correct detections (TP), missed detections (FN), and wrong detections (FP). These can be used to compute common performance metrics, as defined below. Specifically, sensitivity and precision are of high interest. F1-score is used as a combined measure containing information on both sensitivity and precision.

- Sensitivity: Percentage of reference seizures detected by the hypothesis. Computed as:  $TP / (TP + FN)$ .
- Precision: Percentage of correct detections over all hypothesis events. Computed as:  $TP / (TP + FP)$ .
- F1-score: Harmonic mean of sensitivity and recall. Computed as:  $2 \times \text{sensitivity} \times \text{precision} / (\text{sensitivity} + \text{precision})$ .
- False alarms per day: Number of falsely predicted (FP) seizure events, averaged or interpolated to number per day.

We explicitly avoid using metrics that rely on a count of TN, such as specificity and accuracy. This is because in the context of event-based scoring, nonseizure events are ill-defined, and in the context of sample-based scoring, nonseizure samples are much more numerous than seizure samples, given the rarity of seizures, resulting in extremely high scores for specificity and accuracy, with little clinical relevance.



**FIGURE 3** Event-based scoring is based on overlap. It defines a set of rules for event merging, tolerance before and after events, and maximum event duration. Correct detections (green) and false detections (red) are shown. Seizure annotations are indicated in purple.

### 3 | BENCHMARK

The framework described above allows building a standard by which seizure detection algorithms can be compared. Here, we propose a seizure detection benchmark for EEG recordings obtained with a 10–20 system that defines:

- The data that should be used when evaluating algorithms.
- The task and different scenarios that the algorithms should analyze.
- The performance metrics and reporting guidelines for these algorithms.

Datasets should be available to allow for transparent reproduction of results. Currently, four large public datasets are available,<sup>36,37</sup> namely, Physionet CHB-MIT Scalp EEG Database, TUH EEG Seizure Corpus, Physionet Siena Scalp EEG, and SeizIT1. A summary of the data contained in these datasets is provided in Table 1.

The currently available public EEG datasets do not all meet the minimum recording requirements of the framework. To use them, the following manipulations are required:

- EEG signals are resampled to 256 Hz.
- Channels are renamed and rereferenced to 10–20 EEG with a common average reference.
- Annotations are converted to EEG-BIDS-/HED-SCORE-compliant .tsv files.
- Data are reorganized according to EEG-BIDS specifications.
- Some recordings of the TUH EEG Seizure Corpus do not contain all 19 electrodes from the 10–20 system. Missing electrodes are replaced by zero values.

An exception is the Physionet CHB-MIT Scalp EEG Database, which provides only bipolar channels, for which a conversion to the proposed unipolar montage is not possible. This dataset is analyzed with the source bipolar montage.

The machine-learning task can be formulated as a segmentation problem that aims to identify the start and end of each seizure event. Three test scenarios are proposed for the evaluation of seizure detection algorithms:

1. Personalized models.
2. Subject-independent models evaluated on a single dataset.
3. Subject-independent models evaluated across datasets.

Personalized models require sufficient data per subject in terms of number of seizures ( $\geq 3$ ; three seizures allow at least one seizure for training, validation, and test sets) and duration ( $\geq 1$  h 30 min; 2 h corresponds to 30 min of data around each seizure) to be effectively trained and evaluated. For this reason, only the following datasets are considered for personalized models: CHB-MIT, Siena, and SeizeIT (TUH seizure dataset is excluded, as it does not contain enough data [10 min on average] per subject and fewer than three seizures per subject). TSCV with a variable number of data is used. The initial training set includes at least 5 h and a minimum of one seizure. Performance is evaluated on the following hour. The process is repeated by successively adding 1 h of training data and testing on the next hour until the end of the recording. Performance per subject is calculated for sample- and event-based metrics by aggregating all 1-h test sets. The performance of a dataset is computed as the average performance of individual subjects.

Subject-independent models evaluated on a single dataset should use LOO or K-fold cross-validation as long as subject independence is guaranteed. Sample-based metrics aggregate all samples of individual subjects. Overall performance is reported as the average of all subjects. Event-based metrics aggregate all events in the same manner. All four datasets can be evaluated.

