NOVEL GRAPHICAL MODEL AND NEURAL NETWORK FRAMEWORKS FOR AUTOMATED SEIZURE DETECTION, TRACKING, AND LOCALIZATION IN FOCAL EPILEPSY

by
Jeff Craley

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
January 2022

Abstract

Epilepsy is a heterogenous neurological disorder characterized by recurring and unprovoked seizures [1]. It is estimated that 60% of epilepsy patients suffer from focal epilepsy, where seizures originate from one or more discrete locations within the brain. After onset, focal seizure activity spreads, involving more regions in the cortex. Diagnosis and therapeutic planning for patients with focal epilepsy crucially depends on being able to detect epileptic activity as it starts and localize its origin. Due to the subtlety of seizure activity and the complex spatio-temporal propagation patterns of seizure activity, detection and localization of seizure by visual inspection is time-consuming and must be done by highly trained neurologists.

In this thesis, we detail modeling approaches to identify and capture the spatio-temporal ictal propagation of focal epileptic seizures. Through novel multi-scale frameworks, information fusion between signal paths, and hybrid architectures, models that capture the underlying seizure propagation phenomena are developed. The first half relies on graphical modeling approaches to detect seizures and track their activity through the space of EEG electrodes. A coupled hidden Markov model approach to seizure propagation is described. This model is subsequently improved through the addition of convolutional neural network based likelihood functions, removing the reliance on hand designed feature extraction. Through the inclusion of a hierarchical switching chain and localization variables, the model is revised to capture multi-scale seizure onset and spreading information.

In the second half of this thesis, end-to-end neural network architectures for seizure detection and localization are developed. First, combination convolutional and recurrent neural networks are used to identify seizure activity at the level of individual EEG channels.

Through novel aggregation, the network is trained to recognize seizure activity, track its evolution, and coarsely localize seizure onset from lower resolution labels. Next, a multi-scale network capable of analyzing the global and electrode level signals is developed for challenging task of end-to-end seizure localization. Onset location maps are defined for each patient and an ensemble of weakly supervised loss functions are used in a multi-task learning framework to train the architecture.

Thesis Readers

Dr. Archana Venkataraman (Primary Advisor)
John C. Malone Assistant Professor
Department of Electrical and Computer Engineering
Johns Hopkins University

Dr. Sridevi V. Sarma
Associate Professor
Department of Biomedical Engineering
Vice Dean for Graduatae Education
Whiting School of Engineering
Associate Director
Institute for Computational Medicine
Johns Hopkins University

Dr. Sanjeev Khudanpur
Associate Professor
Department of Electrical and Computer Engineering
Johns Hopkins University

Dr. Emily Johnson
Assistant Professor of Neurology
Department of Medicine
Johns Hopkins University

Dr. Richard Leahy
Arthur G. Settle Trust Endowment for USC Leonard Silverman Chair and Professor
Electrical and Computer Engineering, Biomedical Engineering, and Radiology
Department Chair
Ming Hsieh Department of Electrical Engineering-Systems
University of Southern California

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Motivation

Epilepsy is a heterogeneous neurological disorder characterized by recurrent and unprovoked seizures [1]. Epilepsy affects between 1–3% of the world's population, making it one of the most prevalent neurological disorders. While epilepsy can often be controlled with medication, it is estimated that 20–40% of patients are medically refractory and do not respond to anti-epileptic drugs [2]. Alternative therapies for these patients rely on our ability to detect and localize epileptic seizures in the the brain. Epileptic seizures can be be broadly characterized as either focal or generalized. Generalized seizures manifest across the cortex. Conversely, focal seizures originate in a specific onset zone, but may subsequently spread to neighboring regions of the brain until potentially the entire cortex is involved [3]. In medically refractory focal epilepsy, resection of the onset zone may be the only treatment available to completely eliminate seizures.

