Multi-center Assessment of CNN-Transformer with Belief Matching Loss for Patient-independent Seizure Detection in Scalp and Intracranial EEG

Wei Yan Peh  
Nanyang Technological University

Prasanth Thangavel  
Nanyang Technological University

Yuanyuan Yao  
Delft University of Technology

John Thomas  
Montreal Neurological Institute

Yee Leng Tan

Justin Dauwels (✉️ j.h.g.dauwels@tudelft.nl)  
Delft University of Technology

Research Article

Keywords: Patient-independent Seizure Detection, Transformer, Belief Matching, Electroencephalogram

Posted Date: August 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1944396/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.  
Read Full License
Multi-center Assessment of CNN-Transformer with Belief Matching Loss for Patient-independent Seizure Detection in Scalp and Intracranial EEG

Wei Yan Peh\textsuperscript{1,2,3}, Prasanth Thangavel\textsuperscript{1,3}, Yuanyuan Yao\textsuperscript{4}, John Thomas\textsuperscript{5}, Yee Leng Tan\textsuperscript{6}, and Justin Dauwels\textsuperscript{4,*}

\textsuperscript{1}Nanyang Technological University (NTU), Interdisciplinary Graduate School (IGS), Singapore 639798.
\textsuperscript{2}NTU, School of Computer Science and Engineering (SCSE), Singapore 639798.
\textsuperscript{3}NTU, School of Electrical and Electronic Engineering (EEE), Singapore 639798.
\textsuperscript{4}Delft University of Technology (TU Delft), Department of Microelectronics, 2628 CD Delft, Netherlands.
\textsuperscript{5}Montreal Neurological Institute (MNI), McGill University, Montreal, QC H3A 2B4, Canada.
\textsuperscript{6}National Neuroscience Institute (NNI), Singapore 308433.

\*Corresponding Author: J.H.G.Dauwels@tudelft.nl

ABSTRACT

Neurologists typically identify epileptic seizures from electroencephalograms (EEGs) by visual inspection. This process is often time-consuming, especially for EEG recordings that last several hours or even days. To expedite the process, a reliable, automated, and patient-independent seizure detector is essential. However, developing a patient-independent seizure detector is challenging as seizures exhibit diverse morphologies and characteristics across different patients and recording devices. In this study, we propose a patient-independent seizure detector to automatically detect seizures in both scalp EEG (sEEG) and intracranial EEG (iEEG). First, we deploy a convolutional neural network (CNN) with transformers (TRF) and belief matching (BM) loss to detect seizures in single-channel EEG segments (channel-level detection). Next, we extract regional features from the channel-level outputs to detect seizures in multi-channel EEG segments (segment-level detection). At last, we apply postprocessing filters to the segment-level outputs to determine the start and end points of seizures in multi-channel EEGs (EEG-level detection). We introduce the minimum overlap evaluation scoring (MOES) as an evaluation metric that accounts for minimum overlap between the detection and seizure, improving upon existing assessment metrics. We trained the seizure detector on the Temple University Hospital Seizure (TUH-SZ) sEEG dataset and evaluated it on five other independent sEEG and iEEG datasets. On the TUH-SZ dataset, the proposed patient-independent seizure detector achieves a sensitivity (SEN), precision (PRE), average and median false positive rate per hour (aFPR/h and mFPR/h), and median offset of 0.772, 0.429, 4.425, 0, and -2.125s, respectively. Across all four adult datasets (excluding neonatal and paediatric datasets), we obtained SEN of 0.617-1.00, PRE of 0.534-1.00, aFPR/h of 0.425-2.002, and mFPR/h of 0-1.003. Meanwhile, on neonatal and paediatric datasets, we obtained SEN of 0.227-0.678, PRE of 0.377-0.818, aFPR/h of 0.253-0.421, and mFPR/h of 0.118-0.223. The proposed seizure detector can reliably detect seizures in adult EEGs (to less extent in neonatal EEGs) and takes less than 15s for a 30 minutes EEG. Hence, this system could potentially aid the clinicians in reliably identifying seizures expeditiously, allocating more time for devising proper treatment.

Keywords: Patient-independent Seizure Detection, Transformer, Belief Matching, Electroencephalogram.

Introduction

Epilepsy is a brain disorder characterized by the manifestations of sudden unprovoked seizures\textsuperscript{1}. Seizures are diverse, and vary significantly across patients in etiology, severity, and symptoms\textsuperscript{2}. Seizures are characterized by the location they affect and how far they spread\textsuperscript{3}. Most seizures last from 30 seconds to two minutes, where a seizure lasting longer than five minutes is a medical emergency\textsuperscript{4}. Approximately 60% of seizures are convulsive, where the patient experiences uncontrolled shaking. Examples of such seizures are tonic-clonic seizures, which can develop slowly from focal seizures\textsuperscript{5}. The remaining 40% of seizures are non-convulsive, such as absence seizures, where the patient experiences a brief and sudden lapse of consciousness\textsuperscript{6}. Epilepsy is diagnosed when a patient experiences two or more recurring seizures of any types\textsuperscript{7}. Around 1% of the world population is diagnosed with epilepsy\textsuperscript{8}. Moreover, approximately 10% of the population will experience a seizure within their lifetime\textsuperscript{9}. Overall, provoked and unprovoked seizures occur in about 3.5 and 4.2 per 10000 individuals annually, respectively\textsuperscript{8}. After a seizure episode, the likelihood of encountering another seizure event increases to about 50%, bringing the individual to a much greater risk of relapsing\textsuperscript{10}. 
To detect seizures, an electroencephalogram (EEG) can be utilized. An EEG measures the electrical activity in the brain via electrodes attached to the scalp. Scalp EEG (sEEG) records the brain activity by surface electrodes, while intracranial EEG (iEEG) measures the electrical signals directly via implanted electrodes. sEEGs are primarily utilized for patient monitoring, while iEEGs help in surgery planning for epilepsy patients. However, visual inspection of EEGs can be time-consuming. In practice, only small fractions of the entire recordings are inspected. There is a need for automated seizure detectors that can detect seizures reliably and quickly. Most progress has been made toward patient-specific detectors, as seizure morphologies and characteristics strongly vary across patients. Consequently, designing a seizure detector that can detect any type of seizure in any patient can be challenging due to the wide variety of seizure patterns. Nonetheless, such patient-independent seizure detectors would be tremendously helpful for clinicians. There are several commercial seizure detectors available in the market, such as Persyst, Encevis, and BESA. However, these detectors do not generalize well across patients and different EEG datasets, and cannot be applied to both sEEGs and iEEGs.

In recent studies on automated seizure detection from EEG, the seizure detectors are validated mainly on two public seizure datasets: the Temple University Hospital seizure (TUH-SZ) dataset and the Children’s Hospital Boston Massachusetts Institute of Technology (CHB-MIT) dataset. In many studies, different seizure detectors are proposed, including (standard) machine learning models, convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM), transformer, transfer learning models, and temporal graph convolutional networks (TGCNs). The seizure detectors proposed in these studies are similar in architecture or implementation. Most detectors first divide the multi-channel EEG into short multi-channel segments and classify each segment as normal against seizure (segment-level detection). Lastly, they determine the start and end points of the seizures from the segment-level outputs. For most studies, the main innovation lies in the design of the segment-level seizure detector, where most studies propose increasingly deep and complex neural networks with millions of parameters.

However, computationally intensive models may not necessarily improve patient-independent seizure detection, due to the associated increased risk of overfitting. Furthermore, it is shown in the literature that seizure detectors trained on large datasets reported similar results to those trained on smaller datasets. Overall, we observed a fundamental bottleneck that limits the performance of the state-of-the-art patient-independent seizure detectors. To resolve the bottleneck, we require a fresh perspective on this problem. As we will explain in the following, we address certain drawbacks of the existing seizure detector design and resolve some of its weaknesses in this study.

First, most modern seizure detectors identify seizures from short multi-channel segments (segment-level detection) before using the resulting segment-level outputs to determine the start and end points of seizures in the full EEG. Since these detectors are trained on annotated multi-channel EEG segments, they can only handle a fixed number of EEG electrodes (e.g., 21). To apply those models to EEGs with a different number of electrodes (e.g., 32), the models need to be retrained on EEGs with that same number of electrodes. In practice, the number of electrodes may vary; consequently, this limitation is a severe impediment to clinical applications.

To overcome this limitation, we proposed a seizure detector that starts by detecting seizures in single-channel segments (channel-level detection). We evaluate three variations of CNN for the channel-level detector: CNN with softmax loss (CNN-SM), CNN with belief matching (BM) loss (CNN-BM), and a CNN cascaded with a transformer and BM loss (CNN-TRF-BM). The BM loss is used to improve confidence performance and to obtain a more accurate prediction. A model with good calibration behaviour has a distribution of the probability predicted similar to the actual distribution and behaviour of probability observed in training data. In other words, when the output of the system is near 0 or 1, the accuracy should be high. By contrast, when the output is near 0.5, the system seems less confident, and the accuracy is expected to be lower. Therefore, the more confident the system is in its predictions, the more accurate it is supposed to be. Furthermore, the transformer is deployed to extract long-range patterns across the signals, while CNNs typically can only capture short-range patterns. We assess the three CNN models to determine the best model for the channel-level seizure detector. While studies proposed seizure detectors that detect seizures at individual channels, these studies analyzed single-channel EEGs, rather than multi-channel segments from multi-channel EEGs. Consequently, while they performed channel-level seizure detection, they did not conduct segment-level detection. This limitation can severely impact the seizure detector, as multi-channel EEGs yield more information regarding the topological map than single-channel EEG recordings. Additionally, most of these studies performed patient-specific seizure detection.

Second, in the next stage, the outputs from the channel-level detector are aggregated, and from those outputs, seizures are detected in multi-channel EEG segments (segment-level detection). Concretely, we group the outputs into five brain regions and compute statistical features for each region, which can be done for an arbitrary number of electrodes as long as there are two or more electrodes inside each region. At last, leveraging the outputs of the
segment-level seizure detector, the proposed system determines the start and end points of the seizures (if any). As the system starts by processing individual channels and grouping the outputs into regions, independently of the number of electrodes, the proposed seizure detector can be applied to EEGs with an arbitrary number of electrodes. Moreover, not only can it be applied to sEEG, the same system can also be applied to iEEG without retraining. In this study, we trained the proposed seizure detector on a large sEEG dataset (TUH-SZ dataset) and evaluated it on five independent sEEG and iEEG datasets. The independent datasets consist of patients from different age groups (neonatal, paediatric, and adult) and types (humans and dogs). In comparison, existing seizure detectors for sEEGs and iEEGs are often trained and analyzed separately, and are usually trained and tested on the same dataset (with a few exceptions, see \cite{33-35} for example).