Subject-independent models evaluated across datasets are trained on a single dataset and tested on the other datasets to verify generalization properties. Sample-based metrics aggregate all samples of individual subjects, and

**TABLE 1** Publicly available scalp electroencephalographic datasets of people with epilepsy.

Dataset	Overview			Recordings		Data	
	Subjects, <i>n</i>	Duration, h	Seizures, <i>n</i>	Files, <i>n</i>	Average duration, min	Frequency, Hz	Channels, <i>n</i>
CHB-MIT	23	982	198	686	60	256	22–38
TUH	675	1476	4029	7377	10	250–1000	17–128
Siena	14	128	47	41	150	512	35–45
SeizeIT1	42	4211	182	458	612	250	26

Abbreviations: CHB-MIT, Physionet CHB-MIT Scalp EEG Database; Siena, Physionet Siena Scalp EEG Database; TUH, TUH EEG Seizure Corpus.



then calculate mean performance over all subjects. Event-based metrics aggregate all events in the same manner.

The algorithm should report performance for sample-based and event-based scoring including sensitivity, precision, F1-score, and false alarms per day for each individual subject (if possible) and overall average of all subjects. In addition, algorithms should provide enough details to allow result reproducibility, for example, in a model card including model description, software and environment documentation, data used, evaluation metrics, and results.<sup>38</sup> An example of such a model card is provided in [Figure S4](#). To help authors document and report results, we provide a checklist for reproducible SzCORE algorithms, which can be found in [Figure S5](#).

To test the validity of the framework and as an initial contribution to the benchmark, we ran SzCORE with three algorithms. The performance results of these algorithms are presented in [Supplementary material S1](#).

## 4 | OPEN-SOURCE LIBRARY AND BENCHMARK PLATFORM

Along with a description of the framework and benchmark, we provide an open-source code library available on GitHub (<https://github.com/esl-epfl/sz-validation-framework>). At the time of writing, the library provides functionality to perform the following actions.

- Convert EEG data from the main public datasets to standardized EEG-BIDS-compliant format (PyPi package: <https://pypi.org/project/epilepsy2bids/>).
- Convert seizure annotations from the main public datasets to standardized HED-SCORE-compliant format.
- Compute the performance of algorithms using event- or sample-based metrics (PyPi package: <https://pypi.org/project/timescoring/>).

The framework, benchmark, and supporting code library are described on an online platform (<https://eslweb.epfl.ch/epilepsybenchmarks>), which also serves as the central hub for a community-built benchmark of seizure detection algorithms. The platform allows researchers to upload results of a seizure detection algorithm following the framework and benchmark described in this work. All results are presented in comparative tables and charts. The platform is designed to allow continuous improvement by the community. It is expected to grow with community submissions. We invite other researchers to contribute by providing feedback or new datasets, supporting the development of the framework and platform, or simply using the framework/benchmark and submitting their algorithms. Details on different ways to contribute are listed in [Data S1](#).

## 5 | DISCUSSION

In this paper, we present SzCORE, a framework for the validation of EEG-based seizure detection algorithms, and suggest common future research practices, with the aim of allowing fair comparison of performance results and increasing reproducibility of studies. This framework is the result of in-depth discussions with stakeholders from both the medical and computer science communities.

The present framework defines standards for EEG datasets based on existing guidelines and recommendations. It also defines data formats for EEG and seizure annotations that comply with the EEG-BIDS data organization and HED-SCORE nomenclature. It provides recommendations and a checklist for sound cross-validation of algorithms and defines performance metrics for their evaluation.

Based on this framework, we propose the EEG 10–20 seizure detection benchmark. The benchmark defines the dataset, task, and performance metrics to evaluate seizure detection algorithms. Additionally, we provide an open-source library to convert data from the public datasets to a standardized data format along with code that implements the performance metrics.

Previous initiatives compared algorithms in the context of contests associated with signal processing congresses (e.g., Neureka IEEE SPMB 2020,<sup>39,40</sup> ICASSP 2023 seizure detection challenge<sup>41,42</sup>). However, evaluation data were not always available after the event, precluding further elaboration or comparison with subsequent algorithms. In contrast, the present benchmark relies on public datasets, and it provides a fully transparent evaluation framework, which will hopefully enable continuing progress in the field.

