Biomedical Signal Processing and Control 64 (2021) 102360

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Automated inter-patient seizure detection using multichannel Convolutional and Recurrent Neural Networks

Jeff Craley [a,*], Emily Johnson [b], Christophe Jouny [b], Archana Venkataraman [a]

[a] Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, 21218, USA
[b] School of Medicine, Johns Hopkins University, Baltimore, MD, 21205, USA

ARTICLE INFO

Keywords:
Epilepsy
Seizure detection
Convolutional Neural Networks
Long Short-Term Memory networks

ABSTRACT

We present an end-to-end deep learning model that can automatically detect epileptic seizures in multichannel electroencephalography (EEG) recordings. Our model combines a Convolutional Neural Network (CNN) and a Bidirectional Long Short-Term Memory (BLSTM) network to efficiently mine information from the EEG data using a small number of trainable parameters. Specifically, the CNN learns a latent encoding for each one second window of raw multichannel EEG data. In conjunction, the BLSTM learns the temporal evolution of seizure presentations given the CNN encodings. The combination of these architectures allows our model to capture both the short time scale EEG features indicative of seizure activity as well as the long term correlations in seizure presentations. Unlike most prior work in seizure detection, we mimic an in-patient monitoring setting through a leave-one-patient-out cross validation procedure, attaining an average seizure detection sensitivity of 0.91 across all patients. This strategy verifies that our model can generalize to new patients. We demonstrate that our CNN–BLSTM outperforms both conventional feature extraction methods and state-of-the-art deep learning approaches that rely on larger and more complex network architectures.

1. Introduction

Epilepsy is a heterogeneous neurological disorder characterized by spontaneous bursts of neuronal synchrony in the brain that manifest as seizures [1]. Nearly 3.4 million people in the United States, or 1.2% of the population, are believed to have active cases of epilepsy [2]. Worldwide estimates place the number of cases at 50 million, making epilepsy one of the most common neurological disorders with an associated increase in mortality of up to threefold. With its wide prevalence and effect on premature death, epilepsy represents a large and ongoing public health challenge [3]. While the disorder can often be controlled with Anti-Epileptic Drugs (AEDs), and/or diet, roughly 20%–40% of epilepsy patients are medically refractory [4] and do not respond to drug treatment. In these cases, resection or neurostimulation can lead to good outcomes. However these treatments require precise knowledge of the seizure onset zone.

Scalp electroencephalography (EEG) is the first and foremost modality used for epilepsy diagnosis. Determination of the seizure type (focal or general), and the likely onset zone can be made by examining the temporal evolution of a seizure in this modality [5]. To acquire EEG recordings, patients are admitted to an epilepsy monitoring unit, where surface electrodes are applied, typically in the 10/20 or 10/10 international system [6], and any prescribed AEDs are withdrawn. Multichannel EEG data is recorded continuously over several days in order to capture roughly three to five seizures for each patient.

Identification of the seizure in scalp EEG is key for epilepsy diagnosis. However, epileptic activity may occur rarely, requiring days of long term epilepsy monitoring. Analyzing these continuous scalp EEG recordings is time consuming and requires extensive training. In addition, up to 30% of seizures are electrographic with no accompanying behavioral signatures to facilitate seizure identification. As a result of these difficulties, inter-rater agreement among clinicians can be low [7,8] requiring labor-intensive review and discussion. Thus accurate automatic seizure detection has the potential to save clinician time and improve the diagnosis and management of epilepsy.

In this paper we present a deep learning model for accurate seizure detection based on multichannel EEG that overcomes many of the challenges of the current clinical workflow. Our model accurately detects seizures in out of sample patients while maintaining a low number of false positive detections. This generalization ability makes our method uniquely suited for clinical review of prospectively acquired EEG recordings of new patients. Our model combines a Convolutional Neural Network (CNN) encoder with a Bidirectional Long Short-Term Memory (BLSTM) classifier to simultaneously extract predictive features from the EEG data and learn the evolution of seizure presentations

* Corresponding author.
E-mail address: jcraley2@jhu.edu (J. Craley).

https://doi.org/10.1016/j.bspc.2020.102360
Received 11 June 2020; Received in revised form 22 October 2020; Accepted 16 November 2020
Available online 21 November 2020

across time. We evaluate the CNN–BLSTM on EEG data from 34 patients recorded at the Johns Hopkins Hospital. We evaluate our model using leave-one-patient-out cross validation to ensure generalization to new patients. The CNN–BLSTM outperforms baseline methods that rely on hand-crafted features or larger deep neural architectures.

1.1. Machine learning and seizure detection

Seizure detection has been an active area of research for nearly fifty years. While many techniques have been applied to the problem, no standardized methodology has been adopted. Early work focused on rule based systems with hand designed features and thresholding [9]. As computational resources improved, research pivoted to applying signal processing techniques to characterize ictal (e.g. epileptic) EEG for seizure detection. For example, changes in the non-linear dynamics of ictal EEG noted in [10] inspired many researchers to use features derived from chaos theory to differentiate between seizure and baseline EEG. In the past decade, machine learning methods have started to dominate the automated seizure detection literature. In general, these approaches follow a two-stage pipeline. First, feature extraction is performed on windowed segments of EEG data. Second, a classifier is trained to declare each segment as seizure or baseline depending on the features extracted [11]. Below we detail common approaches in feature extraction and classification.

Time frequency-domain features. Brain wave activity is typically analyzed within separate frequency bands, which correspond with normal cognitive processes, such as wakefulness, relaxation, or drowsiness. Changes in activity within these bands can also indicate epileptic seizures [12]. Time–frequency analysis seeks to quantify these changes to detect epileptic events. The Fast Fourier Transform (FFT) is the simplest approach for time–frequency analysis. For example, the authors of [13] and [14] use the FFT to compute power in the 2.5–12 Hz band of each EEG channel. Thresholding techniques developed in [14] were applied to find periods of seizure activity within long-term recordings. In [15,16] components in the theta (1–4 Hz), delta (4–8 Hz), alpha (8–13 Hz), beta (13-30 Hz), and gamma ($\geq$ 30 Hz) brain wave bands were summed for each EEG channel. These features were used within a coupled hidden Markov model framework for tracking the evolution of seizure activity across EEG channels.

A more sophisticated approach uses filter banks to compute the spectral power in different frequency bands. In [17], the EEG signal in each channel was separated into eight evenly spaced frequency bands from 0.5–25 Hz using a filter bank. Patient specific seizure onset detectors were trained using a Support Vector Machine (SVM) classifier. Finally, the hierarchical nature of the wavelet transform has made it a popular representation for seizure detection. In [18], the energy in wavelet subbands from 1–30 Hz was used to create histograms of seizure and non-seizure activity. Changepoint detection was subsequently used to identify seizure onsets. In [19], energy and spectral features were calculated for each wavelet subband and used for classification in an array of classifiers. [20] follows a similar approach, extracting amplitude features for classification after performing a multichannel empirical wavelet transform to the original EEG signal. Similarly, [21] extracted features from non-linear signal processing for each subband of the wavelet transform. An array of classifiers were then compared for their efficacy in the seizure detection task. While these works demonstrate that changes in the EEG frequency content reflects seizure activity, FFT, wavelet, and filter bank based methods ignore phase information between EEG channels. Phase reversals have been long established in the EEG literature to indicate abnormal synchronous firing [12] but cannot be captured by methods that focus on just the power spectrum.

Time domain features. Time domain methods analyze the original EEG signals. As noted above, features from non-linear signal processing and chaos theory have received much attention, as in [22]. Non-linear signal processing techniques quantify the predictability of the system. For example, [22] uses approximate entropy, sample entropy, and phase entropy to measure the similarity of the EEG to its past behavior. It is noted that approximate and sample entropy are lower for non-seizure intervals, indicating a more predictable signal than that of ictal EEG.

Finally, many studies have combined time–frequency and time domain features to leverage advantages from each representation for gains in detection performance. For example, [23] extracts kurtosis, skewness, and correlations computed in the time domain, along with amplitudes and correlations between frequency decompositions taken from the spectral domain. Taking a different approach, the authors of [24,25] and [26] extract non-linear features following the decomposition of the original EEG signal into separate wavelet bands. Both [24] and [25] compute the correlation dimension and largest Lyapunov exponent for each subband after application of the wavelet transform. Similarly, [26] uses the wavelet transform to isolate different frequencies in the EEG signal and applies approximate entropy to each subband to create features for classification. These ensembles of features are borne out of necessity, as each feature on its own generally captures only one signal phenomenon in the original signal. And in fact, there is little evidence that these ensembles are stable across heterogeneous seizure presentations.

Classification. Many classification strategies have been used for seizure detection. SVMs are perhaps the most popular classifier, finding use in [17,27,28]. By comparing each test sample to algorithmically selected data points called support vectors, SVMs are capable of drawing complex decision boundaries via representative samples of the positive and negative classes [29].

The random forest classifier is another popular choice [23,30]. Random forests construct an ensemble of decision trees, such that each tree is trained with a random subset of the input examples, and each node is optimized using a random subset of the input features. This dual randomization provides robustness to overfitting, particularly when training data is limited [31]. Other classifiers for seizure detection include adaptive thresholding [13], which updates a thresholding parameter based on previous intervals, and k-nearest neighbors [30], which computes similarity between an unknown sample and known representatives of the positive and negative classes. While powerful, these classifiers are limited by the discriminative power of their input features. As seizure and baseline EEG morphologies vary widely between patients, traditional approaches often lead to poor generalization.

1.2. Deep learning for seizure detection

Driven by successes in domains such as computer vision and natural language processing, deep neural networks have come to dominate the machine learning field [32]. This interest in deep learning has extended into EEG analysis, finding applications in brain computer interfaces, sleep state analysis, and seizure detection [33]. Inspired by the organization of the brain, deep learning methods use a cascade of primitive functional units to learn arbitrarily complex functions. The network is trained by repeatedly showing it labeled data and updating the parameters according to some desired objective function. Given enough training data and an appropriately designed network architecture, deep methods will often surpass the performance of more traditional machine learning techniques. However, the architecture design of a deep network is a non-trivial problem.

The simplest deep learning architecture is the Multi-Layer Perceptron (MLP), which relies on fully connected neural network layers. While MLPs can learn more complex classification functions than both

SVMs and RFs, their performance is limited by the input features. This behavior is evident in [30], where MLPs, random forests, and SVMs are shown to achieve comparable results when using features extracted from the EEG data. To over come this challenge [34] uses a separate autoencoder to extract features from the raw EEG; the autoencoder output is then fed into the MLP. However this technique ignores the inherently temporal nature of the EEG signal, as each sample of the EEG is analyzed as a separate feature and not an element of an evolving sequence.