For patients with medically refractory epilepsy, scalp EEG plays a critical role in treatment planning. In the case of focal seizures, it can be used to coarsely localize the Seizure Onset Zone (SOZ). Scalp EEG is a natural complement to other noninvasive imaging modalities, such as PET [4] and MRI [5], which can be used to refine the localization. These modalities have the advantage of a higher spatial resolution, but they are more costly to acquire. After multimodal localization is performed using EEG and other noninvasive imaging, the extent of

the onset zone may be identified using more invasive techniques, such as electrocorticography (ECoG) or stereoEEG, just prior to surgical resection. EEG studies play a critical role in this process, affording clinicians a noninvasive and cost efficient means of establishing early information necessary for treatment planning.

Scalp EEG recordings are typically acquired over the course of several days after any medication is withdrawn and an adequate number of seizures are recorded. Visual inspection of the EEG recordings remains the standard procedure for seizure detection. This process is time intensive and requires extensive training. As such, the application of machine learning to EEG inspection has high potential impact in assisting clinicians in this critical component of the clinical workflow.

Where previous applications of machine learning in epilepsy have primarily focused on seizure detection, this approach has limited clinical utility. In this thesis, I present methods developed to move beyond the standard approaches of detecting seizures to include the mapping of spatio-temporal seizure activity as it spreads through the brain. The models presented here are informed by and seek to capture the biological mechanisms of seizure propation in scalp EEG signals. This approach opens a new direction for machine learning methods in epilepsy, by going beyond the standard seizure detection paradigm into the difficult but clinically relevant task of SOZ localization.

The work described here traces a course starting in graphical modeling and ending in neural network based approaches. Using graphical modeling, the behavior of the model can be explicitly controlled such that only certain configurations of states of random variables are allowed. By designing models to admit only states analogous to those observed in seizure spreading, highly specialized models to capture seizure activity will be presented. Two iterations of graphical modeling approaches will be presented. In this section of the thesis, a comparison of traditional signal processing and neural network based likelihoods will also be presented.

The second half of the thesis will present end-to-end neural network approaches. Where

graphical modeling allows models to be defined to capture specific phenomena, neural networks allow complex decision functions to be learned directly from data. By learning to predict seizure spreading activity from the EEG signals, neural networks have the potential to learn accurate representations of seizure activity and eliminate the need for hyperparameter tuning and difficulties in fitting graphical models. To circumvent the black box nature of neural networks, methods for training and analyzing the unstructured neural network outputs to extract clinically relevant seizure detection and localization information are developed.

Outline

Chapter 2 will provide relevant background. Clinical material regarding focal epilepsy and the EEG signal will be discussed. The problem of seizure detection will be introduced and prior work using traditional machine learning pipelines will then be discussed. Technical background regarding graphical modeling and deep learning will be provided along with overviews of their usage in EEG applications. Finally, preliminaries such as the datasets used in this thesis and notation will be discussed.

Chapter 3 will discuss the development of coupled Hidden Markov Models (CHMMs) for seizure activity tracking. The work presented in this chapter was originally presented as a MICCAI paper [6] and subsequently expanded into a journal form in [7]. By allowing latent seizure states in neighboring EEG channels to influence each other, the CHMM can be observed to detect seizure activity as it spreads through the EEG signal.

In Chapter 4, the CHMM presented in Chapter 3 will be extended through the inclusion of Convolutional Neural Network (CNN) likelihood functions. This work was originally presented in conference paper form in [8]. By incorporating CNN likelihoods, powerful likelihood functions can be learned directly from the EEG signal, eliminating the need for hand-designed feature extraction.

Chapter 5 details the Regime-Switching Markov Model for Propagation and Localization

(R-SMMPL) described in [9]. Moving beyond the CHMM approach described in Chapter 3, the R-SMMPL uses hierarchically structured random variables to more accurately capture seizure propagation phenomena while providing localization information for each patient.

In the second half of the thesis, deep learning approaches for seizure detection, tracking, and localization are explored. Chapter 6 describes the SZTrack architecture, currently in review for publication in [10]. By using a combined convolutional and recurrent network structure for detecting seizure activity in each EEG channel, seizures can be tracked and coarse onset zones can be identified. Novel training techniques are described to train SZTrack to identify seizure activity at resolutions higher than the clincal annotations provided.