Third, to measure the effectiveness of seizure detectors, an adequate evaluation metric is necessary. Such metrics score a detection from the automated system based on how much it overlaps with a manually annotated seizure(s), considered to be the ground truth. Unfortunately, most studies do not state what metric has been applied to assess the seizure detectors. Consequently, comparing different detectors proposed in the literature can be challenging, even when evaluated on the same dataset. Several evaluation metrics have been proposed in the literature, including the epoch-based sampling (EBS)\(^{10}\), any-overlap (OVLP)\(^{10}\), time-aligned event scoring (TAES)\(^{10}\), and increased margin scoring (IMS)\(^{10}\). However, these metrics do not reflect real-world clinical requirements. Therefore, we introduce the minimum overlap evaluation scoring (MOES), which requires the detection from the automated system to have a minimum overlap duration of 10s and a minimum overlap of 30% with a ground truth seizure for it to be considered correct (true positive). In contrast, OVLP and TAES require a non-zero (e.g., 0.1%) and perfect (100%) overlap, respectively, which tends to under- or over-penalize the detector, respectively. By requiring a non-trivial overlap, albeit not necessarily a perfect overlap, the MOES metric has a more adequate level of tolerance for clinical practice.

In summary, this paper makes the following contributions:

1. The proposed seizure detection system starts by processing individual EEG channels, and aggregates the outputs at individual channels into several brain regions, independently of the number of electrodes. As a result, the proposed system can be applied to both sEEG and iEEG with an arbitrary number of electrodes.
2. We apply CNN with transformer for seizure detection. While CNNs can extract complex patterns from EEG signals, they are less suitable for extracting long-range patterns. A transformer resolves this limitation as it can identify long-range features. Transformers have rarely been explored for seizure detection from EEG (but see\(^{24,30}\)). However, applying transformers on individual channels had not been proposed before.
3. We utilize a belief matching (BM) loss to improve the calibration performance. Unfortunately, many existing classification algorithms are not optimized for obtaining accurate probabilities, and the predictions they produce may be miscalibrated. Having a proper calibration performance is critical for decision-making. Bayesian approaches are rarely applied in EEG analysis, as most studies favour the traditional softmax (SM) loss.
4. We train the proposed patient-independent seizure detector on one sEEG dataset, and test it on five independent sEEG and iEEG datasets. Seizure detectors are usually not assessed on multiple independent datasets, and especially not on sEEGs and iEEGs simultaneously. Moreover, we obtain promising results on datasets on various EEG types (human and dog EEG) and from various age groups (neonatal, pediatric, adult EEG). This suggests that the proposed seizure detector tends to overfit less and generalizes well across multiple datasets.
5. We introduce the minimum overlap evaluation scoring (MOES) to assess the performance of seizure detectors. In contrast to existing metrics, MOES metric requires a non-trivial but not necessarily perfect overlap between the detection and ground truth seizure(s) for the detection to be considered correct. Existing metrics are either too lenient or strict on the detection overlap criteria, resulting in overly optimistic or pessimistic evaluations.
6. We conduct a detailed literature review on patient-independent seizure detectors and benchmark the proposed patient-independent seizure detectors with several state-of-the-art approaches.

**Results**

**Channel-level Seizure Detection**

We performed single-channel (channel-level) seizure detection with three channel-level detectors: CNN with a softmax (SM) loss (CNN-SM), CNN with a belief matching (BM) loss (CNN-BM), and CNN with a transformer and a BM loss (CNN-TRF-BM). In this setting, each single-channel EEG segment is labelled as “normal” or “seizure”. We summarized the results in Table 1. The precision-recall (PR) curves can be found in Supplementary Figure 5.

On the TUH-SZ dataset, the proposed channel-level detectors achieve high balanced accuracy (BAC), sensitivity (SEN), and specificity (SPE) across all window lengths. Moreover, the expected calibration error (ECE) improved for all window lengths (except for 3s) when the SM loss is replaced with the BM loss (CNN-SM against CNN-BM).
However, the ECE is slightly larger for the CNN-TRF-BM model. The performance peaks at a window length of 20s for all three models. Overall, the CNN-TRF-BM model attained the best results, followed by the CNN-BM and the CNN-SM model. As the channel-level detector attains good detection accuracy on the TUH-SZ dataset, deploying it as the primary training dataset seems to be a promising option.

Next, we assessed the channel-level detector, trained on the TUH-SZ dataset, on the five EEG datasets: the Children’s Hospital Boston Massachusetts Institute of Technology (CHB-MIT) dataset, the Sleep Wake Epilepsy Center at ETH Zurich (SWEC-ETHZ) dataset, the Helsinki University Hospital (HUH) dataset, the International Epilepsy Electrophysiology Portal (IEEGP) dataset, and the Epilepsy iEEG Multicenter (EIM) dataset. The detectors achieve high BACs on the CHB-MIT, SWEC-ETHZ, and EIM datasets, but yield poor BACs on the HUH and IEEGP datasets. For those datasets, seizures have only been annotated on the level of segments instead of channels; therefore, it is impossible to assess the channel detector reliably. Without channel-level annotations, we must assume that all channels within a multi-channel segment contain seizures. However, this is unlikely as seizures sometimes only occur in certain regions. In particular, focal seizures occur only in one hemisphere or at a few electrodes. Consequently, channels that do not exhibit seizures may be mislabelled as “seizure”, leading to errors during training and testing. However, segment-level and EEG-level detection results are reliable for those datasets.

**Segment-level Seizure Detection**

Next, we performed multi-channel segment (segment-level) seizure detection using the outputs from the three channel-level detectors. The segment-level seizure detection results are displayed in Table 2. The precision-recall (PR) curves are displayed in Supplementary Figure 5. We evaluated the three models for segment-level seizure detection on the six sEEG/iEEG datasets.

On the TUH-SZ dataset, the proposed segment-level detectors achieve high BAC, SEN, and SPE across all window lengths, similarly to the channel-level results. However, the ECE reported at segment-level is much greater than the channel-level counterparts. This is anticipated as the segment-level detector model does not minimize ECE, and instead utilizes a traditional loss function. Similarly, the performance peaks at a window length of 20s. Again, the CNN-TRF-BM model outshines the other two models.

Next, we evaluated the pre-trained segment-level seizure detector, trained on the TUH-SZ dataset, on the other five sEEG and iEEG datasets. We obtained excellent performance on all the datasets at various window lengths, except for the HUH dataset. The segment-level detectors obtain high BACs on the IEEGP dataset, even when the channel-level results on this dataset are not satisfactory.

Overall, the performance peaks at different window lengths across the six datasets. This might be due to the discrepancy in seizure types, patient types, and patient age groups across the different datasets. Consequently, the seizure length annotated can be very different for each institution (see Supplementary Figure 3 and Supplementary Table 1). For instance, for datasets with many short seizures, one should deploy a window length of 3s as it can capture shorter seizures, while a window length of 20s would be suboptimal.

**EEG-level Seizure Detection**

Next, we applied a post-processing module to determine the start and end times of the seizures based on the outputs of the segment-level detector. We summarized the results for the six datasets in Table 3. We refer the reader to Supplementary Figure 5 for the precision-recall curves. The EEG-level performance is computed according to the minimum overlap evaluation scoring (MOES), as it is more suitable for clinical practice than existing metrics (see Supplementary materials). We also considered other existing evaluation metrics for comparison in Table 4.

On the TUH-SZ dataset, the CNN-TRF-BM model leads to the most promising results, followed by the CNN-BM and the CNN-SM model. The CNN-TRF-BM EEG-level seizure detector attained a high SEN of 0.772, a decent PRE of 0.429, an average FPR/h (aFPR/h) of 0.425, and a median FPR/h (mFPR/h) of 0, and a median offset of -2.125s. While the aFPR/h is high, the mFPR/h is extremely low. This implies that the aFPR/h is skewed by a small number of EEGs containing an exceptionally huge amount of false detection. While the SEN is similar across all three models, the CNN-TRF-BM model reported the best PRE, which is critical for clinical deployment.

Similarly, we evaluated the EEG-level seizure detectors on the five sEEG and iEEG datasets. The CNN models yield high SEN, decent PRE, and low aFPR/h and mFPR/h on the CHB-MIT, SWEC-ETHZ, and EIM datasets. Meanwhile, on the HUH and IEEGP datasets, the model achieves low SEN (0.254 and 0.450, respectively), high PRE (0.841 and 0.917, respectively), and low mFPR/h (0.347 and 0, respectively). The poorer results on the HUH dataset are in line with our expectations since it is a neonatal dataset. The morphology of neonatal seizures differs vastly from adult seizures. Since the model has been trained on adult sEEG, it struggles to detect seizures in neonatal sEEGs. Meanwhile, the IEEGP dataset contains some dog iEEGs, which could have different seizure patterns from adult humans. However, we observed that the detection performance is comparable for human and
Table 1. Channel-level seizure detection results for different CNN models across six EEG datasets.

| Dataset      | W    | CNN-SM | CNN-BM | CNN-TRF-BM |
|--------------|------|--------|--------|------------|
|              | ECE  | ACC    | SEN    | SPE        | F1          | ECE  | ACC    | SEN    | SPE        | F1          | ECE  | ACC    | SEN    | SPE        | F1          |
| TUB-SZ sEEG  | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.25   | 0.46   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| Adult        | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| CHB-MIT sEEG | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| Paediatric   | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| SWEC-EUTH sEEG | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| Adult        | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| HUH sEEG Neonatal | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| Adult        | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| IEEG sEEG Adult | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| EIM sEEG Adult | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| Average (All) | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |

Table 2. Segment-level seizure detection results for different CNN models across six EEG datasets.

| Dataset      | W    | CNN-SM | CNN-BM | CNN-TRF-BM |
|--------------|------|--------|--------|------------|
|              | ECE  | ACC    | SEN    | SPE        | F1          | ECE  | ACC    | SEN    | SPE        | F1          | ECE  | ACC    | SEN    | SPE        | F1          |
| TUB-SZ sEEG  | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| Adult        | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| CHB-MIT sEEG | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| Paediatric   | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| SWEC-EUTH sEEG | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| Adult        | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| HUH sEEG Neonatal | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| Adult        | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| IEEG sEEG Adult | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| EIM sEEG Adult | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
| Average (All) | 10   | 0.22   | 0.41   | 0.58   | 0.94       | 0.96   | 0.10   | 0.21   | 0.41   | 0.58   | 0.94       | 0.96   |
dog EEGs. Nonetheless, the proposed detector detected some neonate and dog seizures with high PRE, which can be tremendously valuable.

We also determined the detection offset, defined as the average duration between the start time of the seizure and the start time of its corresponding detection. As seen from Table 3, this offset can be negative. A negative offset does not imply that the system can predict seizures in advance, as the EEG data is analyzed offline\textsuperscript{47}. Therefore, data from future time intervals are being considered to decide whether a particular EEG segment is ictal.

In Table 4, we compare results for the CNN-TRF-BM model for different evaluation metrics (IMS, OVL, TAES, and MOES). IMS always leads to the best results, followed by OVL, MOES, and TAES. The results for MOES are similar to OVL and IMS, despite MOES having a more stringent condition. This implies that the proposed seizure detector detects most seizures with at least 10s overlap and with 30% overlap between the seizure and detection. The results for the TAES metric are the lowest: a slight drop in SEN, much lower PRE, and significantly higher aFPR/h and mFPR/h. While there are significant differences across the different performance metrics, the results obtained by MOES are the most appropriate (see Supplementary materials). The detection condition of MOES reflects more accurately to clinical requirements and does lead to overly optimistic or pessimistic results.