Convolutional Neural Networks (CNNs) are a more sophisticated architecture that captures interdependencies between neighboring data points across multiple scales. As such, they have found applications in image and sequence processing, where local similarities are important. Broadly, CNNs apply a set of convolutional kernels with a restricted field of view to form local representations of the signal. By aggregating these local representations and applying subsequent convolutional layers, information at larger and larger scales can be extracted. In this way, CNNs can replace hand designed feature extraction methods with representations learned directly from the data.

CNNs for EEG feature extraction can be divided into two classes, those that use time–frequency representations as input images and those that use the EEG signal as an input time series. Methods that use time–frequency inputs rely on two dimensional convolutions, similar to other computer vision applications. For example, [35] constructed a CNN to operate on Short-Time Fourier Transform (STFT) spectrograms for seizure detection. Similarly, in [36] the authors construct a 2D image of spectrograms taken from each EEG channel and classify these images using several popular CNN architectures from computer vision. Other time–frequency representations have also been considered, such as [37] where a wavelet decomposition was used to build a spectrogram for the CNN input. [38] used a tensor decomposition to find common components in the STFT of each EEG channel before input into a CNN. While CNNs are undoubtedly powerful, applying 2D convolutions to the EEG spectrogram imposes arbitrary structure between neighboring FFT frequency bins that is not present in the original signal space. In addition, as noted above, decomposition of the EEG signal using the FFT disregards important cross-channel phase information that may be indicative of seizure activity.

An alternative to the 2D CNN is to apply one-dimensional convolutions directly on the EEG signals, thus eliminating the need for time–frequency preprocessing. This approach is exemplified by [39–41], where short windows are fed directly into a five one-dimensional CNN followed by fully connected layers for seizure detection. In [39], single channels of intercranial EEG are classified using a one dimensional CNN. In [42] one dimensional convolutions are applied to each EEG channel individually while sharing the same parameters across channels to exploit information from all channels when learning hidden representations of the data. A similar approach was taken in [43], where a one dimensional CNN was applied to each channel individually while fusing information across channels using max pooling in the final classification stage. However, as information across EEG channels is not mixed until the final fully connected layers of the network, phase synchrony between channels may again be lost.

Recurrent Neural Networks (RNNs) are popular architectures for sequence analysis which maintain a hidden representation of the signal at each point in time. This hidden representation is continually updated based on its past value, thus fusing information from neighboring time points. In [44], one second windows of EEG are fed directly into an RNN analogous to the CNNs noted above. The output from the RNN layers is classified using an MLP layer. In [45] a BLSTM network was applied to continuous EEG recordings. The original EEG signal was decomposed using the local mean decomposition applied to each channel. Features were then extracted from each decomposition component. The resulting sequence of features was then classified using a BLSTM network. In addition RNNs can be combined with convolutional networks as in [46] and [47]. In [47], a CNN and LSTM layer were combined to perform seizure detection. Specifically, long windows (101 s) of EEG signal were passed through a 1D CNN. The resulting sequence of hidden representations was fed into a uni-directional LSTM. The output of the LSTM at the final time step was used to detect seizure activity for the entire 101 s sequence.

In contrast the authors of [46] create STFT images that span 30 s and analyzes them using a 2D CNN. This 2D CNN outputs a sequence of hidden states representing small periods of the original STFT which are subsequently fed into a RNN to classify the entire 30 s segment. [48] takes a different approach, using interpolation between EEG channels to create a 2D image of EEG features. A hidden representation for these images is computed using a CNN and 5 s long sequences of hidden states are classified using an LSTM network. While these methods allow for accurate classification of segments of EEG signal, they label sequences of EEG signal and thus have limited temporal resolution. As information from the time of onset is critical to localizing possible seizure foci, rapid and continuous seizure detection is necessary for clinical translation.

Table 1
Patient demographics and clinical attributes for our JHH evaluation dataset (N = 34).

	JHH Dataset
Seizure type	Focal epilepsy
Number of patients	34
Average age	35 ± 16 years
Minimum/Maximum age	6/77 years
Number of Males/Females	16/18
Seizures per Patient	5.9 ± 5.8
Minimum/Maximum seizures per patient	1/24
Average EEG analyzed per patient	1.8 ± 1.8 h
Average seizure duration	112 s
Minimum/Maximum seizure duration	13/979 s

1.3. Our contribution

In this work we present a novel neural network architecture for continuous seizure detection that addresses the critical need for high accuracy with low onset latency. Our model combines a CNN encoding stage and a Bidirectional Long Short-Term Memory (BLSTM) classification stage. The combined architecture contains a relatively small number of trainable parameters, ensuring that our model is computationally efficient. Furthermore, we evaluate our method in a leave-one-patient-out setting in order to evaluate its performance on previously unseen patients.

The CNN feature extraction uses one-dimensional convolutions simultaneously applied across all channels of the EEG recording to automatically learn discriminative representations from one-second windows of the EEG signal. The use of 1D convolutions on the multichannel data ensures that relevant phase information between channels is preserved. The BLSTM aggregates these fine grained CNN representations to learn the longer temporal dependencies of an evolving seizure. The bidirectional nature of our architecture leverages information from both the past and future to perform a window level classification. This learning process mirrors clinical practice, as clinicians generally take into account the temporal evolution of the EEG signal when annotating the beginning and end of a seizure.

While previous approaches have employed both CNNs and BLSTMs, our combined architecture improves upon these approaches in several important ways. First, prior studies have focused on one-dimensional CNNs applied individually to each EEG channel [39–43]. This approach ignores clinically relevant cross-channel phase information, as seizures are often characterized by atypical synchronization between channels [12,49]. Accordingly, our CNN operates on the multichannel EEG recording, preserving this phase information. Prior work using RNNs for seizure detection operates on long sequences of the EEG data, typically on the order of 5–100 s [46–48]. The RNN then provides a

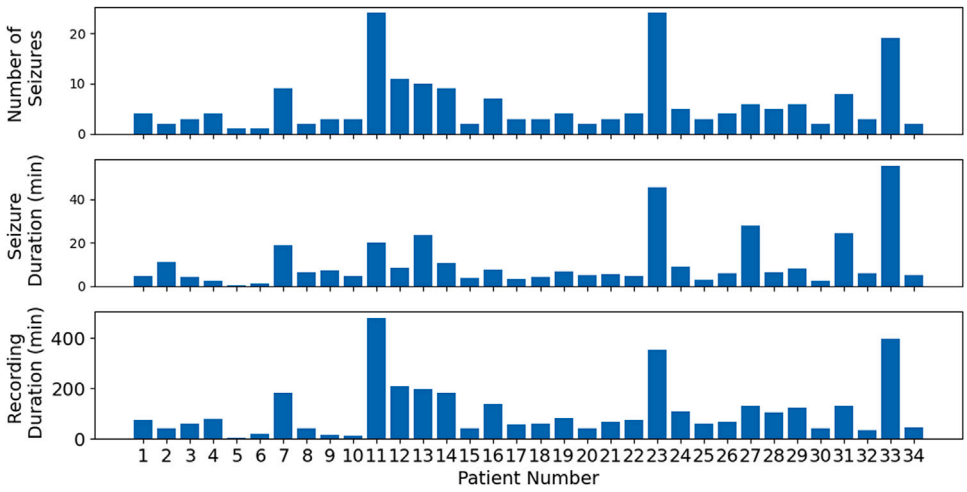


Fig. 1. Distribution of EEG recording statistics by patient.

single classification for the entire sequence. Due to the large sequence duration, this approach can only generate a coarse label for the seizure onset and offset. In contrast, our approach uses the CNN encoding to extract a compact representation for short (one second) windows of the EEG data. We can leverage the bi-directionality of the BLSTM to extract information from the entire recording and make predictions at a fine-grained level.

We demonstrate the generalizability the CNN–BLSTM by performing leave one patient out cross validation on a dataset of clinical EEG recordings. This cross validation method ensures that our network generalizes to new patients with different clinical manifestations. Finally, our CNN–BLSTM is simple with only four convolutional blocks and two recurrent layers. Hence, our model requires less training data than larger deep learning architectures for seizure detection such as [39], and it can easily be integrated into the existing clinical infrastructure.

2. Materials and methods

2.1. EEG data and preprocessing

We validate our model on a dataset of 34 patients, all of whom have focal epilepsy, acquired at the Johns Hopkins Hospital (JHH). The JHH dataset contains a total of 201 multichannel EEG recordings across 34 patients. This dataset has been previously used in [15,16,40,41]. The raw EEG is acquired at 200 Hz using the 10–20 common Ref. [6] and is converted to the longitudinal bipolar montage [12] for this work.

Patient characteristics and seizure presentations are summarized in Table 1. The JHH dataset was collected during clinical workup and contains a high degree of variability in number of seizures collected and duration of recording. Number of seizures collected, total recording duration included, and total seizure duration for each patient is shown in Fig. 1. On average, 5.9 seizures are included for each patient with a minimum of one and a maximum of 24. For each seizure, we include a maximum of 10 min of pre-seizure and post-seizure baseline. An average of 1.8 h of total recording time was included for each patient with a minimum of 4.7 min (patient 5, one seizure) and a maximum of roughly 480 min (patient 11, 24 seizures). While this variety in patient representation complicates model training, validating the model on a diverse dataset ensures that our models generalize to the diversity present in the clinical population.

Each recording is high-pass and low-pass filtered at 1.6 Hz and 30 Hz, respectively. High-pass filtering removes DC trends while low-pass filtering removes physiological artifacts that confound seizure detection. In order to ensure all recordings contain EEG signal of a similar amplitude, we apply a normalization procedure to each recording separately. Each recording was clipped to remove amplitudes larger than two standard deviations from the mean intensity to remove high intensity artifacts such as muscle artifact and electrode popping. The recordings were then normalized to have mean 0 and standard deviation 1 for each channel.

One second non-overlapping windows were extracted from each recording for input into our model (and baselines). Seizure activity in each recording is demarcated by a clinical annotation indicating seizure onset and offset. Any one second window that overlaps this period is considered a positive instance of the seizure class. Conversely, windows containing no seizure activity are labeled as baseline.

2.2. An end-to-end detection framework

2.2.1. CNN–BLSTM architecture

Our model can be conceptualized as a multichannel feature extractor (CNN) followed by a temporal detector (BLSTM). A schematic of the network is shown in Fig. 2. In the feature extractor stage, individual windows $X_t \in \mathcal{R}^{C\times L}$, where C is the number of EEG channels and L is the number of time samples in the window, are fed directly into the CNN. This process generates a sequence of hidden representations $\{h_t\}_{t=1}^{T}$, where T is the length of a given recording, which encode the relevant information for determining whether each window X_t lies within a seizure interval. The representations h_t are learned directly from the data X_t, increasing their discriminative power.