The proposed benchmark could also be compared to existing commercial algorithms, which are still less performant than human experts but have nonetheless already found some use in the clinic.<sup>10,43</sup>

The choice of 10–20 scalp EEG recording content that lies at the core of the present framework is restricted to the minimum recording standards that are recommended for EMU settings.<sup>30</sup> These are, however, not met by some highly promising developments in long-term EEG, particularly ambulatory wearable EEG and subcutaneous EEG, which tend to use a low number of electrodes positioned in nonstandard locations.<sup>6,7</sup> Whereas our choice appears to exclude such recordings, it can be argued that, whenever possible, recording data with the recommended EMU standards in addition to a novel EEG recording setup guarantees high-quality datasets while allowing for the development of specific benchmarks, for example, targeting wearable EEG. This was the case for the SeizeIT dataset

and ICASSP 2023 seizure detection challenge, which included electrodes positioned behind the ear in addition to standard 10–20 EEG electrodes.<sup>42</sup> In the future, we can expect new guidelines for recording EEG in nonstandard locations or different applications that guarantee high-quality datasets. These new recording standards can use the EEG data format defined in this framework such that they integrate seamlessly with the proposed SzCORE evaluation methodology and performance metrics. They will then be used to extend the online platform by setting up new datasets and benchmarks that specifically target those applications.

The presented framework extends previous work that defined seizure scoring<sup>29</sup> by complementing sample-based with event-based scoring. The current choice of parameters for these scoring methods is somewhat arbitrary if pragmatic. Ideally, the choice of these parameters should either correspond to a specific use of seizure detection algorithms or be based on known uncertainty. Specific use may require high accuracy, for example, prompt intervention triggered by seizure alarms. Other uses benefit from high tolerance, for example, offline review of recordings. In addition, human expert labeling (with is the current gold standard) shows variation,<sup>44</sup> resulting in some uncertainty in labeling the start and end time of seizures.<sup>33,34</sup> Our choice in this respect was dictated by the framework, which aims to be generic and fit a wide range of algorithms and applications. Some users of the framework might want to adapt some of the parameters to their own use case.

This work effectively addresses some current key issues relating to the validation of seizure detection algorithms,<sup>27,28</sup> including the difficulty in comparing results from different datasets and risks associated with a lack of data independence in cross-validation. The best level of evidence for validation is reached when based on an independent multicentric dataset with strong generalizability potential. Such a dataset would contain many recordings from different centers from many subjects, including a variety of seizure types, recording equipment, and recording protocols. As this may be difficult to obtain, we give recommendation for cross-validation strategies that ensure independence within a single dataset. Future work from the community should aim at collecting a large multicentric dataset that can be used for the validation of seizure detection algorithms.

## 6 | CONCLUSIONS

This SzCORE framework and benchmark should foster reproducible, transparent, and efficient research. Crucially, they allow the standardization of the validation of seizure

detection algorithms. This will enable direct comparison of reported results that use this benchmark. We also provide well-described performance metrics that are tailored to both the machine-learning and medical communities. The framework, benchmark, and accompanying open-source software libraries lower the technical and domain-specific knowledge required for algorithm developers to work on seizure detection algorithms and test them on multiple datasets. The benchmark will also allow measuring the state of the art of seizure detection algorithms and guiding new research venues.

To encourage the adoption of the framework, we have set up a community online platform to describe it and collect results of algorithms that use it (<https://eslweb.epfl.ch/epilepsybenchmarks>). We welcome any suggestions for new datasets, new tasks, or improvements to the methodology or content.

## AUTHOR CONTRIBUTIONS

**Jonathan Dan:** Conceptualization; methodology; software; validation; data curation; writing; visualization; project administration. **Una Pale:** Conceptualization; methodology; software; investigation; writing; visualization. **Alireza Amirshahi:** Methodology; investigation; writing—original draft. **William Cappelletti:** Methodology. **Thorir Mar Ingolfsson:** Methodology; investigation; writing—original draft. **Xiaying Wang:** Writing—review & editing. **Andrea Cossetti:** Writing—review & editing; supervision. **Adriano Bernini:** Methodology. **Luca Benini:** Writing—review & editing; supervision; funding acquisition. **Sándor Beniczky:** Writing—review & editing; supervision. **David Atienza:** Writing—review & editing; supervision; funding acquisition. **Philippe Ryvlin:** Methodology; writing—review & editing; supervision; funding acquisition.