Finally, to determine the effectiveness of the CNN-TRF-BM-based EEG-level seizure detector (Figure 1), we plot the normalized histograms of the TP and FN of seizures detected sorted by event duration, together with the normalized histogram of SEN, PRE, and FPR/h computed from individual EEGs across the datasets. From Figure 1(a), it can be seen that it is easier to detect a long seizure than a short event. Figure 1(b) and 1(c) reveal that the SEN and PRE are high for most EEGs, with only a minority of the files having a poor detection rate. Lastly, Figure 1(d) confirms that the system does not make false detections in most EEGs, as the mFPR/h is 0. Taken together, these figures seem to suggest that the proposed seizure detector performs well across most EEGs.

In Table 5, we computed the SEN of short (<10s) and long (>10s) seizures across the six datasets for various window lengths. As the CHB-MIT and IEEGP datasets do not contain annotated short seizures, we could not compute the SEN for those datasets. We observe that a window length of 20s leads to drastic drops in SEN for shorter seizures. On the other hand, a shorter window (3s and 5s) can more reliably capture shorter seizures, at the cost of potentially higher FPR/h.

The proposed seizure detectors, specifically the CNN-TRF-BM-based model, can detect patient-independent seizures across various independent sEEG and iEEG datasets without retraining. It takes less than 15s computation time to detect seizures in a 30 minutes EEG. Therefore, the proposed detector can help reduce the time required to annotate seizures in EEGs in clinical settings. However, while the results are appealing for adult human EEG (and dog EEGs), there is room for improvement for neonatal EEG. One may need to perform additional tuning or retraining to achieve better performance for such cases.

**Discussion**

**Comparison with Existing Patient-independent Seizure Detectors**

To compare the proposed seizure detector to the state-of-the-art is challenging, as there is a lack of standardized evaluation metrics, datasets, or training and testing procedures for the problem of seizure detection. The datasets considered in the literature vary in terms of the following:

1. Patient: age group (neonate, pediatric, adult, elderly), type (human or animal), diversity (number of patients, age, gender, race, or ethnicity)
2. Clinical setting: in-patient or out-patient
3. EEG: recording device, type (sEEG or iEEG)
4. Data: quantity (number of EEGs, duration of the EEGs, number of EEG channels), quality (signal-to-noise (SNR) ratio and prevalence of artifacts in the EEG)
5. Annotation: number of annotators, experience of annotators, the granularity of the annotations (e.g., channel-level vs. segment-level annotations)
6. Use case: patient-specific versus patient-independent seizure detection.

It is especially critical to specify the use case. Patient-specific detectors would yield much better performance than their patient-independent counterparts, but cannot be deployed readily clinically. Therefore, comparing those two different types of detectors is meaningless. Consequently, we consider here studies that report results on patient-independent seizure detection on the six datasets analyzed in this paper.
Table 3. EEG-level seizure detection results for different CNN models evaluated with MOES across six EEG datasets.

| Dataset | CNN-TRF-BM | CNN-TRF-BC | ConvLSTM | LSTMAE | SVM | CNN | MOE | CNN-TRF-BM | CNN-TRF-BC | ConvLSTM | LSTMAE | SVM | CNN | MOE |
|---------|------------|------------|-----------|--------|-----|-----|-----|------------|------------|-----------|----------|--------|-----|-----|-----|-----|
| Adult   | 0.846      | 0.846      | 0.846     | 0.846  | 0.846 | 0.846 | 0.846 | 0.846      | 0.846      | 0.846     | 0.846   | 0.846 | 0.846 | 0.846 | 0.846 |
| CHB-MIT | 0.846      | 0.846      | 0.846     | 0.846  | 0.846 | 0.846 | 0.846 | 0.846      | 0.846      | 0.846     | 0.846   | 0.846 | 0.846 | 0.846 | 0.846 |
| Neonatal| 0.846      | 0.846      | 0.846     | 0.846  | 0.846 | 0.846 | 0.846 | 0.846      | 0.846      | 0.846     | 0.846   | 0.846 | 0.846 | 0.846 | 0.846 |
| sEEG    | 0.846      | 0.846      | 0.846     | 0.846  | 0.846 | 0.846 | 0.846 | 0.846      | 0.846      | 0.846     | 0.846   | 0.846 | 0.846 | 0.846 | 0.846 |

Table 4. EEG-level seizure detection results by the CNN-TRF-BM-based EEG-level detector evaluated with IMS.

| Dataset | SEN | PRE | aFPR/h | mFPR/h | Offset |
|---------|-----|-----|--------|--------|--------|
| Adult   | 0.979 | 0.979 | 0.520 | 0 | 10.542  |
| CHB-MIT | 0.979 | 0.979 | 0.520 | 0 | 10.542  |
| Neonatal| 0.979 | 0.979 | 0.520 | 0 | 10.542  |
| sEEG    | 0.979 | 0.979 | 0.520 | 0 | 10.542  |

Table 5. EEG-level seizure detection results by the CNN-TRF-BM-based EEG-level detector evaluated with IMS.

| Dataset | SEN | PRE | aFPR/h | mFPR/h | Offset |
|---------|-----|-----|--------|--------|--------|
| Adult   | 0.979 | 0.979 | 0.520 | 0 | 10.542  |
| CHB-MIT | 0.979 | 0.979 | 0.520 | 0 | 10.542  |
| Neonatal| 0.979 | 0.979 | 0.520 | 0 | 10.542  |
| sEEG    | 0.979 | 0.979 | 0.520 | 0 | 10.542  |

OUP, TAPS, and MOES across six EEG datasets.
| Dataset          | Seizure duration (in s) | Sensitivity (SEN) | Precision (PRE) | FP rate per hour (FPR/h) |
|------------------|-------------------------|-------------------|-----------------|-------------------------|
| TUH-SZ sEEG Adult|                         |                   |                 |                         |
| CHB-MIT sEEG Pediatric |                 |                   |                 |                         |
| SWEC-ETHZ iEEG Adult |                 |                   |                 |                         |
| HUH sEEG Neonatal |                         |                   |                 |                         |
| IEEGP iEEG Adult  |                         |                   |                 |                         |
| EIM iEEG Adult    |                         |                   |                 |                         |

**Figure 1.** EEG-level seizure detection results for the CNN-TRF-BM model across different datasets. (a) normalized histograms of TPs and FNs sorted by seizure duration; (b-d) normalized histograms of the sensitivity (SEN), precision (PRE), and false positive rate per hour (FPR/h) for individual EEGs, respectively.
Table 5. SEN of short (<10s) and long (>10s) seizures detected by the CNN-TRF-BM-based model across the six datasets according to MOES metric.

| Dataset | W  | SEN of short SZ | SEN of long SZ |
|---------|----|-----------------|----------------|
| TUH-SZ  | 3  | 0.772           | 0.532          |
|         | 5  | 0.751           | 0.548          |
|         | 10 | 0.671           | 0.334          |
|         | 20 | 0.671           | 0.334          |
| CHB-MIT | 3  | 0.653           | 0.421          |
|         | 5  | 0.571           | 0.334          |
|         | 10 | 0.678           | 0.343          |
|         | 20 | 0.679           | 0.343          |
| SWEC-ETHZ | 3 | 0.933           | 0.532          |
|         | 5  | 0.933           | 0.532          |
|         | 10 | 0.857           | 0.912          |
|         | 20 | 0.849           | 0.912          |
| HUH     | 3  | 0.545           | 0.318          |
|         | 5  | 0.253           | 0.123          |
|         | 10 | 0.227           | 0.123          |
|         | 20 | 0.254           | 0.123          |
| EIM     | 3  | 0.671           | 0.671          |
|         | 5  | 0.671           | 0.671          |
|         | 10 | 0.5             | 0.5            |
|         | 20 | 0.45            | 0.45           |
|         | 10 | 0.931           | 0.955          |
|         | 20 | 0.953           | 0.955          |

Seizure Detection on the TUH-SZ Dataset

Numerous patient-independent seizure detectors have been evaluated on the TUH-SZ dataset. Chatzichristos et al. designed an L TSM that achieved a SEN and FPR/h of 0.1237 and 0.06\(^{20}\). Roy et al. utilized different machine learning models and reported a SEN and FPR/h of 0.916 and 137.31\(^{21}\). Meanwhile, Shah et al. applied an L TSM to detect seizures at segment-level and obtained SEN between 0.33-0.37 and FPR/h between 1.24-20.8\(^{18}\). Ayodele et al. trained a VGGNet and evaluated it on 24 EEGs, attaining a SEN, FPR/h, and offset of 0.7835, 0.9, and 2.32s, respectively\(^{19}\).

Most results reported are not suitable for clinical application; either the SEN of those detectors is too low, or the FPR/h is too high. Additionally, most studies did not report the seizure evaluation metrics. When they do, they utilize EBS and OVLP metrics, which fail to appropriately represent the requirements of a seizure detector. In contrast, the proposed CNN-TRF-BM seizure detector achieved superior results calculated with MOES (SEN, PRE, aFPR/h, and mFPR/h of 0.772, 0.429, 0.425, and 0, respectively), which is suitable for clinical applications. To the best of the author’s knowledge, no existing studies have reported the PRE, although it is an essential metric in clinical practice. Moreover, only few studies reported the offset.

Seizure Detection on the CHB-MIT Dataset

Similarly, many patient-independent detectors proposed in the literature have been assessed on the CHB-MIT dataset. Generally, seizure detectors evaluated in a patient-independent manner on the CHB-MIT dataset reported poor performance, especially for certain patients. More details can be found in the supplementary notes.

In the following, we briefly review the results of the CHB-MIT dataset reported in the literature. Furbass et al. deployed epileptiform wave sequence (EWS) analysis to classify seizure features and obtained a SEN and FPR/h of 0.67 and 0.32, respectively\(^{22}\). Gomez et al. applied a CNN classifier and achieved a SEN, SPE, and FPR/h of 0.531, 0.931, and 7.8, respectively\(^{23}\). Ayodele et al. employed two datasets (CHB-MIT and TUH-SZ dataset) and reported a SEN, FPR/h, and offset of 0.745, 0.2, and 2.32s, respectively\(^{19}\). Mansouri et al. trained their detector on the CHB-MIT (with five patients removed) and the TUH-SZ (only 24 patients) dataset, and evaluated the detector on the CHB-MIT dataset\(^{23}\). They attained a SEN, SPE, and FPR/h of 0.83, 0.96, and 8, respectively\(^{23}\).

No studies reported the PRE nor offset.

The proposed CNN-TRF-BM model achieves better results on the CHB-MIT dataset, with SEN, PRE, aFPR/h, mFPR/h, and offset of 0.678, 0.377, 0.421, 0.118, and 4.684s, respectively. An important difference between our approach and the ones from the literature is that the proposed CNN-TRF-BM model is trained on the TUH-SZ dataset and tested on the CHB-MIT dataset. In contrast, the models from the literature are both trained and tested on the CHB-MIT dataset. The TUH-SZ dataset contains more seizures (3,055 events) compared to CHB-MIT (185 events), which may explain why the CNN-TRF-BM model outshines the state-of-the-art models. On the other hand, the training and test sets are different; therefore, it is not a priori clear that training on a larger but different dataset
helps improve the performance. The numerical results show that the model generalizes well across the different datasets and can take advantage of the larger number of seizures in the TUH-SZ dataset for training purposes.