The CNN is composed of four successive blocks containing two layers each as shown in Fig. 2. Each layer includes a one dimensional convolution with a length three kernel, with a stride and pooling of one. After two repetitions of the convolution, LeakyReLu, and batch normalization, a max pooling operation is applied with a kernel size of two, effectively halving the length of the representation after each block. This succession of convolutional layers distills information from the EEG signal into higher order features. As in the VGGNet [50], we double the number of channels in each block after each max pooling. This process prevents an overall loss of information and ensures that each convolutional block requires roughly the same amount of computation. The convolution and max pooling procedures are illustrated in Fig. 3. As illustrated in the figure, the number of kernels doubles while the length of the representation is halved after each convolution block

Global average pooling is applied to the representation generated after the final convolution block. Effectively the output of each CNN kernel is averaged across the EEG window, resulting in a single feature for each kernel. This procedure has a regularizing effect on the network; broadly, it reduces overfitting, as the subsequent recurrent layers receive information pooled across the entire one second window,

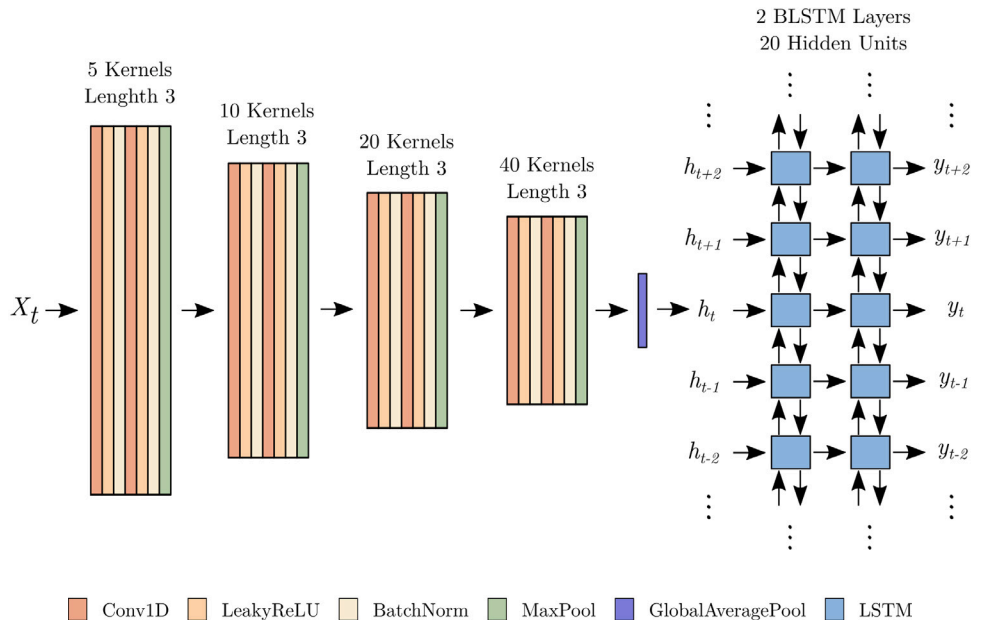


Fig. 2. Our CNN–BLSTM architecture for inter-patient seizure detection. A convolutional encoder converts EEG signal X_t to hidden representations h_t. These representations are classified by a two layer BLSTM to predict seizure labels y_t.

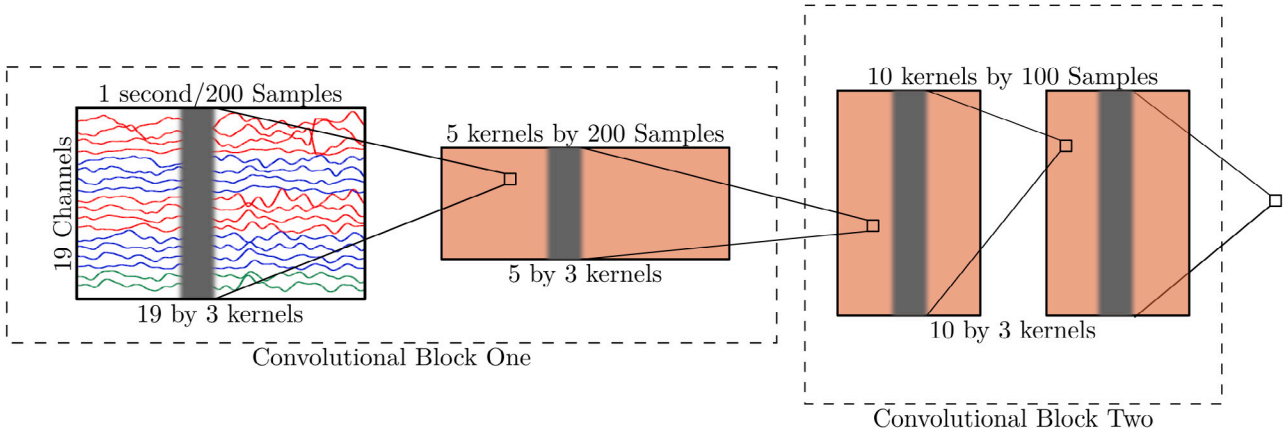


Fig. 3. The first two convolutional blocks of the CNN encoder. One second of preprocessed EEG signal is fed directly into the first layer of the CNN. An example input for each convolution is shown in gray while the corresponding output of the convolution is shown in the next layer as a square. Each block contains two convolutional layers. Between blocks, the number of convolutional kernels is doubled, while the length of the sequences is halved. LeakyReLU activations and batch norms not pictured.

thus mitigating overfitting to isolated data irregularities. As the final CNN layer contains 40 kernels, the output of the CNN feature extraction stage is reduced to a length 40 feature vector.

Following the CNN feature extraction stage, the sequence of hidden vectors $\{h_t\}_{t=1}^{T}$ is classified into a sequence of binary predictions $\{y_t\}_{t=1}^{T}$. The BLSTM architecture concatenates the output of two LSTMs, one operating on the sequence in the forward direction, and the other operating backward. Thus the BLSTM hidden state at any given time point includes information from both the past and the future of that time point. The bidirectional architecture allows the network to learn the temporal evolution of a seizure. By using the entire recording in the network at one second intervals, we learn the full progression from baseline to seizure and back, ensuring high temporal resolution and low latency. Two BLSTM hidden layers are used before outputting a final prediction y_t.

2.2.2. Postprocessing

To combat the noisy seizure versus baseline classification, we apply temporal smoothing to the sequence of predictions. Specifically, we average the network outputs y_t over a 20 sample window to enforce temporal contiguity in seizure detections. Near the beginning and end of the recording, any indices outside the data window are ignored when computing this average. As the output of our models is a continuous value between 0 and 1, it is important to establish a threshold at which to declare a positive (seizure versus baseline) detection. The setting of this threshold effectively controls the trade off between false positives and the sensitivity of our model. In this work, we opt to calibrate the CNN–BLSTM to a seizure detection threshold based on a user-specified duration of false seizure detection. For the experiments presented here, the seizure detection threshold was set such that each model is allowed only 2 min of false positives per hour. This threshold is computed from the training set after training. The computed threshold is subsequently applied to the test set.

2.2.3. Training and implementation

The flexibility and expressiveness of our CNN–BLSTM network makes it prone to overfitting. When trained in an end-to-end fashion, we observe the network to be able to exactly learn the presentation of specific seizures in the training set while failing to generalize to new data. In addition, RNNs can be notoriously difficult to train due to the

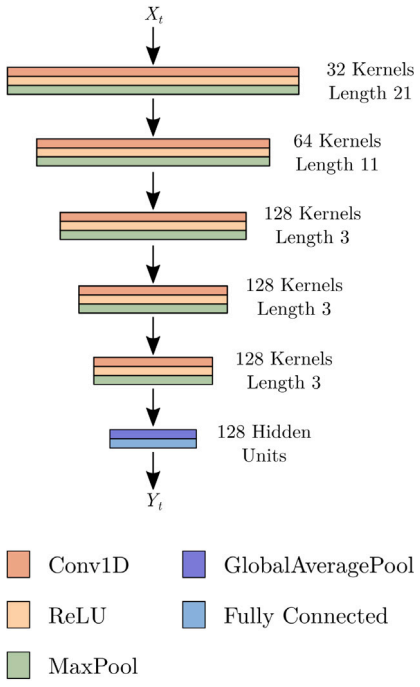


Fig. 4. Wei–CNN baseline Architecture.

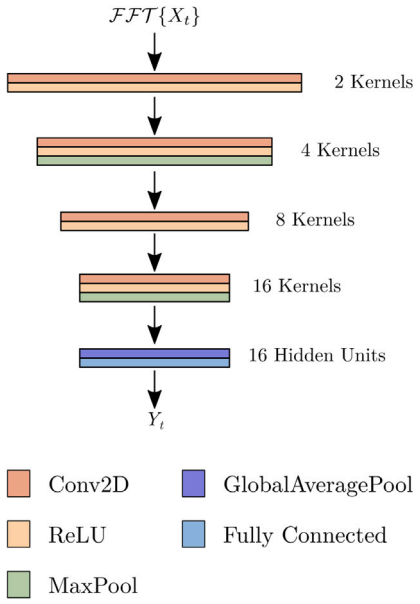


Fig. 5. CNN–2D FFT image baseline architecture.

vanishing and exploding gradient problem [32]. Furthermore, while our dataset contains hours of EEG recordings, we have in total only 201 seizure presentations. As the BLSTM operates on full recordings, this limits the number of examples in our dataset to a relatively small number for deep learning.

To address these concerns, we adopt a two stage training strategy to combat both overfitting and the difficulties in RNN training. In the first stage, the CNN is pre-trained by appending a simple fully connected MLP (two layers of 20 hidden units). To classify individual one second windows, we train the CNN for 10 epochs using a batch size of 32 windows, a learning rate of 0.01, and the ADAM optimizer [51]. In the pre-training stage, we train using the cross entropy loss. As the dataset contains a high imbalance between seizure and non-seizure classes, each class is weighted according to the inverse proportion of its prevalence in the dataset. In this fashion we leverage the large recording time of EEG signal in our dataset while sidestepping the limited number of total seizures. Thus, this pre-training ensures that the CNN learns discriminative feature representations from the raw EEG signal prior to the training of the BLSTM network for temporal classification.

In the second stage of training, the MLP is removed and the BLSTM layers are appended to the network. The full CNN–BLSTM is then trained in an end-to-end fashion. As the CNN has already learned to extract discriminative features, this stage of training focuses on learning the temporal evolution of seizures in the BLSTM layers. During this phase, entire seizure recordings are used as samples and fed to the CNN–BLSTM in their entirety. When training the BLSTM, we use the cross entropy loss applied to each individual window of the recording with the same weighting as applied in the pre-training stage. Thus each window contributes to the loss for the entire recording. We use a batch size of 2, indicating that a gradient step is taken after two recordings are passed through the network. The network is trained with a learning rate of 0.005 using the ADAM optimizer [51]. As the computational power of the BLSTM greatly increases the chance of overfitting to the limited number of total seizures in the training data, we adopt an early stopping strategy and only train the combined model for a single epoch. Using this training technique, we are able to fully utilize the data in our dataset to train the BLSTM–CNN.