## ACKNOWLEDGMENTS

This work was supported by the Swiss National Science Foundation Sinergia grant #193813: PEDESITE—Personalized Detection of Epileptic Seizure in the Internet of Things (IoT) Era. The Pedesite consortium participated in this study through critical feedback on the proposed methodology. In addition, we would like to thank the many international collaborators who participated in discussions that helped build this work, in particular the participants of the Fourth International Congress on Mobile Health and Digital Technology in Epilepsy (2023): Christos Chatzichristos, Lauren Swinnen, Jaiver Macea, and Nick Seeuws from KU Leuven (Belgium), and Bernard Dan and Karine Pelc from ULB (Belgium). Open access funding provided by Ecole Polytechnique Federale de Lausanne.

**CONFLICT OF INTEREST STATEMENT**

None of the authors has any conflict of interest to disclose. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

**DATA AVAILABILITY STATEMENT**

The data that support the findings of this study are openly available in the following repositories:- <https://github.com/esl-epfl/sz-validation-framework>- <https://github.com/esl-epfl/epilepsy2bids>- <https://github.com/esl-epfl/timescoring>- <https://zenodo.org/records/10259996>- <https://zenodo.org/records/10640762>.

**ORCID**

Jonathan Dan  <https://orcid.org/0000-0002-2338-572X>

Sándor Beniczky  <https://orcid.org/0000-0002-6035-6581>

Philippe Ryvlin  <https://orcid.org/0000-0001-7775-6576>

**REFERENCES**

- Baumgartner C, Koren JP. Seizure detection using scalp-EEG. *Epilepsia*. 2018;59:14–22. <https://doi.org/10.1111/EPI.14052>
- Kamitaki BK, Yum A, Lee J, Rishty S, Sivaraaman K, Esfahanizadeh A, et al. Yield of conventional and automated seizure detection methods in the epilepsy monitoring unit. *Seizure*. 2019;69:290–5. <https://doi.org/10.1016/j.seizure.2019.05.019>
- Hasan TF, Tatum WO. Ambulatory EEG usefulness in epilepsy management. *J Clin Neurophysiol*. 2021;38(2):101–11. <https://doi.org/10.1097/WNP.0000000000000601>
- Tatum WO, Desai N, Feyissa A. Ambulatory EEG: crossing the divide during a pandemic. *Epilepsy Behav Rep*. 2021;16:100500. <https://doi.org/10.1016/J.EBR.2021.100500>
- Japaridze G, Loeckx D, Buckinx T, Armand Larsen S, Proost R, Jansen K, et al. Automated detection of absence seizures using a wearable electroencephalographic device: a phase 3 validation study and feasibility of automated behavioral testing. *Epilepsia*. 2023;64:S40–S46. <https://doi.org/10.1111/epi.17200>
- Macea J, Bhagubai M, Broux V, De Vos M, Van Paesschen W. In-hospital and home-based long-term monitoring of focal epilepsy with a wearable electroencephalographic device: diagnostic yield and user experience. *Epilepsia*. 2023;64(4):937–50. <https://doi.org/10.1111/EPI.17517>
- Weisdorf S, Duun-Henriksen J, Kjeldsen MJ, Poulsen FR, Gangstad SW, Kjær TW. Ultra-long-term subcutaneous home monitoring of epilepsy—490 days of EEG from nine patients. *Epilepsia*. 2019;60:2204–14. <https://doi.org/10.1111/EPI.16360>
- Andrzejak RG, Zaveri HP, Schulze-Bonhage A, Leguia MG, Stacey WC, Richardson MP, et al. Seizure forecasting: where do we stand? *Epilepsia*. 2023;64:S62–S71. <https://doi.org/10.1111/EPI.17546>