Seizure Detection on the SWEC-ETHZ Dataset

No existing seizure detectors had been evaluated on the SWEC-ETHZ dataset in a patient-independent manner. However, two existing studies performed seizure detection on the SWEC-ETHZ dataset in a patient-specific manner. Burrello et al. utilized hyperdimensional computing for EEG-level seizure detection and obtained a SEN and SPE of 0.960 and 0.948, respectively45. Wang et al. trained a 1D CNN to detect seizures in multi-channel EEG segments and obtained a SEN, SPE, and ACC of 0.901, 0.998, and 0.997, respectively12. At the EEG-level, Wang et al. attained a SEN, FPR/h, and offset of 0.975, 0.07, and 13.2s. Overall, Wang et al. reported better results than Burrello et al., while the results obtained by Wang et al. more similar to the ones reported in the current study. However, we reiterate that Wang et al. detected seizures in a patient-specific manner, which is expected to yield better performance than a patient-independent approach. It is interesting to note that Wang et al. also evaluated their seizure detector on both sEEGs (from CHB-MIT) and iEEGs (from SWEC-ETHZ). They found that the same seizure detector pipeline could be applied to both sEEG and iEEG. However, they trained and tested on the sEEG and iEEG separately; in other words, they did not test their model on an independent test set. In our study, we trained the model on one dataset (TUH-SZ dataset) and tested it on five independent EEG datasets.

Seizure Detection on the HUH Dataset

No seizure detectors have so far been evaluated on the HUH dataset in a patient-independent manner. Existing studies only evaluated patient-specific seizure detection. O'Shea et al. detected seizures at individual channels, achieving an AUC between 0.955-0.95650. Similarly, Açıkoğlu et al. performed segment-level seizure detection and attained an ACC of 0.988. Consequently, the current study is the first to perform patient-independent seizure detection at EEG-level on the HUH dataset. Moreover, we applied a seizure detector trained on adult sEEGs to detect seizures in neonatal sEEGs, and attained promising results. The current study demonstrates that a seizure detector trained on adult seizures may capture neonatal seizures with a high PRE, despite the substantial age gap. We reiterate that the morphology of neonatal seizures differs vastly from adult seizures. As the model has been trained on adult sEEG, it struggles to detect seizures in neonatal sEEGs.

Seizure Detection on the IEEGP Dataset

Few studies investigated seizure detection on the IEEGP dataset. The biggest study on the IEEGP dataset is in the context of a competition on patient-specific seizure detection organized by the University of Pennsylvania and Mayo Clinic50. This study is centred on segment-level detection in 1s segments rather than EEG-level seizure detection. The top 5 models achieved AUCs between 0.954 to 0.963. However, AUC is a poor metric for an imbalanced dataset, as seen in the IEEGP dataset where the non-ictal class outnumbers the ictal class 10 to 1. Meanwhile, Shen et al. performed patient-specific segment-level seizure detection on the IEEGP dataset and obtained accuracies between 0.59 to 0.7951. Similarly, accuracy is a poor metric for an imbalanced dataset. Therefore, the current study can be the baseline for patient-independent seizure detection on the IEEGP dataset.

Seizure Detection on the EIM Dataset

No earlier studies on automated seizure detection have been conducted on the EIM dataset. The existing studies aim to predict surgical outcomes (positive or negative)52. The current study is the first to analyze the EIM dataset for patient-independent seizure detection.

Comparison with Commercial Seizure Detectors

Several commercial seizure detectors are available in the market, such as Persyst14, Encevis15, and BESA15. Commercial seizure detectors can assist clinicians in reviewing EEG recordings, which may help to reduce time and boost the annotation accuracy53. Earlier studies by Reus et al.16 and Koren et al.17 have compared the performance of Persyst, Encevis, and BESA. We summarized their findings against the performance of the proposed detector in Table 6. Reus et al. and Koren et al. evaluated the commercial seizure detectors on adult sEEG datasets; hence, we concentrate on the TUH-SZ dataset in this section.

Overall, the proposed model significantly outperforms the three commercial detectors in terms of higher SEN and lower FPR/h. The results for the proposed model are substantially better than for the three commercial systems in the study conducted by Reus et al. by a significant margin. The proposed system outperforms Persyst and BESA in the study by Koren et al., with Encevis reporting similar results to the current study. However, we report results for MOES, OVLP, and IMS metric, while Reus et al. and Koren et al. introduced an evaluation metric that seems to be more lenient. Indeed, Reus et al. consider a detection to be correct as long as the detection is within 30s before the
start or after the end of the seizure. This approach is called the increased margin scoring (IMS) metric. Koren et al. implemented an altered version of IMS, where the margin is increased to 120s. These two evaluation metrics are less stringent than OVLP, as it allows lopsided detection by introducing a margin for error. Hence, if we were to compute the performance based on their metrics, we would report even better results (see Table 4).

Table 6. Performance of commercial seizure detectors against the proposed CNN-TRF-BM detector.

| Author          | Dataset | Number of Patients | Number of EEG | Number of Seizures | Duration (in hours) | Seizure Metrics | SEN       | FPR/h   |
|-----------------|---------|--------------------|---------------|-------------------|--------------------|-----------------|-----------|---------|
| Reus et al.     | Private | 283                | 286           | 249               | 8771               | IMS             |           |         |
| Persyst 1.4.1   |         |                    |               |                   |                    | 0.588           | 0.071     |
| Encevis 1.9.2   |         |                    |               |                   |                    | 0.158           | 0.229     |
| BESA 2.0        |         |                    |               |                   |                    | 0.100           | -         |
| Koren et al.    | Private | 81                 |               | 790               | 6900               | IMS             |           |         |
| Persyst 1.4.1   |         |                    |               |                   |                    | 0.816           | 0.9       |
| Encevis 1.7     |         |                    |               |                   |                    | 0.778           | 0.2       |
| BESA 2.0        |         |                    |               |                   |                    | 0.766           | 0.7       |
| Current study   | TUH-SZ  | 637                | 5610          | 3055              | 922.673            | CNN-TRF-BM     | 0.772     | 0.325   |
| MOES            |         |                    |               |                   |                    | CNN-TRF-BM     | 0.75      | 0.423   |
| OVLP            |         |                    |               |                   |                    | CNN-TRF-BM     | 0.75      | 0.423   |
| IMS             |         |                    |               |                   |                    | CNN-TRF-BM     | 0.797     | 0.412   |

Transformer for Seizure Detection

We identified two studies that apply transformers for seizure detection. In other words, these systems studies did not implement a channel-level detector. Those systems do not analyze individual channels separately, but instead, only consider multi-channel EEG segments. The current study is the first to implement a channel-level seizure detector by means of transformers.

Bhattacharya et al. utilized a transformer for patient-specific seizure detection on the CHB-MIT and IEEGP dataset. For the CHB-MIT and IEEGP datasets, they attained an average SEN of 0.985 and 0.948, and FPR/h of 0.124 and 0, respectively. Overall, there are several differences in the study performed by Bhattacharya et al. as compared to the current study:

- We followed a patient-independent approach, while Bhattacharya et al. designed a patient-specific detector.
- The proposed system can detect seizures at individual channels, while the systems in Bhattacharya et al. can only detect seizures in multi-channel segments.
- We implemented BM loss to enhance generalization and calibration. Meanwhile, Bhattacharya et al. utilized the traditional SM loss.
- We performed seizure detection on all available EEGs. In contrast, Bhattacharya et al. cropped the EEGs to 30 minutes and discarded the remaining recordings. Consequently, they rejected a significant amount of EEG data from the CHB-MIT dataset, as all the EEGs in this dataset are at least 1 hour long.

As Bhattacharya et al. detected seizures in a patient-specific manner, it is not an appropriate benchmark for the models proposed in the current study. Nonetheless, we achieve decent SEN with low FPR/h on both datasets for patient-independent seizure detection. However, the results yielded from the IEEGP dataset were much poorer.

On the other hand, Pedoeem et al. trained and evaluated a patient-independent transformer-based seizure detector on the TUH-SZ dataset. They computed their EEG-level results according to three seizure evaluation metrics: EBS, OVLP, and TAES. Pedoeem et al. attained the best results with OVLP metric, i.e., SEN and FPR/h of 0.09 and 1.301, respectively. However, their results are inferior to the ones achieved by the proposed CNN-TRF-BM model (SEN of 0.775, PRE of 0.43, AFPR/h of 0.423, and mFPR/h of 0, computed with OVLP). A possible explanation for the significant performance improvement could be the fact that the proposed systems detect, in the first stage, seizures at individual channels. Such spatial information might boost the accuracy of seizure detection.

Training the Seizure Detector on the TUH-SZ dataset Only

Patient-independent seizure detectors that do not require retraining or optimization can be readily deployed, which is convenient for clinical practice. In this study, we only trained the seizure detectors on the TUH-SZ dataset to replicate this scenario. By contrast, most existing studies utilize the same dataset for training and testing, and retrain a separate model for each dataset. Earlier, we showed that the proposed seizure detectors yield good performance on six EEG datasets. However, when testing on an independent dataset, we do not know whether the model trained on the TUH-SZ dataset would yield better performance than a model trained on the test dataset itself.

To address this question, we train and test the seizure detectors on the same dataset, i.e., the CHB-MIT dataset. We selected the CHB-MIT dataset as the total length of EEG in that dataset is comparable to the total duration of
As previously mentioned, most seizure detectors proposed in the literature deployed a similar pipeline. Most existing seizure detectors do not analyze EEG signals at individual channels; therefore, they cannot localize seizures at individual channels. Instead, they first process multi-channel EEG segments, then proceed to detect the start and end points of seizures in the EEG. The main innovation in those studies lies in improving the deep neural networks that classify the multi-channel EEG segments as ictal or normal. These deep neural networks typically contain numerous layers (often 10+) and millions of parameters. Such models require substantial computational power for training and testing. Moreover, such networks tend to overfit to specific datasets, leading to poor generalization capability. We explore whether deeper models lead to better seizure detection performance.

In Table 8, we list existing deep learning systems for seizure detection, and provide information about their complexity as well as the datasets they were trained on. The complexity of these models varies widely, with some requiring significant resources for training and others producing sub-optimal results. We conclude that there is a clear need for further research into the development of more efficient and generalizable models for seizure detection.
Figure 2. Performance of various seizure detectors as a function of (a) parameters (in millions) and (b) layers in the deep learning model. Each red indicates a model (all three models) deployed in the current study on the TUH-SZ dataset, while each black dot denotes a model in literature. In all the plots, the x-axis is in the logarithmic scale of base 2. We displayed all nine performance metrics, as studies tend to report different metrics, making comparison difficult.

Next, we examined the correlation between model size and performance for seizure detection. Due to the vast variability in the metrics reported by various studies, it is challenging to make a comparison. Hence, we plotted all nine metrics (AUC, AUPRC, ACC, BAC, SEN, SPE, PRE, F1, and FPR/h) versus the number of parameters (in millions) and the number of layers in the neural networks in Figure 2. Generally, AUC, BAC, AUPRC, SPE, and F1 scores are used in segment-level analysis, while the SEN, PRE, and FPR/h are used in EEG-level detection. The proposed seizure detector models reported higher SEN and lower FPR/h than most models with more parameters and layers. The AUC, ACC, BAC, and F1 were comparable, while the SPE was poorer in our model. However, SPE is only computed in segment-level classification, which is not an EEG-level detection metric. Moreover, the proposed models obtained better AUPRC, SEN, and FPR/h than most existing models with fewer parameters and layers. These studies deployed different metrics, not limited to EBS, IMS, OVLP, and TAES. Nevertheless, we only reported the results with MOES in Figure 2, as MOES is more appropriate as an evaluation metric. However, we can refer to Table 4 for results obtained from other metrics (EBS, IMS, OVLP, and TAES).