2.3. Baseline comparison methods

2.3.1. Feature based classification

Our first set of baseline methods employs the two-stage feature selection and classification pipeline discussed in Section 1.1. While many approaches to seizure detection have been presented in the literature, variations in implementation, datasets used, and experiment design make direct comparisons difficult. As such, we opt to construct our baseline comparisons using feature extraction techniques representative of the major approaches in the field of seizure detection as discussed in Section 1.1. From the time domain, we compute total signal power, sample entropy, Largest Lyapunov Exponent (LLE), and line length on a channel-wise basis. The features are extracted independently for each one second window of raw EEG data. Mathematically, let $X_t^j[i]$ denote sample i of channel j at time t. We calculate power in a single channel using the expression $\frac{1}{L}\sum_{i=1}^{L}(X_t^j[i])^2$. Line length is computed using the expression $\sum_{i=2}^{L}|X_t^j[i]-X_t^j[i-1]|$. Sample entropy and LLE computations are performed following [22] and [24,25], respectively. Intuitively sample entropy measures the degree to which similar trajectories remain similar to previously observed paths. Likewise, LLE measures the rate at which similar trajectories diverge from each other. We calculate these features using the freely available Python nolds package [52]. These time domain features contribute a single scalar for each channel, resulting in a total of 54 time domain features for each one second window. In the time–frequency domain, we compute the filter bank power in each channel by passing X_t^j through a set of 10 evenly spaced order four Butterworth bandpass filters from 0 to 30 Hz. This results in a total of 180 time–frequency domain features.

The variety of classifiers used in the seizure detection literature mirrors the variety of feature extraction techniques. We limit our baseline investigations to MLP classifiers, as these classifiers have shown high seizure detection efficacy in recent literature and lead to complementary comparisons with our CNN–BLSTM models. We construct a MLP classifier to determine whether or not each EEG window X_t lies within a seizure interval. This classification is done based on (i) time–domain only, (ii) time–frequency domain only, and (iii) the combined feature set. Features extracted for each channel are concatenated and fed directly into the MLP classifier. The MLP baseline includes two layers of ten hidden units each. During training, dropout of 0.5 is applied after each layer. Due to the noisiness of classifications made on

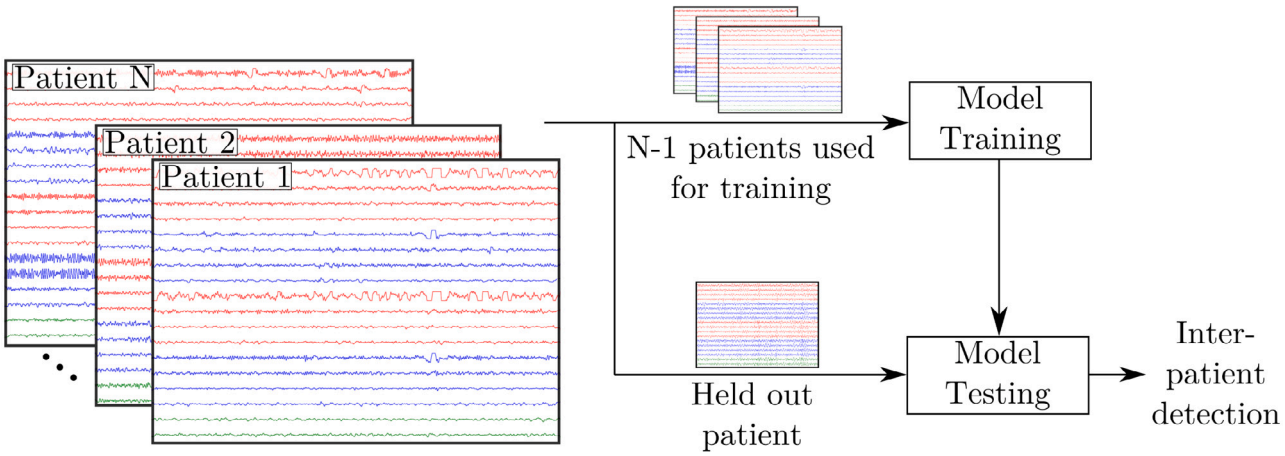


Fig. 6. Cross validation procedure, in which one patient is left-out for testing while the rest of the dataset is used for training. This procedure is repeated for each patient and the performance is averaged across all N folds.

single seconds of EEG signal, we apply the same temporal smoothing and calibration described in 2.2.2 to limit false positives.

In addition, we implement the wavelet-based feature extraction and SVM classifier from Kaleem et al. [19] (Kaleem–SVM). Using this classification pipeline, the authors report a sensitivity, specificity, and accuracy of 99.7, 99.2, and 99.4, respectively, on the CHB-MIT dataset in a patient specific seizure classification task. A five level DWT is performed on 4 s windows of EEG signals. Energy, sparsity of the amplitude spectrum, and the sum of the derivative of the amplitude spectrum are calculated for each subband. These features are concatenated and classified using a linear SVM.

2.3.2. Convolutional models

We implement the CNN network from Wei et al. [39] (Wei–CNN). This architecture has been shown to perform well in the literature, achieving a sensitivity, specificity, and accuracy of 0.7211, 0.9589, and 0.8400, respectively, on the publicly available CHB-MIT pediatric epilepsy dataset [53,54]. While this work trained the CNN models by leaving out a single test patient, recordings from this left out patient were used as a validation set for early stopping. As such these results represent performance under less restrictive conditions than the leave-one-patient-out cross validation paradigm considered in this work. The architecture of this network, as shown in Fig. 4, contains five convolutional and max pooling layers before two fully connected layers. Each layer of the Wei–CNN uses a one-dimensional CNN kernel with a stride of 1. The first through fifth layers of the network use decreasing kernel sizes of 21, 11, 3, 3, and finally 3. Zero padding of 11, 6, 2, 2, and 2, respectively, is used. Due to the larger size of the network, max pooling in the Wei–CNN uses a kernel size and stride of 3. Designed for windows five seconds in length, this network is accordingly much bigger. This baseline will assess the performance when using longer time windows, as opposed to a temporal evolution model. We also evaluate results using the CNN–MLP network in our pre-training section. By comparing our model to this network, the increase in performance from the BLSTM is directly quantifiable. Again, we apply temporal smoothing and calibration as described in Section 2.2.2.

Finally, we implement a two dimensional CNN model (CNN–2D) that operates on the FFT features in an image format. The CNN–2D architecture is detailed in Fig. 5. The EEG signal is windowed into one second non-overlapping segments and an FFT is calculated. FFT amplitudes from 0 to 30 Hz are arranged into an 2D image with channel along one axis and frequency along the other. These 2D images are input into a 4 layer CNN, where the number of kernels is doubled after each layer. ReLU operations are applied at each layer, and max pooling is applied after the second and fourth layers. Each convolution uses a kernel of size 3 with stride 1 and no padding. In addition, each max pooling operation uses a kernel size and stride of 2. Finally, global average pooling is applied, followed by frame-wise classification using a single fully connected layer. This approach was inspired by [35,37,38] where time–frequency decompositions are used in conjunction with 2D convolutions.

2.4. Cross validation

Most studies optimize patient-specific seizure detectors, in which a single recording is set aside for testing, and a detector is trained on the remaining recordings. This method of evaluation assumes that seizure recordings for a given patient are available *a priori* [55]. This patient specific approach is appropriate for settings such as responsive neurostimulation or in developing seizure alert systems for a particular patient. However, during clinical review, a clinician would like to prospectively detect seizures with no *a priori* EEG data from the patient. During this phase of the clinical workflow, long continuous EEG recordings are retrospectively analyzed for seizure content by trained neurologists, requiring considerable time.

Patient agnostic or inter-patient seizure detection trains detectors based at the population level. This leave-one-patient-out procedure is shown in Fig. 6. To ensure that trained models generalize to new patients, we perform cross validation by removing a single patient from the dataset. This patient is used as a test subject while models are trained on the remaining patients. In this way we mimic a clinical review setting, where previously trained models are applied to newly admitted patients on-the-fly. A similar cross validation was used in [56] to reduce bias in estimating the generalization error of a neonatal seizure detection algorithm. We emphasize that leave-one-patient-out is a far more challenging paradigm than the patient-specific evaluations used in prior work due to the variable seizure presentations across individuals. Hence, the performance metrics are expected to be lower.

2.5. Evaluation

We evaluate performance of our detectors both at the level of individual EEG windows X_t and at the level of seizures. At the window level, each snippet X_t is labeled as belonging to the seizure or baseline class $y_t \in \{0, 1\}$. We evaluate the Area Under the Receiver Operating Curve (AUC-ROC) and the Area Under the Precision–Recall Curve (AUC-PR). These metrics provide summary scores that capture behavior at a range of detection thresholds. In addition, we include the sensitivity and specificity of computed based on the thresholds computed during the calibration phase. While these metrics are less clinically relevant than those evaluated at the seizure level, they offer a convenient illustration of each model's overall performance.

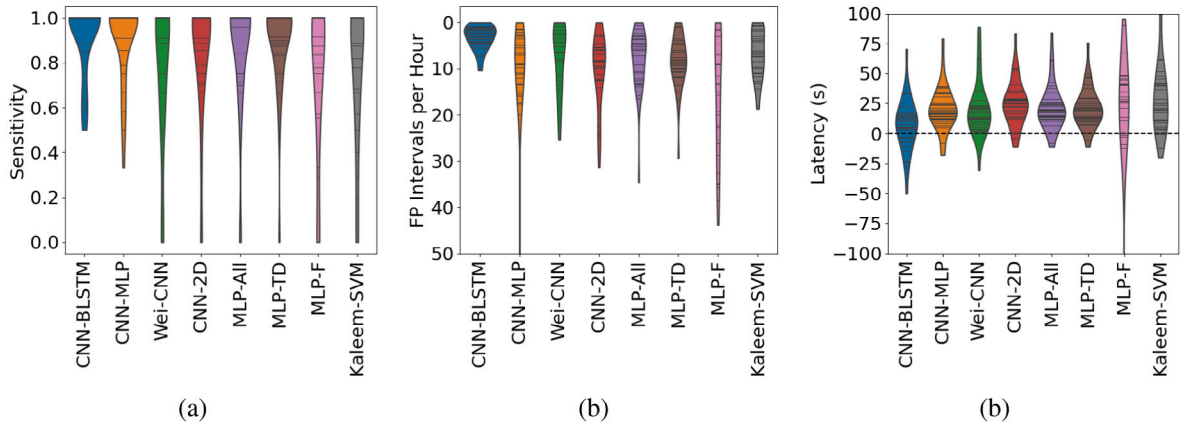


Fig. 7. Violin plots depicting seizure level (a) sensitivity (b) false positives per hour (c) latency for each model. Horizontal lines indicate single datapoints from each trial of leave-one-patient-out cross validation. Width of the violin shows the smoothed distribution of each metric.