- Handa P, Mathur M, Goel N. EEG datasets in machine learning applications of epilepsy diagnosis and seizure detection. *SN Computer Sci*. 2023;4(5):1–11. <https://doi.org/10.1007/S42979-023-01958-Z>
- Reus EEM, Visser GH, van Dijk JG, Cox FME. Automated seizure detection in an emu setting: are software packages ready for implementation? *Seizure*. 2022;96:13–7. <https://doi.org/10.1016/J.SEIZURE.2022.01.009>
- Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. p. 248–55 2010. <https://doi.org/10.1109/CVPR.2009.5206848>
- Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000+ questions for machine comprehension of text. In: Su J, Duh K, Carreras X, editors. Proceedings of the 2016 conference on empirical methods in natural language processing. Austin, TX: Association for Computational Linguistics; 2016. p. 2383–92. <https://doi.org/10.18653/v1/D16-1264>
- Schrimpf M, Kubilius J, Lee MJ, Murty NAR, Ajemian R, DiCarlo JJ. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*. 2020;108(3):413–23. <https://doi.org/10.1016/j.neuron.2020.07.040>
- Chatzichristos C, Swinnen L, Macea J, Bhagubai M, Van Paesschen W, De Vos M. Multimodal detection of typical absence seizures in home environment with wearable electrodes. *Front Signal Process*. 2022;2:1014700. <https://doi.org/10.3389/FRSIP.2022.1014700>
- Dan J, Vandendriessche B, Van Paesschen W, Weckhuysen D, Bertrand A. Computationally-efficient algorithm for real-time absence seizure detection in wearable electroencephalography. *Int J Neural Syst*. 2020;30(11):2050035. <https://doi.org/10.1142/S0129065720500355>
- Thuwajit P, Rangpong P, Sawangjai P, Autthasan P, Chaisaen R, Banluesombatkul N, et al. EEGWaveNet: multiscale Cnn-based spatiotemporal feature extraction for EEG seizure detection. *IEEE Trans Industr Inform*. 2022;18(8):5547–57. <https://doi.org/10.1109/TII.2021.3133307>
- Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and Physionet: components of a new research resource for complex physiologic signals. *Circulation*. 2000;101(23):E215. <https://doi.org/10.1161/01.CIR.101.23.E215>
- Shoeb AH. Application of machine learning to epileptic seizure onset detection and treatment. Massachusetts, USA: Massachusetts Institute of Technology; 2009.
- Shah V, von Weltin E, Lopez S, McHugh JR, Veloso L, Golmohammadi M, et al. The Temple University Hospital seizure detection corpus. *Front Neuroinform*. 2018;12:357250. <https://doi.org/10.3389/FNINF.2018.00083>
- Detti P. Siena scalp EEG database V1.0.0. *Phys Ther*. 2020. <https://doi.org/10.13026/5d4a-j060>
- Detti P, Vatti G, Zabalo Manrique de Lara G. EEG synchronization analysis for seizure prediction: a study on data of noninvasive recordings. *PRO*. 2020;8(7):846. <https://doi.org/10.3390/PR8070846>
- Beniczky S, Aurlen H, Brøgger JC, Hirsch LJ, Schomer DL, Trinka E, et al. Standardized computer-based organized reporting of EEG: SCORE—second version. *Clin Neurophysiol*. 2017;128(11):2334–46. <https://doi.org/10.1016/J.CLINPH.2017.07.418>
- Gorgolewski KJ, Auer T, Calhoun VD, Cameron Craddock R, Das S, Duff EP, et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*. 2016;3(1):1–9. <https://doi.org/10.1038/sdata.2016.44>