Overall, the proposed models outshine models with vastly more parameters. Therefore, this study suggests that designing ever-bigger neural networks for patient-independent seizure detection may not be a fruitful avenue for research. Instead, alternative pipelines with substantially fewer parameters may perform comparably to the state-of-the-art or even better. In this study, we demonstrated in particular that by first detecting seizures at individual channels, one can not only localize seizures at specific channels but also vastly reduce the number of parameters (to about 4 million in total) while achieving the same or increased level of performance.

Benefits of Channel-level Seizure Detection
In this study, we proposed to detect seizures starting from single-channel segments (channel-level detection). However, many seizure detectors in the literature detect seizures directly from multi-channel segments (segment-level detection).
Table 8. Comparison of the proposed seizure detector models to existing models from the literature in terms of complexity and detection performance.

| Author | Model | Layers | Parameters (in millions) | Input Size | AUC | AUPRC | ACC | BAC | SEN | SPE | F1 | FPR/h |
|--------|-------|--------|--------------------------|------------|-----|-------|-----|-----|-----|-----|----|-------|
| Asif et al. 23 | SeizureNet | 133 | 45.94 | 190,528 | - | - | - | - | - | - | 0.896 | - |
| Raghu et al. 22 | AlexNet | 25 | 62 | 151,529 | - | 0.769 | - | - | - | - | - | 0.853 | - |
| | VGG16 | 41 | 138 | 50,176 | - | 0.833 | - | - | - | - | - | - | - |
| | VGG19 | 47 | 138 | 50,176 | - | 0.818 | - | - | - | - | - | - | - |
| | SqueezeNet | 68 | 1.2 | 151,529 | - | 0.851 | - | - | - | - | - | - | - |
| | GoogleNet | 144 | 7 | 50,176 | - | 0.747 | - | - | - | - | - | - | - |
| | Inceptionv3 | 316 | 24 | 89,401 | - | 0.883 | - | - | - | - | - | - | - |
| | DenseNet201 | 709 | 20 | 50,176 | - | 0.851 | - | - | - | - | - | - | - |
| | ResNet18 | 72 | 11 | 50,176 | - | 0.862 | - | - | - | - | - | - | - |
| | ResNet50 | 177 | 23 | 50,176 | - | 0.862 | - | - | - | - | - | - | - |
| | ResNet101 | 347 | 29.4 | 50,176 | - | 0.861 | - | - | - | - | - | - | - |
| Covert et al. 23 | TCGC | 30 | 5.5 | 415,107 | 0.926 | - | - | 0.809 | 0.648 | 0.970 | - | - |
| | TGCN | 26 | 5 | 415,107 | 0.935 | - | - | 0.808 | 0.645 | 0.970 | - | - |
| | TGCN | 24 | 4.5 | 415,107 | 0.927 | - | - | 0.808 | 0.645 | 0.970 | - | - |
| | TGCN | 20 | 4 | 415,107 | 0.917 | - | - | 0.812 | 0.653 | 0.970 | - | - |
| | TGCN | 16 | 3.5 | 415,107 | 0.931 | - | - | 0.820 | 0.669 | 0.970 | - | - |
| | TGCN | 12 | 3 | 415,107 | 0.928 | - | - | 0.803 | 0.635 | 0.970 | - | - |
| Hossain et al. 26 | CNN | 4 | 0.04 | 11,648 | - | - | - | 0.908 | 0.880 | 0.917 | - | - |
| | VggNet | 6 | 0.34 | 17,664 | 0.907 | 0.944 | - | - | - | - | - | - |
| | Zhen et al. 28 | CNN | 3 | 0.4 | 5,888 | - | - | - | 0.595 | 0.616 | 0.742 | - | - |
| | ChromaNet | 10 | 12.7 | 40,688 | 0.940 | - | - | 0.773 | 0.650 | 0.890 | - | - |
| | Emani et al. 27 | VGG16 | 41 | 138 | 50,176 | - | - | - | 0.740 | - | - | 0.260 | - |
| | Anti et al. 27 | CNN | 23 | 0.0076 | 54,000 | 0.830 | - | - | 0.770 | - | - | 0.900 | - |
| Khadkikar et al. 29 | ResNet | 72 | 11 | 65,566 | - | - | - | 0.792 | - | - | - | 0.241 |
| | CNN-LSTM | 7 | 0.5 | 77,440 | - | 0.648 | 0.308 | 0.969 | - | - | - | 0.281 |
| | CNN-MLP | 7 | 0.5 | 77,440 | - | - | - | 0.580 | 0.391 | 0.768 | - | - |
| Shah et al. 28 | IPCA-LSTM | 2 | 0.5 | 77,440 | - | 0.553 | 0.330 | 0.776 | - | - | - | 3.063 |
| | HM1-LSTM | 2 | 0.5 | 77,440 | - | 0.513 | 0.301 | 0.865 | - | - | - | 2.538 |
| | Him-LMA | 2 | 0.5 | 77,440 | - | 0.564 | 0.354 | 0.734 | - | - | - | 3.225 |
| Gomez et al. 25 | CNN | 12 | 0.314 | 24,504 | 0.927 | 0.929 | 0.741 | 0.598 | 0.931 | 0.541 | 0.461 | 7.800 |
| Yang et al. 28 | L-STM | 6 | 0.4 | 34,038 | - | 0.920 | - | - | - | - | - | 0.267 |
| Bhatnagar et al. 27 | L-STM | 6 | 0.4 | 199,009 | - | 0.939 | - | - | - | - | - | - |
| Chen et al. 27 | U-Net | 10 | 5.88 | 74,830 | - | 0.720 | - | - | 0.676 | - | - | - |
| | CNN | 7 | 0.16 | 384 | - | - | - | 0.713 | - | - | - | 0.287 |
| | CNN | 7 | 0.26 | 640 | - | - | - | 0.701 | - | - | 0.497 | 0.578 |
| | CNN | 7 | 0.52 | 1,280 | - | - | - | 0.701 | - | - | 0.512 | 0.592 |
| | CNN | 7 | 1 | 2,560 | - | - | - | 0.578 | - | - | - | 0.592 |
| Current study | CNN-TRF | 15 | 2.3 | 384 | - | - | - | 0.712 | - | - | 0.429 | 0.552 |
| Current study | CNN-TRF | 15 | 2.5 | 640 | - | - | - | 0.653 | - | - | 0.476 | 0.551 |
| Current study | CNN-TRF | 15 | 2.8 | 1,280 | - | - | - | 0.671 | - | - | 0.534 | 0.595 |
| Current study | CNN-TRF | 15 | 3.5 | 2,560 | - | - | - | 0.655 | - | - | 0.520 | 0.580 |

This approach usually requires a fixed number of channels, an important limiting factor for clinical practice where the number of channels may vary depending on the EEG cap and machine. Moreover, the clinical setting may vary (inpatient vs. outpatient), while EEG caps for inpatient and outpatient recordings typically have a different number of electrodes. Moreover, this approach may have poor generalization performance, as the system might be strongly overfitted to a particular EEG electrode configuration and dataset. Here, we refer to seizure detector pipelines that detect seizures starting from single-channel segments as 1D models, while detectors that detect seizures starting from multi-channel segments as 2D models.

To further evaluate the benefits of the three proposed 1D seizure detectors, we designed two 2D seizure detectors that directly perform segment-level classification from the EEG signals. Those 2D models are identical to the 1D ones, except that the 1D convolutional filters are replaced with 2D filters. We optimized the 2D CNNs on the T UH SZ dataset with the SM and BM loss, leading to the two different 2D seizure detectors. As all the sEEGs in the T UH SZ dataset contain 20 common channels, we fixed the number of channels to 20. Hence, the input of the 2D models on the CHB-MIT dataset, and compared them to the models trained on the T UH SZ dataset.

We display the segment- and EEG-level results for the T UH SZ and CHB-MIT datasets in Table 9, where the EEG-level results are computed by the MOES metric. When trained and evaluated on the T UH SZ dataset, the 2D models attain much weaker results for both segment- and EEG-level classification than the 1D models (see Table 3
for comparison). Those models also perform poorly on the CHB-MIT dataset, leading to substantially lower SEN and PRE scores than the 1D models. Moreover, when the 2D models were trained and evaluated on the CHB-MIT dataset, we obtained the worst results thus far, with PRE lower than 5% for all cases. These numerical results are in line with many 2D models reported in the literature. Overall, 2D models underperform compared to the 1D models by a considerable margin. In conclusion, the channel-level detector appears to be vital for achieving superior generalization performance. Moreover, with the channel-level outputs, as seen in Supplementary Figure 6, we can identify the channels that exhibit the seizure by tracing the channels with the highest seizure probabilities. This would be impossible with a direct segment-level detection approach, as in the 2D models.