Table 2
IID window level results.

	AUC-ROC	AUC-PR	Sensitivity	Specificity	Number of parameters
CNN–BLSTM	**0.9042**	**0.6491**	**0.6304**	0.9295	30k
CNN–MLP	0.8624	0.6031	0.5370	0.9555	11k
Wei–CNN	0.7642	0.4880	0.4048	0.9209	174k
CNN–2D	0.8243	0.5268	0.4695	0.9491	1.5k
MLP–All	0.8448	0.5895	0.5138	0.9532	8k
MLP–Time	0.8380	0.5614	0.5096	0.9540	2k
MLP–Filterbank	0.7135	0.3761	0.3371	0.9148	6k
Kaleem–SVM	0.7054	0.4304	0.3643	0.9454	–

At the seizure level, we consider contiguous seizure classifications produced by each model. Namely, if the model prediction exceeds the threshold determined in Section 2.2.2, a seizure onset is marked. This seizure classification continues until the model output once again falls below the threshold. Any detections of this kind that fall within an annotated seizure are considered true positives. Conversely, any contiguous detections that do not overlap with an annotated seizure are considered false positives. We quantify the sensitivity (true positives divided by total number of seizures), latency of seizure detection, and False Positive Rate (FPR) of each model. The goal in a clinical setting is to achieve high accuracy with low FPR.

3. Experimental results

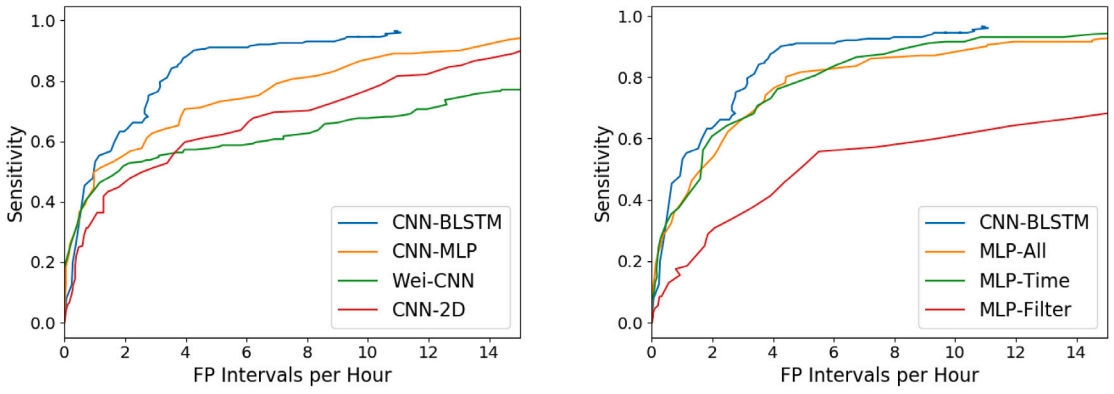
3.1. Window level accuracy

Table 2 reports the window-level detection performance along with the number of trainable parameters for each model. We observe that the CNN–BLSTM model outperforms all competing models achieving an AUC-ROC and AUC-PR of 0.9042 and 0.6491, respectively. This model is followed by the CNN–MLP (AUC-ROC 0.8620, AUC-PR 0.6017) and CNN–2D (AUC-ROC 0.8243, AUC-PR 0.5268). The Wei–CNN baseline performs the worst of all end-to-end models with an AUC-ROC of 0.7642 and AUC-PR of 0.4880. Of the feature-based MLP baselines, the network trained using all features performs best with an AUC-ROC of 0.8448 and an AUC-PR of 0.5895. The MLP trained with time domain features achieves slightly lower but still comparable performance measures. The network trained using only filter bank features performs significantly worse, with an AUC-ROC and AUC-PR of 0.7135 and 0.3761, respectively. While the MLP–All model achieves decent performance with roughly 8 k parameters, the CNN–MLP model outperforms it while increasing the parameter count by only roughly 3 k. The addition of the BLSTM network enlarges the model to roughly 30 k parameters with an accompanying increase in AUC-ROC and AUC-PR. This threefold increase in parameters between the CNN–MLP and CNN–BLSTM is justified by this gain in performance, while the CNN–BLSTM remains significantly smaller than the much larger Wei–CNN model. The Kaleem–SVM performed worst of all, achieving AUC-ROC and AUC-PR of 0.7054 and 0.4304, retrospectively.

In addition, Table 2 includes sensitivity and specificity measures for each model computed on a window-wise basis. We observe that the CNN–BLSTM model outperforms all other baselines in sensitivity, achieving a sensitivity of 0.6304. The CNN–MLP, Wei–CNN, and CNN–2D all exhibited lower sensitivities, with 0.5370, 0.4048, and 0.4695, respectively. Again we observe that the MLP–All model achieves a decent performance in these metrics, with a sensitivity and specificity of 0.5138 and 0.9532. All models exhibit specificities above 0.9, with the CNN–MLP achieving the highest specificity of 0.9555.

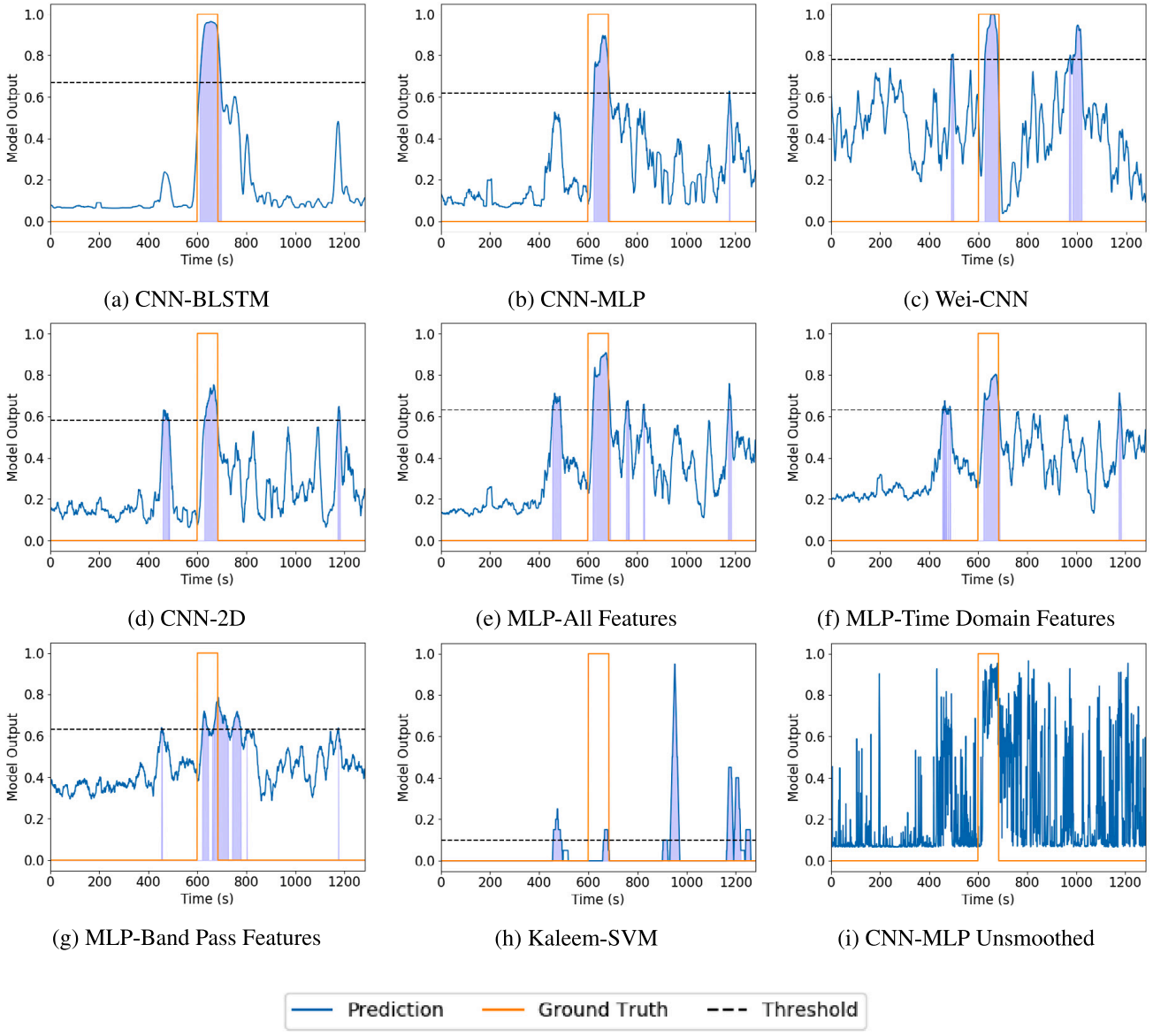
3.2. Seizure level results

Fig. 7 depicts violin plots for sensitivity, false positive rate, and latency for each model. In these plots, metrics computed from each left-out patient are indicated by horizontal lines within the violin. The width of the violin represents the distribution of the computed metrics across all patients. In Fig. 7(a) we see that the CNN–BLSTM maintains high sensitivity across the dataset, while baseline models fail to generalize to some patients. In addition, Fig. 7(b) shows that CNN–BLSTM false positive rates cluster near 3 false positives per hour. In contrast, the baselines exhibit higher false positive rates in some patients. Tables A.3 and A.4 in the appendix report the patient-specific performance metrics. Table A.3 shows performance for the CNN-based models (CNN–BLSTM, CNN–MLP, Wei–CNN, CNN–2D) while Table A.4 shows results for MLP-based models (MLP All Features, MLP–Time Domain Features, and MLP–Filterbank Features). When averaged across left-out patients the CNN–BLSTM achieves an average sensitivity of 0.91 while allowing an average of 3.3 FPs/h. The CNN–MLP, Wei–CNN, and CNN–2D all exhibit lower sensitivities at 0.90, 0.77, and 0.84, respectively, and FPs/h of 9.6, 7.5, and 10.2, respectively. Thus the CNN–BLSTM achieves the highest sensitivity with the lowest false positive rate. Finally, Fig. 7(c) shows the spread of onset latencies for



(a) End-to-End Baselines (b) Feature-Based MLP Baselines

Fig. 8. Sensitivity versus false positive rate curves for each model. The metrics are calculated as the seizure detection threshold is swept from 0 to 1 for each patient. The threshold sweep is performed globally and not calibrated separately for each patient.