24. Pernet CR, Appelhoff S, Gorgolewski KJ, Flandin G, Phillips C, Delorme A, et al. EEG-Bids, an extension to the brain imaging data structure for electroencephalography. *Sci Data*. 2019;6(1):1–5. <https://doi.org/10.1038/s41597-019-0104-8>
25. Attia TP, Robbins K, Beniczky S, Bosch-Bayard J, Delorme A, Lundstrom BN, et al. Hierarchical event descriptor library schema for EEG data annotation. 2023. <https://doi.org/10.48550/arXiv.2310.15173>
26. Refaeilzadeh P, Tang L, Liu H. Cross-validation. In: Liu L, Tamer Özsu M, editors. *Encyclopedia of database systems*. Boston, MA: Springer US; 2009. p. 532–8. [https://doi.org/10.1007/978-0-387-39940-9\\_565](https://doi.org/10.1007/978-0-387-39940-9_565)
27. Shafiezadeh S, Duma GM, Mento G, Danieli A, Antoniazzi L, Cristaldi FDP, et al. Methodological issues in evaluating machine learning models for EEG seizure prediction: good cross-validation accuracy does not guarantee generalization to new patients. *Appl Sci*. 2023;13(7):4262. <https://doi.org/10.3390/AP13074262>
28. Pale U, Teijeiro T, Atienza D. Importance of methodological choices in data manipulation for validating epileptic seizure detection models. *Proceedings of the 2023 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Sidney, Australia: IEEE; 2023. <https://doi.org/10.13039/501100011033>
29. Shah V, Golmohammadi M, Obeid I, Picone J. Objective evaluation metrics for automatic classification of EEG events. *Biomed Signal Process*. 2021;223–55. [https://doi.org/10.1007/978-3-030-67494-6\\_8](https://doi.org/10.1007/978-3-030-67494-6_8)
30. Peltola ME, Leitinger M, Halford JJ, Vinayan KP, Kobayashi K, Pressler RM, et al. Routine and sleep EEG: minimum recording standards of the International Federation of Clinical Neurophysiology and the International League Against Epilepsy. *Epilepsia*. 2023;64(3):602–18. <https://doi.org/10.1111/EPI.17448>
31. Scheffer IE, Berkovic S, Capovilla G, Connolly MB, French J, Guilhoto L, et al. ILAE classification of the epilepsies: position paper of the ILAE commission for classification and terminology. *Epilepsia*. 2017;58(4):512–21. <https://doi.org/10.1111/EPI.13709>
32. Hastie T, Tibshirani R, Friedman J. *Model assessment and selection. The elements of statistical learning: data mining, inference, and prediction*. New York: Springer New York; 2009. p. 219–59. [https://doi.org/10.1007/978-0-387-84858-7\\_7](https://doi.org/10.1007/978-0-387-84858-7_7)
33. Maimaiti B, Meng H, Lv Y, Qiu J, Zhu Z, Xie Y, et al. An overview of EEG-based machine learning methods in seizure prediction and opportunities for neurologists in this field. *Neuroscience*. 2022;481:197–218. <https://doi.org/10.1016/j.neuroscience.2021.11.017>
34. Shoeb A, Kharbouch A, Soegaard J, Schachter S, Gutttag J. A machine-learning algorithm for detecting seizure termination in scalp EEG. *Epilepsy Behav*. 2011;22:S36–S43. <https://doi.org/10.1016/j.yebeh.2011.08.040>
35. Trinka E, Cock H, Hesdorffer D, Rossetti AO, Scheffer IE, Shinnar S, et al. A definition and classification of status epilepticus—report of the ILAE task force on classification of status epilepticus. *Epilepsia*. 2015;56(10):1515–23. <https://doi.org/10.1111/EPI.13121>
36. Chatzichristos C, Miguel Claro B. SeizeIT1. KU Leuven RDR 2023. [1048804/P5Q00J](https://doi.org/10.1007/978-3-030-67494-6_8)
37. Wong S, Simmons A, Rivera-Villicana J, Barnett S, Sivathamboo S, Perucca P, et al. EEG datasets for seizure detection and prediction—a review. *Epilepsia Open*. 2023;8(2):252–67. <https://doi.org/10.1002/EPI4.12704>
38. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machinery; 2019. p. 220–9. <https://doi.org/10.1145/3287560.3287596>
39. Chatzichristos C, Dan J, Mundanad Narayanan A, Seeuws N, Vandecasteele K, De Vos M, et al. Epileptic seizure detection in EEG via fusion of multi-view attention-gated U-net deep neural networks. *IEEE Signal Process Med Biol Symp*. 2020;1. <https://doi.org/10.1109/SPMB50085.2020.9353630>
40. Neureka IEEE SPMB 2020. 2020. <https://neureka-challenge.com>
41. Al-Hussaini I, Mitchell CS. SeizFt: interpretable machine learning for seizure detection using wearables. *Bioengineering*. 2023;10(8):918. <https://doi.org/10.3390/bioengineering10080918>
42. Seizure Detection Challenge—IEEE ICASSP. 2023. <https://signalprocessingsociety.org/publications-resources/data-challenges/seizure-detection-challenge-icassp-2023>
43. Koren J, Hafner S, Feigl M, Baumgartner C. Systematic analysis and comparison of commercial seizure-detection software. *Epilepsia*. 2021;62(2):426–38. <https://doi.org/10.1111/epi.16812>
44. Halford JJ, Shiao D, Desrochers JA, Kolls BJ, Dean BC, Waters CG, et al. Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG recordings. *Clin Neurophysiol*. 2015;126(9):1661–9. <https://doi.org/10.1016/J.CLINPH.2014.11.008>

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Dan J, Pale U, Amirshahi A, Cappelletti W, Ingolfsson TM, Wang X, et al. SzCORE: Seizure Community Open-Source Research Evaluation framework for the validation of electroencephalography-based automated seizure detection algorithms. *Epilepsia*. 2024;00:1–11. <https://doi.org/10.1111/epi.18113>