| Table 9. Results of 2D seizure detectors on the TUH-SZ and CHB-MIT dataset. |
|-------------------------------------------------|
| **Testing Dataset** | **Training Dataset** | **Model** | **W** | **Segment-level** | **EEG-level** |
|---------------------|---------------------|-----------|-------|-------------------|---------------|
| **TUH-SZ EEG**      | **TUH-SZ**          | 2D        | 3     | 0.106 0.769 0.722 0.717 0.827 0.779 | 0.544 0.659 0.463 2.555 0 9.125 |
| Adult               | CNN-SM              | 5         | 0.119 0.791 0.769 0.672 0.866 0.788 | 0.538 0.674 0.448 2.764 0 7.000 |
|                     |                     | 10        | 0.149 0.849 0.751 0.566 0.937 0.842 | 0.530 0.656 0.444 3.020 0 7.500 |
|                     |                     | 20        | 0.160 0.859 0.734 0.534 0.933 0.854 | 0.510 0.521 0.499 1.741 0 3.250 |
| **TUH-SZ EEG**      | **TUH-SZ**          | 2D        | 3     | 0.106 0.801 0.805 0.753 0.857 0.801 | 0.774 0.625 0.569 2.979 0 5.250 |
| Adult               | CNN-BM              | 5         | 0.119 0.816 0.791 0.678 0.904 0.812 | 0.584 0.668 0.519 2.068 0 7.875 |
|                     |                     | 10        | 0.149 0.857 0.782 0.635 0.929 0.854 | 0.559 0.658 0.486 2.402 0 3.750 |
|                     |                     | 20        | 0.160 0.868 0.711 0.468 0.955 0.856 | 0.528 0.506 0.553 1.499 0 5.250 |
| **CHB-MIT Paediatric sEEG** | **CHB-MIT** | 2D        | 3     | 0.173 0.717 0.729 0.681 0.797 0.71 | 0.078 0.520 0.422 2.221 0 13.875 |
| sEEG                | CNN-SM              | 5         | 0.143 0.732 0.748 0.638 0.856 0.714 | 0.079 0.503 0.434 3.120 0 25.00 |
|                     |                     | 10        | 0.122 0.783 0.753 0.639 0.873 0.765 | 0.076 0.503 0.441 2.894 0 16.00 |
|                     |                     | 20        | 0.168 0.872 0.786 0.636 0.960 0.861 | 0.086 0.509 0.474 2.947 0 16.250 |
| **CHB-MIT Paediatric sEEG** | **CHB-MIT** | 2D        | 3     | 0.165 0.716 0.742 0.687 0.776 0.707 | 0.082 0.510 0.444 2.221 0 7.500 |
| sEEG                | CNN-BM              | 5         | 0.145 0.733 0.748 0.655 0.841 0.716 | 0.061 0.505 0.322 2.951 0 7.69 |
|                     |                     | 10        | 0.112 0.779 0.75 0.633 0.868 0.759 | 0.063 0.508 0.034 2.892 0 10.23 |
|                     |                     | 20        | 0.16 0.868 0.783 0.635 0.932 0.858 | 0.067 0.510 0.036 2.692 0 8.225 |
| **CHB-MIT Paediatric sEEG** | **TUH-SZ** | 2D        | 3     | 0.233 0.584 0.662 0.365 0.959 0.547 | 0.303 0.439 0.231 0.391 0 2.40 |
| sEEG                | CNN-SM              | 5         | 0.159 0.677 0.680 0.429 0.931 0.646 | 0.292 0.626 0.190 1.372 0 4.65 |
|                     |                     | 10        | 0.148 0.744 0.714 0.290 0.981 0.690 | 0.383 0.536 0.298 0.821 0 2.42 |
|                     |                     | 20        | 0.083 0.843 0.658 0.365 0.951 0.798 | 0.370 0.376 0.365 0.113 0 5.184 |
| **CHB-MIT Paediatric sEEG** | **TUH-SZ** | 2D        | 3     | 0.383 0.524 0.541 0.084 0.998 0.416 | 0.115 0.078 0.218 0.072 0 0.395 |
| sEEG                | CNN-BM              | 5         | 0.383 0.524 0.541 0.084 0.998 0.416 | 0.126 0.368 0.505 0.129 0 5.126 |
|                     |                     | 10        | 0.156 0.750 0.648 0.307 0.989 0.696 | 0.441 0.524 0.380 0.396 0 0.05 |
|                     |                     | 20        | 0.068 0.853 0.691 0.341 0.987 0.817 | 0.474 0.461 0.488 0.183 0 8.526 |

Conclusion

This study proposed patient-independent seizure detectors that identify seizures on three EEG scales (see Supplementary Figure 1): single-channel EEG segments (channel-level detection), multi-channel EEG segments (segment-level detection), and entire EEGs (EEG-level detection). Firstly, the channel-level detectors detect seizures in single-channel segments through a CNN-based deep learning model (CNN-SM, CNN-BM, or CNN-TRF-BM). We perform channel-level detection at all channels of multi-channel EEG segments. Next, we extract statistical features from the channel-level outputs based on different scalp regions. Then, we apply a machine learning model to classify the segment-level features as normal or ictal. At last, we apply post-processing filters to the segment-level outputs to determine the start and end times of any detected seizures.

We trained and tested the proposed seizure detectors on the TUH-SZ sEEG dataset, before evaluating the pretrained detectors on five independent sEEG and iEEG datasets without retraining. We introduced a new metric named MOES to measure the EEG-level seizure detection performance, and compare it to existing metrics. MOES addresses some shortcomings of the latter. To the best of the author’s knowledge, this study is one of the first to incorporate a channel-level detector within the seizure detection system. Moreover, we implemented a pipeline that can detect seizures in both sEEGs and iEEGs with any number of electrodes. Furthermore, we demonstrated that a channel-level detector is essential for reliable seizure detection and boosting the generalization performance. Finally, the proposed seizure detector is computationally efficient, with a computation time of less than 15s for a 30 minutes EEG. Hence, the detector may help accelerate and improve EEG annotation in clinical practice.

In future work, we will address the problem of detecting artifacts before seizure detection. The artifact detector will be designed to reduce FPR/h and improve PRE of the seizure detector. Consequently, it might be able to reject artifacts without eliminating important cerebral signals, such as slow waves, sharp waves, and seizures in EEGs.
Methods

Dataset

We analyze six public sEEG and iEEG datasets in this study:

1. Temple University Hospital Seizure (TUH-SZ) dataset\(^{61}\)
2. Children’s Hospital Boston Massachusetts Institute of Technology (CHB-MIT) dataset\(^{62}\)
3. Sleep Wake Epilepsy Center at ETH Zurich (SWEC-ETHZ) dataset\(^{48}\)
4. Helsinki University Hospital (HUH) dataset\(^{63}\)
5. International Epilepsy Electrophysiology Portal (IEEGP) dataset\(^{64}\)
6. Epilepsy iEEG Multicenter (EIM) dataset\(^{52}\)

Table 10. Information on the six sEEG and iEEG datasets analyzed in the study.

| Information                  | Details | TUH-SZ | CHB-MIT | SWEC-ETHZ | HUH | IEEGP | EIM |
|------------------------------|---------|--------|---------|-----------|-----|-------|-----|
| Patient Type                 | Patient Age Group | Human | Human | Human | Human | Human/Adult | Human | Adult |
| EEG Details                  | EEG Type | sEEG | sEEG | iEEG | iEEG | iEEG | iEEG |
| F_s (Hz)                     | 250-1000 | 256 | 512 | 256 | 400-5000 | 250-1000 |
| Channel Name                 | Available | Available | Unavailable | Available | Unavailable | Unavailable |
| Channel-level Annotation     | Yes | No | No | No | No | No |
| Seizure Label, Type          | Yes, 8 | No | No | No | No | No |
| No of Channels               | 19.21 | 23.24.26 | 36-100 | 21 | 16-72 | 53-216 |
| Number of Patients and EEGs  | Patients | 637 | 24 | 10 | 75 | 12 | 31 |
| Duration                     | All EEGs | 5,610 | 684 | 100 | 75 | 12 | 102 |
|                            | Non-Seizure EEGs | 4,450 | 545 | 0 | 22 | 0 | 0 |
|                            | Seizure EEGs | 1,150 | 138 | 100 | 54 | 12 | 102 |
|                            | Seizure Events | 3,050 | 185 | 100 | 517 | 12 | 102 |
|                            | All EEGs (in hours) | 922 | 980 | 13.5 | 114 | 7.20 | 7.96 |
|                            | Non-SZ EEGs (in hours) | 681 | 792 | 0 | 35.0 | 0 | 0 |
|                            | SZ EEGs (in hours) | 242 | 388 | 13.5 | 78.6 | 7.20 | 7.96 |
|                            | Average (All) (in minutes) | 9.84 | 86.1 | 8.1 | 89.64 | 36 | 4.68 |

Information about the six datasets is summarized in Table 10. We refer to the supplementary notes for more details on the six datasets. The TUH-SZ dataset is the largest among those six datasets, with the most annotated seizure events and eight seizure types (see Supplementary Figure 2). As that dataset contains the most seizure data, we utilize it as the primary source to train the entire seizure detector pipeline. Firstly, the seizure detector is trained and evaluated with the TUH-SZ dataset via 4-fold cross-validation (CV), where we assign approximately the same number of patients and seizures to each of the four-folds. We train the proposed seizure detector on the TUH-SZ dataset, and assess it on the other five independent EEG datasets. In this way, we can examine the generalizability of the detector on different EEG datasets with different EEG types (sEEG and iEEG) and patient age groups (neonates, paediatrics, and adults).

For all the EEGs, a 4\(^{th}\) order Butterworth notch filter at 60Hz (USA) and 50Hz (EU) is applied to remove electrical interference\(^{65}\). Next, a 1Hz high-pass filter (4\(^{th}\) order) is implemented to reject DC shifts and baseline fluctuations\(^{66}\). Finally, all the EEGs are downsampled to a sampling frequency F_s of 125Hz. At last, we convert all sEEGs to bipolar montage, as the TUH-SZ dataset is annotated in the bipolar montage. As the montage for the iEEGs is not compatible with the bipolar montage, we keep the montage of the iEEGs at monopolar.

Seizure Detector Pipeline

We perform seizure detection first at individual channels (channel-level detection), next at multi-channel segments (segment-level detection), and at last, we detect the start and end points of the seizures in the entire multi-channel EEG (EEG-level detection)\(^{65–67}\) (see Supplementary Figure 1). The pipeline of the proposed seizure detector is displayed in Figure 3. The pipeline consists of a channel-level deep learning classifier, segment-level machine learning classifier, and multiple EEG-level post-processing modules. The seizure detectors are implemented on NVIDIA GeForce GTX1080 GPUs in Keras 2.2.0 and TensorFlow 2.6.0.

Channel-level Seizure Detector

The channel-level seizure detector computes the seizure probability for single-channel EEG segments. For seizure detection, the window length W adopted in the literature ranges between 1s to 30s. However, we believe that 1s is too short to capture long-range seizure morphology, while 30s is too long to capture short seizures. Therefore, we
tested window lengths $W$ of 3s, 5s, 10s, and 20s. In this study, we deploy three channel-level seizure detectors based on convolutional neural networks (CNN):

1. CNN with softmax (SM) loss: CNN-SM
2. CNN with belief matching (BM) loss: CNN-BM
3. CNN with transformer and BM loss: CNN-TRF-BM.

**CNN-SM Model**
The CNN-SM model is a typical CNN with a SM loss function. The input of the CNN-SM model is the raw single-channel signal of length $W \times F_s$. The CNN architecture contains 5 convolutional layers with 8, 16, 32, 64, and 128 filters, respectively, with two fully connected layers. The architecture is summarized in Supplementary Table 2. To minimize the loss, we applied the Adam optimizer$^{58}$ with an initial learning rate equal to $10^{-4}$. The batch size during training is set to 1000. Also, we implemented class weights that are inversely proportional to the class frequency in the training data during training. This allows us to optimize the loss function on an imbalanced dataset without overfitting$^{37}$. Finally, we optimized parameters within the CNN via nested CV on the training data, with an 80:20% split for training and validation.

**CNN-BM Model**
The CNN-BM model has the same architecture as the CNN-SM model, except that the BM loss replaces the SM loss. Additional details of the BM loss can be found in the supplementary methods. We implemented the BM loss as it has shown to improve uncertainty estimation and calibration compared to the traditional SM loss$^{39}$. Additionally, the BM loss tends to improve generalization performance, which is an important property required for seizure detection.

**CNN-TRF-BM Model**
The CNN-TRF-BM model contains two components: the CNN and the transformer. The architecture is the same as in the CNN-BM model, but we insert an additional transformer encoder between the final convolutional layer and the flattening layer (see Figure 4(a) and (b)). We implemented a transformer in tandem with the CNN, as the CNN alone cannot model correlations between distant data points, such as seizure morphologies. The transformer
can compensate for this limitation by extracting long-range information from the CNN features. The transformer encoder contains eight heads, and the number of hidden layer neurons in the forward feed network (FFN) is 1024. As input to the transformer, we extract 1s segments with 25% overlap from the W-second single-channel segment. More details on the architecture of the transformer can be found in the supplementary methods.