(a) CNN-BLSTM (b) CNN-MLP (c) Wei-CNN

(d) CNN-2D (e) MLP-All Features (f) MLP-Time Domain Features

(g) MLP-Band Pass Features (h) Kaleem-SVM (i) CNN-MLP Unsmoothed

Fig. 9. Model outputs for a representative seizure recording. Seizure prediction scores for each window of the EEG recording are pictured for the duration of the recording. Time proceeds along the x-axis while seizure prediction certainty is shown on the y-axis. 0 indicates non-seizure baseline while 1 denotes seizure, while higher values indicate increasing model confidence in seizure activity. Seizure prediction thresholds for each model calculated during calibration are shown as a horizontal dashed line. Any predictions crossing this threshold are considered positive seizure predictions and are shown in blue. True labels are shown in orange, where 0 indicates baseline and 1 indicates seizure.

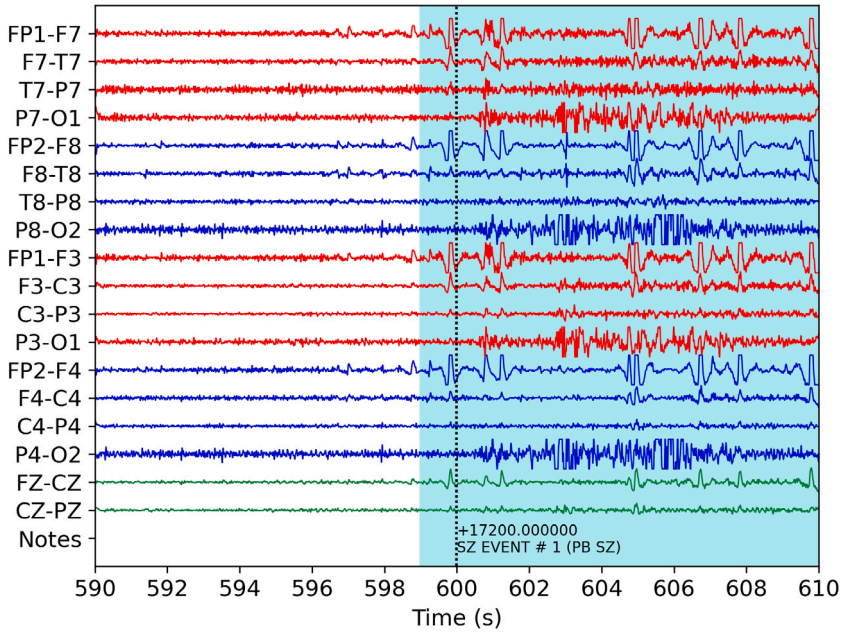


Fig. 10. EEG recording and CNN–BLSTM classification corresponding to Fig. 9(a). Seizure onset annotation is depicted by the vertical dashed line at 600 s. CNN–BLSTM seizure classification is shown shaded in light blue. The CNN–BLSTM declares the onset of a seizure at 599 s, in accordance with the clinical annotation.

each model. In the CNN–BLSTM onset latency is distributed around 10 s while other models report a higher average latency.

Sensitivity versus FPR plots are shown in Fig. 8, grouped according to the baseline model type. In this plot we sweep the threshold globally across each left-out patient and compute the overall sensitivity and the number of false positive intervals per hour across all testing runs. As optimal calibration points differ for each model, these plots do not correspond directly averaged metrics given in Tables A.3 and A.4. Despite this fact, we observe several important trends. The CNN–BLSTM achieves much higher sensitivities at lower FPRs when compared to baseline methods. Only when false positives are increased to much higher levels do baseline methods achieve the level of sensitivity of the CNN–BLSTM method at lower FPR.

Fig. 9 shows the classifications for a representative seizure recording. In each figure, time proceeds along the x-axis while the model output is shown on the y-axis. This output ranges continuously from 0 (baseline) to 1 (seizure). The calibration threshold for each model is indicated by the horizontal dashed black line. Regions containing positive seizure detections are shaded blue. As seen the CNN–BLSTM exhibits a high degree of certainty in the seizure label throughout the entire seizure, activating slightly after the annotated onset and continuing past its annotated duration. This extension past the end of the seizure is less clinically relevant than accurate onset detection and is likely due to artifact in the EEG recording occurring past the offset annotation.

The prediction output of baseline methods are shown in Fig. 9(b)–(h). While most baselines correctly detect seizure activity during the seizure interval, this detection generally occurs much later than the onset. Also notable is the presence of false positive detections, such as in Fig. 9(c), where the Wei–CNN makes three spurious false positive detections throughout the recording. Fig. 9(i) shows the unsmoothed prediction output for the CNN–MLP model. When comparing this image to the smoothed CNN–MLP output in Fig. 9(b), the effect of temporal smoothing is clear. After smoothing, the temporally contiguous positive seizure classification during the true seizure event remains with high certainty, while the more sporadic deviations away from baseline are averaged resulting in a lower certainty of seizure.

The raw EEG signal and CNN–BLSTM classification for a representative seizure is shown in Fig. 10. In this image, EEG signals in the longitudinal bipolar montage are arranged vertically while time proceeds horizontally. Annotated onset in this recording corresponds to a patient push button alarm occurring 600 s after the start of the recording, in the figure indicated by the vertical dashed black line. The CNN–BLSTM detects the seizure at 599 s, one second prior to the push button alarm annotation. Thus the onset time detected by the CNN–BLSTM corresponds closely to the annotated onset of the seizure.

4. Discussion

We have developed a novel CNN–BLSTM network for robust inter-patient epileptic seizure detection in long windows of continuously acquired EEG. Our model uses a CNN to extract discriminative hidden representations directly from the EEG signal. These representations are then classified using a recurrent BLSTM network, which learns the temporal evolution of seizure presentations by fusing information from the past and future. The combination of these two elements yields a detection performance with high sensitivity and low error. We validate our model on a challenging dataset of focal epilepsy patients, in which the seizures exhibit a high degree of heterogeneity. To evaluate the clinical utility of our model, we train and test our CNN–BLSTM network using leave-one-patient-out cross validation. Thus we ensure that our model can generalize to new patients in a continuous epilepsy monitoring setting.

Our model achieves higher sensitivity than numerous baseline comparison methods, correctly classifying 0.955 of seizures averaged across patients. This performance is mirrored in the AUC-ROC and AUC-PR scores, where our model again outperforms competing methods. At the patient level, we see in Table A.3 that our model correctly detects all seizures for many patients. The lowest patient sensitivity is 0.5, indicating that one half of the seizures are still correctly classified. By calibrating the model's detection threshold on the training set, we restrict the amount of false positive to two minutes per hour. This low FPR generalizes across patients, as the average number of FPs/h during testing was 3.3.

When examining the output of an individual model in Fig. 9(i), we observe a high degree of noise. This behavior can be effectively ameliorated by applying temporal smoothing to the output of each model, as seen in Fig. 9(b). However, we note that our BLSTM network

Table A.3
JHH CNN seizure results by patient.

Patient	CNN–BLSTM			CNN–MLP			Wei–CNN			CNN–2D		
	FPs/h	Sensitivity	Latency (s)	FPs/h	Sensitivity	Latency (s)	FPs/h	Sensitivity	Latency (s)	FPs/h	Sensitivity	Latency (s)
Patient 1	2.5	1.00	22.00	9.1	1.00	30.75	11.6	1.00	12.00	14.0	1.00	28.00
Patient 2	1.5	1.00	3.50	1.5	1.00	11.50	13.5	1.00	12.50	3.0	1.00	12.00
Patient 3	1.0	1.00	3.33	2.0	1.00	24.00	0.0	1.00	22.67	6.0	1.00	25.00
Patient 4	7.1	0.50	−3.00	14.9	0.50	5.00	12.6	0.50	9.00	18.1	0.50	29.00
Patient 5	0.0	1.00	−24.00	0.0	1.00	−8.00	0.0	1.00	3.00	0.0	1.00	−4.00
Patient 6	0.0	1.00	15.00	3.0	1.00	24.00	0.0	1.00	62.00	6.0	1.00	27.00
Patient 7	4.7	0.89	−7.12	7.0	1.00	16.89	19.3	1.00	−30.89	5.3	1.00	27.67
Patient 8	1.5	1.00	22.00	9.0	1.00	34.00	13.5	1.00	3.50	22.5	1.00	29.00
Patient 9	0.0	0.67	7.50	0.0	0.67	18.50	0.0	0.67	17.00	0.0	0.67	22.00
Patient 10	0.0	1.00	34.00	0.0	1.00	39.33	14.4	1.00	20.33	9.6	1.00	41.00
Patient 11	2.1	1.00	5.21	2.5	1.00	14.00	0.0	0.00	0.00	3.0	0.71	25.41
Patient 12	1.7	0.64	12.29	5.5	0.91	15.90	0.9	0.91	18.70	6.9	0.91	16.70
Patient 13	5.2	0.60	10.00	5.5	0.50	33.40	8.3	0.50	88.60	12.6	0.70	54.14
Patient 14	3.0	1.00	14.00	12.0	1.00	19.89	14.3	0.89	21.12	31.3	0.89	24.88
Patient 15	0.0	1.00	13.50	3.0	1.00	18.00	4.5	1.00	10.50	3.0	1.00	22.00
Patient 16	1.7	0.86	13.83	3.5	0.86	14.83	0.9	0.14	23.00	8.3	0.86	12.17
Patient 17	4.4	1.00	4.00	6.6	0.67	23.50	8.8	0.67	27.50	24.2	1.00	−11.00
Patient 18	4.0	1.00	13.67	15.9	1.00	37.67	7.9	0.67	62.50	9.9	0.33	36.00
Patient 19	9.6	1.00	16.00	15.5	1.00	16.75	5.2	1.00	13.25	14.0	1.00	18.00
Patient 20	1.4	1.00	−10.50	1.4	1.00	12.00	0.0	1.00	27.50	5.8	1.00	16.50
Patient 21	3.7	1.00	33.33	11.0	1.00	38.67	2.7	1.00	51.33	5.5	1.00	35.00
Patient 22	4.5	1.00	17.00	20.0	1.00	10.00	25.4	1.00	7.00	2.7	1.00	28.75
Patient 23	3.4	0.58	−3.50	6.5	0.58	25.21	6.5	0.67	21.44	9.3	0.75	34.06
Patient 24	1.1	1.00	−14.80	5.0	1.00	6.40	1.7	1.00	5.00	5.5	1.00	0.60
Patient 25	4.1	1.00	−27.00	13.3	1.00	12.67	5.1	1.00	12.33	24.6	1.00	14.33
Patient 26	1.8	0.75	11.33	17.6	0.75	8.33	21.9	0.75	1.33	12.3	0.75	13.33
Patient 27	2.3	1.00	70.33	2.7	1.00	79.00	2.3	1.00	82.67	2.7	1.00	52.17
Patient 28	5.3	1.00	5.80	8.8	1.00	18.60	6.4	0.80	13.00	4.7	0.80	22.75
Patient 29	7.4	1.00	−6.83	13.4	1.00	16.50	0.0	0.67	27.50	12.4	1.00	19.17
Patient 30	5.6	0.50	−50.00	11.3	1.00	−18.00	4.2	0.00	0.00	8.5	0.00	0.00
Patient 31	0.9	1.00	33.75	4.6	1.00	57.12	21.0	1.00	27.50	8.2	1.00	54.38
Patient 32	10.3	1.00	−6.00	68.9	0.33	21.00	3.4	0.00	0.00	27.5	0.00	0.00
Patient 33	2.9	0.84	33.81	9.7	0.79	34.87	17.2	1.00	14.26	11.4	0.79	53.93
Patient 34	6.7	1.00	−23.50	17.3	1.00	−13.50	2.7	0.50	36.00	9.3	1.00	83.50
Average	**3.3**	**0.91**	**7.03**	9.6	0.90	20.55	7.5	0.77	21.27	10.2	0.84	25.39

further suppresses this classification noise by directly learning the evolution of a seizure over time. This temporal suppression is especially evident when comparing results between the CNN–MLP and CNN–BLSTM, as the former includes the discriminative feature extraction of the CNN architecture without the temporal element granted by the BLSTM. When calibrated identically, the CNN–MLP achieves a similar sensitivity of 0.90 with a much higher rate of 9.6 false positives per hour. This behavior is evident in Fig. 9(b), where the CNN–MLP correctly identifies the seizure but exhibits less confidence in non-seizure and makes a spurious false positive detection.

As is evident in Fig. 7(c), the average onset latency for the CNN–BLSTM is evenly distributed around 10 s. Comparison methods exhibit higher positive latencies, indicating that the seizure detection occurs after the annotated onset. While these later detections can still be useful for identifying seizure in long recordings, often the seizure onset is most important for diagnosis. As seen in Fig. 10, the CNN–BLSTM responds to electrographic signatures of epilepsy prior to the push button alarm seizure annotation. Thus we observe that the CNN–BLSTM is capable of recognizing clinically relevant epileptic and detecting seizures with low latency.

Table 2 shows the approximate number of trainable parameters for each of the networks used. With roughly 30k parameters, the CNN–BLSTM network is nearly an order of magnitude smaller than the Wei–CNN, which contains roughly 174k trainable parameters. Smaller still is the CNN–MLP, which contains only 11k. It is notable that the CNN–MLP and the Wei–CNN perform comparably in summary statistics AUC-ROC and F1 given that the CNN–MLP model is roughly 15 times smaller. Smaller still, the feature based MLP network contains only approximately 2.5k trainable parameters. However, when comparing the pre-computed features to the end-to-end CNNs it is clear that extracting encodings directly from the multichannel EEG time series results in performance gains. As such the CNN–BLSTM achieves the best tradeoff between number of trainable parameters and performance.

The increase in discriminative power when using a CNN feature extractor comes with little, if any, extra computational requirement. To heuristically evaluate computational load, we timed feature extraction on a roughly 4 min sample of EEG. Bandpass, FFT, line-length, and power features combined could be computed in less than 5 s. However, using a freely available Python package, the non-linear features sample entropy and LLE took roughly 60 and 320 s, respectively, far too long for use in a clinical environment. By comparison, the CNN–BLSTM took roughly 0.15 s to classify this recording when running on the CPU (i.e. without GPU acceleration), indicating that the computational complexity is on par with the least expensive feature extraction techniques.

Extensions to the work presented here could further leverage advances in deep learning to provide greater translational benefits. As in all deep learning research, increases in dataset size lead directly to performance gains. Collecting more annotated continuous EEG recordings promise to facilitate the development of more powerful models. While accurate seizure detection is important in clinical practice, this task is only an intermediate step in diagnosing epilepsy subtypes and identifying possible focal onset zones. Future extensions could provide onset localization alongside detection to further assist the clinician. In addition, specific EEG morphologies, such as rhythmicity, slowing, and phase reversals, are often useful in diagnosis. Models capable of annotating EEG for this content could provide further utility in long term epilepsy monitoring.

5. Conclusions

We have presented a CNN–BLSTM network for inter-patient seizure detection that is optimized for use in the epilepsy monitoring unit. Our

Table A.4
JHH MLP seizure results by patient.

Patient	MLP–All			MLP–Time Domain Features			MLP–Filterbank Features			KaleemSVM		
	FPs/h	Sensitivity	Latency (s)	FPs/h	Sensitivity	Latency (s)	FPs/h	Sensitivity	Latency (s)	FPs/h	Sensitivity	Latency (s)
Patient 1	9.1	1.00	28.50	8.3	1.00	29.00	13.2	1.00	25.75	13.2	1.00	19.50
Patient 2	4.5	1.00	11.50	10.5	1.00	10.50	6.0	1.00	9.00	0.0	1.00	110.00
Patient 3	3.0	1.00	24.33	1.0	1.00	25.33	5.0	1.00	23.00	3.0	1.00	45.67
Patient 4	15.7	0.50	10.00	14.9	0.75	9.67	14.9	1.00	9.75	0.8	0.00	0.00
Patient 5	0.0	1.00	−8.00	0.0	1.00	−5.00	0.0	1.00	−10.00	12.7	1.00	−12.00
Patient 6	6.0	1.00	24.00	9.0	1.00	25.00	6.0	1.00	23.00	0.0	1.00	21.00
Patient 7	6.0	1.00	6.78	6.0	1.00	7.33	9.3	1.00	3.56	11.0	0.89	23.75
Patient 8	16.5	1.00	25.00	13.5	1.00	29.50	13.5	1.00	23.00	1.5	0.50	25.00
Patient 9	0.0	0.67	19.50	0.0	0.67	19.50	0.0	0.67	16.50	8.4	0.00	0.00
Patient 10	4.8	1.00	39.67	4.8	1.00	40.67	4.8	1.00	37.67	0.0	1.00	39.67
Patient 11	0.6	0.96	17.78	1.0	0.92	14.73	1.5	0.96	15.61	0.6	0.88	11.43
Patient 12	5.7	1.00	11.55	11.8	1.00	11.82	6.9	1.00	10.45	6.3	0.82	9.11
Patient 13	5.2	0.70	60.86	9.2	0.90	50.00	10.1	0.80	45.00	1.8	0.20	11.00
Patient 14	34.7	1.00	22.44	29.3	0.89	26.88	36.3	1.00	6.11	18.0	0.78	36.57
Patient 15	4.5	1.00	18.00	7.5	1.00	19.00	13.5	1.00	17.00	0.0	1.00	62.00
Patient 16	5.2	0.71	12.00	4.4	0.71	13.60	8.3	0.86	11.33	6.1	0.57	5.50
Patient 17	12.1	1.00	12.33	16.5	1.00	11.67	25.3	1.00	11.00	9.9	0.67	42.00
Patient 18	10.9	1.00	37.67	16.9	1.00	47.33	12.9	1.00	33.33	6.9	0.67	81.00
Patient 19	12.6	1.00	20.25	8.1	1.00	20.75	14.8	1.00	18.75	9.6	1.00	20.25
Patient 20	1.4	1.00	15.00	4.3	1.00	19.00	1.4	1.00	12.50	0.0	1.00	52.00
Patient 21	9.2	1.00	34.33	10.1	1.00	37.67	14.7	1.00	31.67	6.4	1.00	40.33
Patient 22	3.6	1.00	23.50	6.4	1.00	23.75	10.9	1.00	20.25	7.3	1.00	3.50
Patient 23	10.5	0.75	25.50	5.9	0.83	29.20	14.4	0.79	23.00	3.9	0.29	41.29
Patient 24	4.4	1.00	10.20	3.9	1.00	12.40	8.3	1.00	8.00	11.0	1.00	25.20
Patient 25	7.2	0.67	14.50	8.2	0.67	19.00	14.4	0.67	13.00	14.4	0.67	17.50
Patient 26	3.5	0.75	14.33	3.5	0.75	13.33	8.8	0.75	13.00	3.5	1.00	41.25
Patient 27	2.3	1.00	84.00	0.9	1.00	22.33	5.9	1.00	75.83	18.7	1.00	−0.67
Patient 28	13.4	1.00	19.00	11.7	1.00	21.00	20.4	1.00	17.20	9.9	0.40	48.00
Patient 29	10.9	1.00	13.67	9.9	1.00	9.67	13.4	1.00	11.33	7.4	0.67	28.50
Patient 30	2.8	0.00	0.00	9.9	0.50	7.00	8.5	0.50	8.00	12.7	1.00	4.00
Patient 31	5.9	1.00	61.25	6.4	0.88	75.57	9.6	1.00	56.00	4.1	0.88	105.43
Patient 32	13.8	0.00	0.00	8.6	0.00	0.00	22.4	0.00	0.00	0.0	0.00	0.00
Patient 33	13.1	0.84	43.00	11.8	0.84	46.38	13.7	0.89	38.35	11.7	0.68	61.31
Patient 34	13.3	1.00	−11.00	10.7	1.00	−11.00	14.7	1.00	−13.00	6.7	1.00	−20.00
Average	8.0	0.87	21.81	8.4	0.89	21.55	11.3	0.91	18.97	6.7	0.75	29.38

model uses a CNN network to learn a discriminative representation EEG data on one-second windows. These representations are scored using a BLSTM which analyzes the entire seizure recording. The CNN–BLSTM network contains a relatively small number of trainable parameters, making it appropriate for clinical applications.

We show that even when limiting false positives, the CNN–BLSTM provides clinically useful sensitivity. We further show that our method generalizes to new patients via leave-one-patient-out cross validation. Finally, our CNN–BLSTM outperforms larger models with more parameters. Taken together, our CNN–BLSTM has the potential to facilitate clinical review of multichannel scalp EEG.

CRediT authorship contribution statement

Jeff Craley: Conceptualization, Methodology, Software, Writing - original draft. **Emily Johnson:** Data curation, Writing - review & editing, Funding acquisition. **Christophe Jouny:** Data curation, Writing - review & editing. **Archana Venkataraman:** Conceptualization, Methodology, Writing - review & editing, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by a JHU Discovery Award, USA (Venkataraman/Johnson) and NSF CAREER, USA 1845430 (Venkataraman).