Segment-level Seizure Detector
Next, we rely on the outputs of the channel-level detectors to detect seizures in multi-channel segments. The channel-level detectors yield seizure probabilities for each EEG channel, which we arrange into regions according to the scalp topology: frontal, central, occipital, and parietal. Besides those four local regions, we also define a "global" region containing all channels. From each region, we extract seven statistical features: mean, median, standard deviation, maximum value, minimum value, and value at 25% and 75% percentile. As there are five regions, we extract $5 \times 7 = 35$ features. From all channel-level outputs, we compute the normalized histogram features (5 bins, range set at [0,1]), and include them into the feature set, bringing the total features to 40.

In the iEEGs, the channel locations are unavailable. Therefore, we cannot group the iEEG channels into local regions. For consistency, we replace the four local regions with the global region. In this scenario, only 12 features are unique, and the remaining ones are duplicates. In any case, the number of segment-level features is 40, regardless of the total number of channels or the availability of the channel locations. This approach ensures that the number of features is consistent during the training and evaluation of any dataset. The features will be the inputs to a machine learning classifier for training and validation. We employed six machine learning models as candidates for the segment-level classifier: logistic regression (LR), support vector machine (SVM), gradient boosting (GB), AdaBoost (AB), random forest (RF), and XGBoost (XGB). We determined the best model and hyperparameters via grid search CV.

Channel- and Segment-level Evaluation Metric
We assess the channel- and segment-level seizure classifiers through the following metrics: accuracy (ACC), balanced accuracy (BAC), sensitivity (SEN), specificity (SPE), F1 score (F1), and expected calibration error (ECE). As the seizure and non-seizure classes are imbalanced, we evaluate the results mainly in terms of BAC.

EEG-level Seizure Detector
Finally, we perform seizure detection on full multi-channel EEGs (EEG-level detection). Specifically, we determine the start and end time of the seizures, if any. First, we apply a sliding window of length $W$ with an overlap duration...
After collecting all the start and end times of the detected seizures, we assess the accuracy of those detections. The first option is too lenient in practice, while the latter is too strict.

The minimum overlap evaluation scoring (MOES) determines the overlap duration $T_{\text{overlap}}$ between the detection ($T_{\text{detection}} = [d_{\text{start}}, d_{\text{end}}]$) and seizure ($T_{\text{seizure}} = [s_{\text{start}}, s_{\text{end}}]$) window, and vice versa, before deciding if the detection is correct or the seizure is captured. Based on existing literature, only seizures of at least 10s are annotated typically\textsuperscript{71,72}. Therefore, the minimum overlap duration of the detection(s) with the seizure should be 10s. However, these criteria do not account for the duration of the seizure nor the detection. Even if the detection correctly detected over 10s of a seizure, if the majority of the detection did not capture any seizure, the system should be penalized. To resolve this, we compute the detection overlap (DOL) and the seizure overlap (SOL), which measures the fraction of the detection that overlaps with any seizures, and vice versa, as:

$$DOL_i = \frac{\sum_T T_{\text{overlap},s,i}}{d_{\text{end},i} - d_{\text{start},i}}, \quad (1)$$

$$SOL_j = \frac{\sum_T T_{\text{overlap},d,j}}{s_{\text{end},j} - s_{\text{start},j}}, \quad (2)$$

where $i$ and $j$ is the index of a detection and a seizure, respectively, $\sum T_{\text{overlap},s,i}$ is the sum of all the overlaps with any seizures with detection $i$, and $\sum T_{\text{overlap},d,j}$ is the sum of all the overlaps with any seizures with seizure $j$.

In this study, we set a minimum DOL and SOL of 0.3 (30%), to ensure that a significant portion of the detection overlaps with the seizures and vice versa. In OVLP metric, the DOL is set to be 0+%, while in TAES it is 100%. The first option is too lenient in practice, while the latter is too strict.

A high DOL implies that the detection overlaps well with the seizure(s). Meanwhile, a high SOL indicates that the seizure is well captured by the detection(s). If the DOL is low, the detection should be discarded and treated as a false positive (FP). Similarly, if the SOL is low, the seizure should be treated as a false negative (FN). This approach allows us to consider different cases (see Figure 5):
Case 1: The detection window encapsulates the seizure window almost perfectly. Therefore, the SOL\(_j\) = 1, while DOL\(_i\) > 0.3 (close to 1). In this case, the seizure\(_j\) is a TP as the detection\(_i\) is correct.

Case 2: The detection window is encapsulated by the seizure window almost perfectly. Therefore, the DOL\(_i\) = 1, while SOL\(_j\) > 0.3 (close to 1). In this case, the seizure\(_j\) is a TP as the detection\(_i\) is correct.

Case 3: The detection window overlaps with the seizure window, however, the detection window protrudes the seizure window by a significant margin. In this case, while the SOL\(_j\) can be greater than 0.3, the DOL\(_i\) is low (less than 0.3). As DOL\(_i\) < 0.3, we consider the detection\(_i\) as a false alarm (FP). As a result, the seizure\(_j\) is considered as a FN.

Case 4: The seizure window overlaps the detection window, however, the seizure window protrudes the detection window by a significant margin. In this case, while the DOL\(_i\) can be greater than 0.3, the SOL\(_j\) is low (less than 0.3). As SOL\(_j\) < 0.3, we consider the seizure\(_j\) as missed (FN). As a result, the detection\(_i\) is considered as a FP.

Case 5: Multiple detection windows (1, 2, 3) overlap with the annotated seizure. The majority of the seizure\(_j\) is detected, hence the SOL\(_j\) is high (greater than 0.3). However, the DOL\(_i\) vary for each detection, though all of them clipped the seizure to a certain extent.

1. Detection\(_1\) would have DOL\(_1\) ≈ 0.5, hence it is a correct detection.
2. Detection\(_2\) would have DOL\(_2\) = 1, hence it is a correct detection.
3. Detection\(_3\) would have DOL\(_3\) < 0.3, hence it is a false detection.

As SOL\(_j\) > 0.3, we consider the seizure\(_j\) as captured, hence a true positive. Meanwhile, the detection\(_1\) and detection\(_2\) are correct (TP) and detection\(_3\) is considered as a false positive.

Case 6: Multiple seizure windows (1, 2, 3) overlap with a detection window. The majority of the detection\(_i\) had capture seizures, hence the DOL\(_i\) is high (greater than 0.3). However, the SOL\(_j\) vary for each seizure, though all of them clipped the detection to a certain extent.

1. Seizure\(_1\) would have SOL\(_1\) ≈ 0.5, hence the seizure is detected well.
2. Seizure\(_2\) would have SOL\(_2\) = 1, hence the seizure is detected well.
3. Seizure\(_3\) would have SOL\(_3\) < 0.3, hence the seizure is not detected.

As DOL\(_i\) > 0.3, we consider the detection\(_i\) as correct, hence it is not a false positive. Meanwhile, the seizure\(_1\) and seizure\(_2\) are one TP each and detection\(_3\) is missed and is considered as a false negative.

Figure 5. Different cases of seizure detection that may be encountered.
properly detected. The two detections could have very low DOL. Consequently, the seizure in that EEG will be considered a FN instead of a TP.

Next, by investigating the detections, we compute the FPs. We first determine what seizures overlap with detection. The detection is considered a FP, as long as any of the following two conditions are met:

1. $DOL_i < 0.3$
2. $SOL_j < 0.3$ for all seizures overlapping with detection $i$.

Note that it is important to compute TPs from the perspective of the seizure. Indeed, multiple detections may overlap with the same seizure (see Figure 5 Case 6). However, as there is only one seizure event, we only can have one TP or one FN associated with a seizure event. Therefore, we need to compute the TPs from the perspective of the seizures. Computing TP from the perspective of the detection windows may result in multiple TPs for a single seizure event, which is undesirable.

Finally, the detection may start earlier or later than the actual annotation. Hence, one should compute the detection offset. We compute the detection offset as:

$$T_{\text{offset}} = d_{\text{start}} - s_{\text{start}} + W,$$

where $W$ are the duration of the window length, $d_{\text{start}}$ is the start time of the detection, $s_{\text{start}}$ is the start time of the annotated seizure. We added $W$ in the offset as we require a minimum window of length $W$ to detect seizures. To more accurately detect the onset of a seizure, one may slide the window in smaller steps around the onset of a detection. However, this goes beyond the scope of this work, as we are mainly interested in detecting seizures, irrespective of their onset times.

**EEG-level Seizure Detection Performance Metrics**

We measure the performance of EEG-level seizure detection by the sensitivity (SEN), precision (PRE), false positive per hour (FPR/h), and median detection offset (in seconds). We report both the average FPR/h (aFPR/h) and median FPR/h (mFPR/h). We mainly focus on the median FPR/h in this study as they are more robust to outliers compared to the average.