Appendix. Results by patient

See Tables A.3 and A.4

References

[1] R.S. Fisher, C. Acevedo, A. Arzimanoglou, A. Bogacz, J.H. Cross, C.E. Elger, J. Engel Jr, L. Forsgren, J.A. French, M. Glynn, et al., Ilae official report: a practical clinical definition of epilepsy, Epilepsia 55 (4) (2014) 475–482.

[2] M. Zach, et al., National and state estimates of the numbers of adults and children with active epilepsy – United States, 2015, CDC MMWR 66 (2017) 821–825.

[3] W.H. Organization, Epilepsy fact sheet, 2010, https://www.who.int/news-room/fact-sheets/detail/epilepsy (Accessed: 2010-03-10).

[4] J.A. French, Refractory epilepsy: clinical overview, Epilepsia 48 (2007) 3–7.

[5] H.O. Lüders, I. Najm, D. Nair, P. Widdess-Walsh, W. Bingman, The epileptogenic zone: general principles, Epileptic Disord. 8 (2) (2006) 1–9.

[6] V. Jurcak, et al., 10/20, 10/10, and 10/5 systems revisited: their validity as relative head-surface-based positioning systems, Neuroimage 34 (4) (2007) 1600–1611.

[7] S.B. Wilson, R. Emerson, Spike detection: a review and comparison of algorithms, Clin. Neurophysiol. 113 (12) (2002) 1873–1881.

[8] C.A. van Donselaar, R.-J. Schimsheimer, A.T. Geerts, A.C. Declerck, Value of the electroencephalogram in adult patients with untreated idiopathic first seizures, Arch. Neurol. 49 (3) (1992) 231–237.

[9] J. Gotman, Automatic recognition of epileptic seizures in the eeg, Electroencephalogr. clin. Neurophysiol. 54 (5) (1982) 530–540.

[10] R.G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, C.E. Elger, Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state, Phys. Rev. E 64 (6) (2001) 061907.

[11] I. Osorio, H.P. Zaveri, M.G. Frei, S. Arthurs, Epilepsy: The Intersection of Neurosciences, Biology, Mathematics, Engineering, and Physics, CRC press, 2016.

[12] L.V. Marcuse, M.C. Fields, J.J. Yoo, Rowan's Primer of EEG E-Book, Elsevier Health Sciences, 2015.

[13] R. Hopfengärtner, B.S. Kasper, W. Graf, S. Gollwitzer, G. Kreiselmeyer, H. Stefan, H. Hamer, Automatic seizure detection in long-term scalp eeg using an adaptive thresholding technique: a validation study for clinical routine, Clin. Neurophysiol. 125 (7) (2014) 1346–1352.

[14] R. Hopfengärtner, F. Kerling, V. Bauer, H. Stefan, An efficient, robust and fast method for the offline detection of epileptic seizures in long-term scalp eeg recordings, Clin. Neurophysiol. 118 (11) (2007) 2332–2343.

[15] J. Craley, E. Johnson, A. Venkataraman, A spatio-temporal model of seizure propagation in focal epilepsy, IEEE Trans. Med. Imaging (2019) 1, http://dx.doi.org/10.1109/TMI.2019.2950252.

[16] J. Craley, E. Johnson, A. Venkataraman, A novel method for epileptic seizure detection using coupled hidden markov models, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2018, pp. 482–489.

[17] A.H. Shoeb, J.V. Guttag, Application of machine learning to epileptic seizure detection, in: Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010, pp. 975–982.

[18] A.S. Zandi, et al., Automated real-time epileptic seizure detection in scalp eeg recordings using an algorithm based on wavelet packet transform, IEEE Trans. Biomed. Eng. 57 (7) (2010) 1639–1651.

[19] M. Kaleem, A. Guergachi, S. Krishnan, Patient-specific seizure detection in long-term eeg using wavelet decomposition, Biomed. Signal Process. Control 46 (2018) 157–165.

[20] A. Bhattacharyya, R.B. Pachori, A multivariate approach for patient-specific eeg seizure detection using empirical wavelet transform, IEEE Trans. Biomed. Eng. 64 (9) (2017) 2003–2015.

[21] O. Faust, U.R. Acharya, H. Adeli, A. Adeli, Wavelet-based eeg processing for computer-aided seizure detection and epilepsy diagnosis, Seizure 26 (2015) 56–64.

[22] U. Acharya, et al., Automated diagnosis of epileptic eeg using entropies, Biomed. Signal Process. Control 7 (4) (2012) 401–408.

[23] D. Wu, Z. Wang, L. Jiang, F. Dong, X. Wu, S. Wang, Y. Ding, Automatic epileptic seizures joint detection algorithm based on improved multi-domain feature of ceeg and spike feature of aeeg, IEEE Access 7 (2019) 41551–41564.

[24] S. Ghosh-Dastidar, H. Adeli, N. Dadmehr, Mixed-band wavelet-chaos-neural network methodology for epilepsy and epileptic seizure detection, IEEE Trans. Biomed. Eng. 54 (9) (2007) 1545–1551.

[25] H. Adeli, S. Ghosh-Dastidar, N. Dadmehr, A wavelet-chaos methodology for analysis of eegs and eeg subbands to detect seizure and epilepsy, IEEE Trans. Biomed. Eng. 54 (2) (2007) 205–211.

[26] H. Ocak, Automatic detection of epileptic seizures in eeg using discrete wavelet transform and approximate entropy, Expert Syst. Appl. 36 (2) (2009) 2027–2036.

[27] M. Bandarabadi, C.A. Teixeira, J. Rasekhi, A. Dourado, Epileptic seizure prediction using relative spectral power features, Clin. Neurophysiol. 126 (2) (2015) 237–248.

[28] V. Sridevi, M.R. Reddy, K. Srinivasan, K. Radhakrishnan, C. Rathore, D.S. Nayak, Improved patient-independent system for detection of electrical onset of seizures, J. Clin. Neurophysiol. 36 (1) (2019) 14.

[29] K.P. Murphy, Machine Learning: A Probabilistic Perspective, The MIT Press, 2012.

[30] E. Alickovic, J. Kevric, A. Subasi, Performance evaluation of empirical mode decomposition, discrete wavelet transform, and wavelet packed decomposition for automated epileptic seizure detection and prediction, Biomed. Signal Process. Control 39 (2018) 94–102.

[31] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[32] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[33] A. Craik, Y. He, J.L. Contreras-Vidal, Deep learning for electroencephalogram (eeg) classification tasks: a review, J. Neural Eng. 16 (3) (2019) 031001.

[34] A.-M. Tăuţan, M. Dogariu, B. Ionescu, Detection of epileptic seizures using unsupervised learning techniques for feature extraction, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 2377–2381.

[35] Y. Yuan, G. Xun, K. Jia, A. Zhang, A multi-view deep learning framework for eeg seizure detection, IEEE J. Biomed. Health Inform. 23 (1) (2018) 83–94.

[36] Y. Gao, B. Gao, Q. Chen, J. Liu, Y. Zhang, Deep convolutional neural network-based epileptic electroencephalogram (eeg) signal classification, Front. Neurol. 11 (2020).

[37] H. Khan, L. Marcuse, M. Fields, K. Swann, B. Yener, Focal onset seizure prediction using convolutional networks, IEEE Trans. Biomed. Eng. 65 (9) (2017) 2109–2118.

[38] M. Taherisadr, M. Joneidi, N. Rahnavard, Eeg signal dimensionality reduction and classification using tensor decomposition and deep convolutional neural networks, in: 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2019, pp. 1–6.

[39] Z. Wei, J. Zou, J. Zhang, J. Xu, Automatic epileptic eeg detection using convolutional neural network with improvements in time-domain, Biomed. Signal Process. Control 53 (2019) 101551.

[40] J. Craley, E. Johnson, C. Jouny, A. Venkataraman, Automated noninvasive seizure detection and localization using switching Markov models and convolutional neural networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 253–261.

[41] J. Craley, E. Johnson, A. Venkataraman, Integrating convolutional neural networks and probabilistic graphical modeling for epileptic seizure detection in multichannel eeg, in: International Conference on Information Processing in Medical Imaging, Springer, 2019, pp. 291–303.

[42] L. Zou, X. Liu, A. Jiang, X. Zhousp, Epileptic seizure detection using deep convolutional network, in: 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP), IEEE, 2018, pp. 1–4.

[43] A. O'Shea, G. Lightbody, G. Boylan, A. Temko, Neonatal seizure detection from raw multi-channel eeg using a fully convolutional architecture, Neural Netw. 123 (2020) 12–25.

[44] L. Vidyaratne, A. Glandon, M. Alam, K.M. Iftekharuddin, Deep recurrent neural network for seizure detection, in: 2016 International Joint Conference on Neural Networks (IJCNN), IEEE, 2016, pp. 1202–1207.

[45] X. Hu, S. Yuan, F. Xu, Y. Leng, K. Yuan, Q. Yuan, Scalp eeg classification using deep bi-lstm network for seizure detection, Comput. Biol. Med. 124 (2020) 103919.

[46] A. Affes, A. Mdhaffar, C. Triki, M. Jmaiel, B. Freisleben, A convolutional gated recurrent neural network for epileptic seizure prediction, in: International Conference on Smart Homes and Health Telematics, Springer, 2019, pp. 85–96.

[47] W. Liang, H. Pei, Q. Cai, Y. Wang, Scalp eeg epileptogenic zone recognition and localization based on long-term recurrent convolutional network, Neurocomputing 396 (2020) 569–576.

[48] K. Ayodele, W. Ikezogwo, M. Komolafe, P. Ogunbona, Supervised domain generalization for integration of disparate scalp eeg datasets for automatic epileptic seizure detection, Comput. Biol. Med. (2020) 103757.

[49] G.L. Krauss, R.S. Fisher, The Johns Hopkins Atlas of Digital EEG: An Interactive Training Guide, The Johns Hopkins University Press, 2007.

[50] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[51] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.

[52] C. Schölzel, Nonlinear measures for dynamical systems, Zenodo, 2019, http://dx.doi.org/10.5281/zenodo.3814723.

[53] A.H. Shoeb, Application of machine learning to epileptic seizure onset detection and treatment (Ph.D. thesis), Massachusetts Institute of Technology, 2009.

[54] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.-K. Peng, H.E. Stanley, Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals, circulation 101 (23) (2000) e215–e220.

[55] H. Qu, J. Gotman, A seizure warning system for long-term epilepsy monitoring, Neurology 45 (12) (1995) 2250–2254.

[56] S.B. Nagaraj, N.J. Stevenson, W.P. Marnane, G.B. Boylan, G. Lightbody, Neonatal seizure detection using atomic decomposition with a novel dictionary, IEEE Trans. Biomed. Eng. 61 (11) (2014) 2724–2732.