**References**

1. Jirsa, V. K., Stacey, W. C., Quilichini, P. P., Ivanov, A. I. & Bernard, C. On the nature of seizure dynamics. *Brain* **137**, 2210–2230 (2014).
2. Nunes, V. D., Sawyer, L., Neilson, J., Sarri, G. & Cross, J. H. Diagnosis and management of the epilepsies in adults and children: summary of updated nice guidance. *Bmj* **344** (2012).
3. Misulis, K. E. & Murray, E. L. *Essentials of Hospital Neurology* (Oxford University Press, 2017).
4. Britton, J. W. *et al.* Electroencephalography (EEG): An introductory text and atlas of normal and abnormal findings in adults, children, and infants (American Epilepsy Society, Chicago, 2016).
5. Gogou, M. & Cross, J. H. Seizures and epilepsy in childhood. *CONTINUUM: Lifelong Learn. Neurol.* **28**, 428–456 (2022).
6. Rinaldi, V. E., Di Cara, G., Mencaroni, E. & Verrotti, A. Therapeutic options for childhood absence epilepsy. *Pediatr. Reports* **13**, 658–667 (2021).
7. Milán-Tomás, Á., Persyko, M., Del Campo, M., Shapiro, C. M. & Farcnik, K. An overview of psychogenic non-epileptic seizures: etiology, diagnosis and management. *Can. J. Neurol. Sci.* **45**, 130–136 (2018).
8. Organization, W. H. *et al.* Atlas: epilepsy care in the world (World Health Organization, 2005).
9. Ferri, F. F. *Ferri’s Clinical Advisor 2020 E-Book: 5 Books in 1* (Elsevier Health Sciences, 2019).
10. Berg, A. T. Risk of recurrence after a first unprovoked seizure. *Epilepsia* **49**, 13–18 (2008).
11. Mornmann, F., Andrzejak, R. G., Elger, C. E. & Lehnertz, K. Seizure prediction: the long and winding road. *Brain* **130**, 314–333 (2007).
12. Jayakar, P. *et al.* Diagnostic utility of invasive eeg for epilepsy surgery: indications, modalities, and techniques. *Epilepsia* **57**, 1735–1747 (2016).
13. Geut, I., Weenink, S., Knotterius, I. & van Putten, M. J. Detecting interictal discharges in first seizure patients: ambulatory eeg or eeg after sleep deprivation? *Seizure* **51**, 52–54 (2017).
14. Sierra-Marcos, A., Scheuer, M. L. & Rossetti, A. O. Seizure detection with automated eeg analysis: a validation study focusing on periodic patterns. *Clin. neurophysiology* **126**, 456–462 (2015).
15. Rommens, N., Geertsema, E., Holleboom, L. J., Cox, F. & Visser, G. Improving staff response to seizures on the epilepsy monitoring unit with online eeg seizure detection algorithms. *Epilepsy & Behav.* **84**, 99–104 (2018).
16. Reus, E., Visser, G., van Dijk, J. & Cox, F. Automated seizure detection in an emu setting: are software packages ready for implementation? *Seizure* (2022).
17. Koren, J., Hafner, S., Feigl, M. & Baumgartner, C. Systematic analysis and comparison of commercial seizure-detection software. *Epilepsia* **62**, 426–438 (2021).
18. Shah, V. *et al.* Optimizing channel selection for seizure detection. In *2017 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 1–5 (IEEE, 2017).
19. Ayodele, K., Ikezogwo, W., Komolafe, M. & Ogumbona, P. Supervised domain generalization for integration of disparate scalp eeg datasets for automatic epileptic seizure detection. *Comput. Biol. Medicine* **120**, 103757 (2020).
20. Chatzichristos, C. *et al.* Epileptic seizure detection in eeg via fusion of multi-view attention-gated u-net deep neural networks. In *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 7 (2020).
21. Roy, S. *et al.* Evaluation of artificial intelligence systems for assisting neurologists with fast and accurate annotations of scalp electroencephalography data. *EBioMedicine* 103275 (2020).
22. Fürbass, F. *et al.* Prospective multi-center study of an automatic online seizure detection system for epilepsy monitoring units. *Clin. Neurophysiol.* **126**, 1124–1131 (2015).
23. Mansouri, A., Singh, S. P. & Sayood, K. Online eeg seizure detection and localization. *Algorithms* **12**, 176 (2019).
24. Gómez, C. *et al.* Automatic seizure detection based on imaged-eeg signals through fully convolutional networks. *Sci. reports* **10**, 1–13 (2020).
25. Valdez, A. B., Hickman, E. N., Treiman, D. M., Smith, K. A. & Steinmetz, P. N. A statistical method for predicting seizure onset zones from human single-neuron recordings. *J. neural engineering* **10**, 016001 (2012).
26. Soare, I. & Escudero, J. Evaluation of eeg dynamic connectivity around seizure onset with principal component analysis. In *44th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (IEEE Xplore, 2022).
27. Ansari, A. H. *et al.* Neonatal seizure detection using deep convolutional neural networks. *Int. journal neural systems* **29**, 1850011 (2019).
28. Yang, Y., Truong, N. D., Maher, C., Nikpour, A. & Kavehei, O. Two-channel epileptic seizure detection with blended multi-time segments electroencephalography spectrogram (2020).
29. Thyagachandran, A., Kumar, M., Sur, M., Aghoram, R. & Murthy, H. Seizure detection using time delay neural networks and lstms. In *2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 1–5 (IEEE, 2020).
30. Bhattacharya, A., Baweja, T. & Karri, S. Epileptic seizure prediction using deep transformer model. *Int. J. Neural Syst.* **2150058** (2021).
31. Pedoeem, J., Abittan, S., Yosef, G. B. & Keene, S. Tabs: Transformer based seizure detection. In *2020 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 1–6 (IEEE, 2020).
32. Raghu, S., Sriram, N., Temel, Y., Rao, S. V. & Kubben, P. L. Eeg based multi-class seizure type classification using convolutional neural network and transfer learning. *Neural Networks* **124**, 202–212 (2020).
33. Saab, K., Dunnmon, J., Ré, C., Rubin, D. & Lee-Messer, C. Weak supervision as an efficient approach for automated seizure detection in electroencephalography. *NPJ digital medicine* **3**, 1–12 (2020).
34. Emami, A. *et al.* Seizure detection by convolutional neural network-based analysis of scalp electroencephalography plot images. *NeuroImage: Clin.* **22**, 101684 (2019).
35. Covert, I. C. *et al.* Temporal graph convolutional networks for automatic seizure detection. In *Machine Learning for Healthcare Conference*, 160–180 (PMLR, 2019).
36. Roy, S., Kiral-Kornek, I. & Harrer, S. Chrononet: a deep recurrent neural network for abnormal eeg identification. In *Conference on Artificial Intelligence in Medicine in Europe*, 47–56 (Springer, 2019).

37. Shellhaas, R. A. & Clancy, R. R. Characterization of neonatal seizures by conventional eeg and single-channel eeg. *Clin. Neurophysiol.* **118**, 2156–2161 (2007).

38. Wusthoff, C., Shellhaas, R. & Clancy, R. Limitations of single-channel eeg on the forehead for neonatal seizure detection. *J. Perinatol.* **29**, 237–242 (2009).

39. Lu, Y., Ma, Y., Chen, C. & Wang, Y. Classification of single-channel eeg signals for epileptic seizures detection based on hybrid features. *Technol. Heal. Care* **26**, 337–346 (2018).

40. Ammar, S. & Senouci, M. Seizure detection with single-channel eeg using extreme learning machine. In *2016 17th international conference on sciences and techniques of automatic control and computer engineering (STA)*, 776–779 (IEEE, 2016).

41. Liu, J. & Woodson, B. Deep learning classification for epilepsy detection using a single channel electroencephalography (eeg). In *Proceedings of the 2019 3rd International Conference on Deep Learning Technologies*, 23–26 (2019).

42. Wang, X. *et al.* One dimensional convolutional neural networks for seizure onset detection using long-term scalp and intracranial eeg. *Neurocomputing* **459**, 212–222 (2021).

43. Henriksen, J. *et al.* Automatic seizure detection: going from seeg to ieeg. In *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, 2431–2434 (IEEE, 2010).

44. Zhang, Z. & Parhi, K. K. Low-complexity seizure prediction from ieeg/seeg using spectral power and ratios of spectral power. *IEEE transactions on biomedical circuits systems* **10**, 693–706 (2015).

45. Rout, S. K., Sahani, M., Dash, P. & Biswal, P. K. Multifuse multilayer multikernel rvlh+ of process modes decomposition and approximate entropy data from ieeg/seeg signals for epileptic seizure recognition. *Comput. Biol. Medicine* **132**, 104299 (2021).

46. Ziyabari, S., Shah, V., Golmohammadi, M., Obeid, I. & Picone, J. Objective evaluation metrics for automatic classification of eeg events. *arXiv preprint arXiv:1712.10107* (2017).

47. Cook, M. J. *et al.* Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study. *The Lancet Neurol.* **12**, 563–571 (2013).

48. Burrello, A., Schindler, K., Benini, L. & Rahimi, A. Hyperdimensional computing with local binary patterns: one-shot learning of seizure onset and identification of ictogenic brain regions using short-time ieeg recordings. *IEEE Transactions on Biomed. Eng.* **67**, 601–613 (2019).

49. O’Shea, A., Lightbody, G., Boylan, G. & Temko, A. Neonatal seizure detection from raw multi-channel eeg using a fully convolutional architecture. *Neural Networks* **123**, 12–25 (2020).

50. Brinkmann, B. H. *et al.* Crowdsourcing reproducible seizure forecasting in human and canine epilepsy. *Brain* **139**, 1713–1722 (2016).

51. Shen, Y. Machine learning based epileptic seizure detection for responsive neurostimulator system optimization. In *Journal of Physics: Conference Series*, vol. 1453, 012089 (IOP Publishing, 2020).

52. Li, A. *et al.* Neural fragility as an eeg marker of the seizure onset zone. *bioRxiv* 862797 (2021).

53. Cox, F., Reus, E., Widman, G., Zwemmer, J. & Visser, G. Epilepsy monitoring units can be safe places; a prospective study in a large cohort. *Epilepsy & Behav.* **102**, 106718 (2020).

54. Asif, U., Roy, S., Tang, J. & Harrer, S. Seizurenet: Multi-spectral deep feature learning for seizure type classification. In *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology*, 77–87 (Springer, 2020).

55. Hossain, M. S., Amin, S. U., Alsualaiman, M. & Muhammad, G. Applying deep learning for epilepsy seizure detection and brain mapping visualization. *ACM Transactions on Multiimed. Comput. Commun. Appl. (TOMM)* **15**, 1–17 (2019).

56. Yuan, Y., Xun, G., Jia, K. & Zhang, A. A multi-view deep learning framework for eeg seizure detection. *IEEE journal biomedical health informatics* **23**, 83–94 (2018).

57. Zhou, M. *et al.* Epileptic seizure detection based on eeg signals and cnn. *Front. neuroinformatics* **12**, 95 (2018).
58. Khalkhali, V. et al. Low latency real-time seizure detection using transfer deep learning. In 2021 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), 1–7 (IEEE, 2021).

59. Shah, V. Improved Segmentation for Automated Seizure Detection Using Channel-Dependent Posteriors (Temple University, 2021).

60. Peh, W. Y., Yao, Y. & Dauwels, J. Transformer convolutional neural networks for automated artifact detection in scalp eeg. arXiv preprint arXiv:2208.02405 (2022).

61. Shah, V. et al. The temple university hospital seizure detection corpus. Front. neuroinformatics 12, 83 (2018).

62. Shoeb, A. et al. Patient-specific seizure onset detection. Epilepsy & Behav. 5, 483–498 (2004).

63. Stevenson, N., Tapani, K., Lauronen, L. & Vanhatalo, S. A dataset of neonatal eeg recordings with seizure annotations. Sci. data 6, 1–8 (2019).

64. Wagenaar, J. B., Brinkmann, B. H., Ives, Z., Worrell, G. A. & Litt, B. A multimodal platform for cloud-based collaborative research. In 2013 6th international IEEE/EMBS conference on neural engineering (NER), 1386–1389 (IEEE, 2013).

65. Thomas, J. et al. Automated adult epilepsy diagnostic tool based on interictal scalp electroencephalogram characteristics: A six-center study. Int. J. Neural Syst. 2050074 (2021).

66. Thangavel, P. et al. Time–frequency decomposition of scalp electroencephalograms improves deep learning-based epilepsy diagnosis. Int. J. Neural Syst. 2150032 (2021).

67. Peh, W. Y. et al. Multi-center validation study of automated classification of pathological slowing in adult scalp electroencephalograms via frequency features. Int. J. Neural Syst. 2150016 (2021).

68. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

69. Joo, T., Chung, U. & Seo, M.-G. Being bayesian about categorical probability. In International Conference on Machine Learning, 4950–4961 (PMLR, 2020).

70. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In International Conference on Machine Learning, 1321–1330 (PMLR, 2017).

71. Afra, P., Jouny, C. C. & Bergey, G. K. Duration of complex partial seizures: an intracranial eeg study. Epilepsia 49, 677–684 (2008).

72. Krystal, A. D. et al. Comparison of seizure duration, ictal eeg, and cognitive effects of ketamine and methohexital anesthesia with ect. The J. neuropsychiatry clinical neurosciences 15, 27–34 (2003).

Author contributions statement
J.D. and W.Y.P. conceived and designed the experiments, W.Y.P. T.P., and Y.Y.Y. conducted the experiments, W.Y.P., J.D., T.P., J.T., and Y.L.T. analyzed and discussed the results. J.D. and W.Y.P. wrote the manuscript. All authors provided critical feedback and helped shape the research, analysis and manuscript.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SeizureDetectionJournalSI20220809.pdf