Chapter 7 presents SZLoc, an end-to-end neural network designed to identify SOZs directly from scalp EEG recordings. Extending approaches from Chapter 6, SZLoc employs both global and electrode signal paths to consider multi-scale information for localizing seizure activity. Specialized weakly supervised loss functions are developed to train SZLoc to localize seizures at a resolution higher than SZTrack using a multi-task learning framework. This work is currently being prepared for submission to MIDL.

In Chapter 8, a discussion of the work presented in this thesis is provided. Overall themes, commonalities, and differences between the approaches presented in each chapter will be discussed. Potential improvements and avenues for futher research will be provided, and the thesis will be concluded.

Work in seizure detection using a hybrid convolutional and recurrent neural network is presented in the Appendix. Originally published in [11], this work demonstrates that by combining convolutional and recurrent neural networks seizure detection performance can be improved. This result motivates the development of fully neural network approaches to seizure activity tracking. However as the network presented analyzes the multi-channel EEG signal exclusively, the network cannot be used for identification of seizure activity at the electrode channel and is excluded from the main text of the thesis.

Chapter 2

Background

This chapter will introduce relevant background material necessary for the exposition of the work presented in the subsequent chapters. The chapter is divided into five sections. In the first section an overview of the EEG signal and its use for diagnosis will be presented. The second section will introduce the problem of automated seizure detection and provide an overview of the major themes and approaches in the literature. The third section will present technical background regarding graphical modeling and present seizure detection techniques using this modeling approach. Section four provides a brief overview of deep learning and highlights its use in the seizure detection and EEG analysis literature to date.

2.1 EEG and Epilepsy

2.1.1 Physiology of the EEG Signal

Scalp electroencephalography was developed by Hans Berger in 1924. While his early reports of changes in electric activity corresponding to changing mental state or cerebral injury were received with some skepticism, the independent confirmation of his findings established EEG as a promising new methodology to investigate cortical activity [12]. From the 1930s to the 1970s, EEG served as the primary noninvasive study for localizing cerebral abnormalities. While CT and MRI scanning have emerged as primary noninvasive methods for localization structural abnormalities, EEG remains a central part of the diagnostic process for seizure

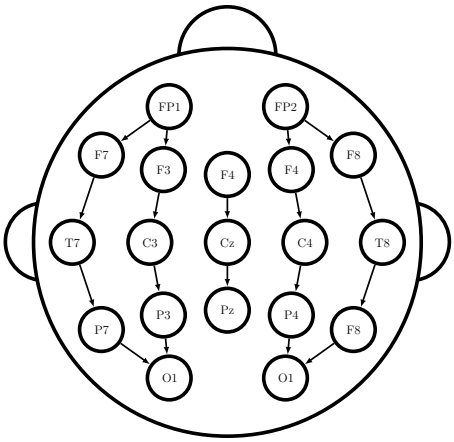


Figure 2-1. 10-20 System. EEG electrodes are placed in a standardized array. Difference channels used in the longitudinal bipolar montage are shown using arrows.

disorders [13].

While many electrical phenomena in the brain may contribute to its generation, the primary source of the EEG is believed to be excitatory and inhibitory post synaptic potentials (PSPs). As a neuron fires, an action potential travels down the axon to the synapse, triggering the release of a neurotransmitter. While this process is responsible for the direct communication between neurons, the short 10 ms spike of the action potential generates too weak a field to be recorded on the scalp. PSPs are responsible for raising and lowering the potential difference between the extracellular matrix and the cell body of the neuron. This action effectively adjusts the threshold of the signal required to trigger a spike in the subsequent neuron, leading to increased or decreased activity. As PSPs last 50–200 ms, populations of neurons exhibiting synchronized activity are capable of generating potential fields strong enough to be measured on the scalp [14].

To record EEG signals, electrodes are placed on the scalp in a standardized array known as the 10-10 or 10-20 system [15]. Electrodes are placed evenly on the scalp in order to measure the electrical signal from different regions of the cortex. The 10-20 electrode placement system is shown in Figure 2-1. As single electrode potentials must be referenced to a voltage, in many cases the average potential of all electrodes will be subtracted from each electrode. Alternatively, difference channels between electrodes can be taken, allowing cortical activity

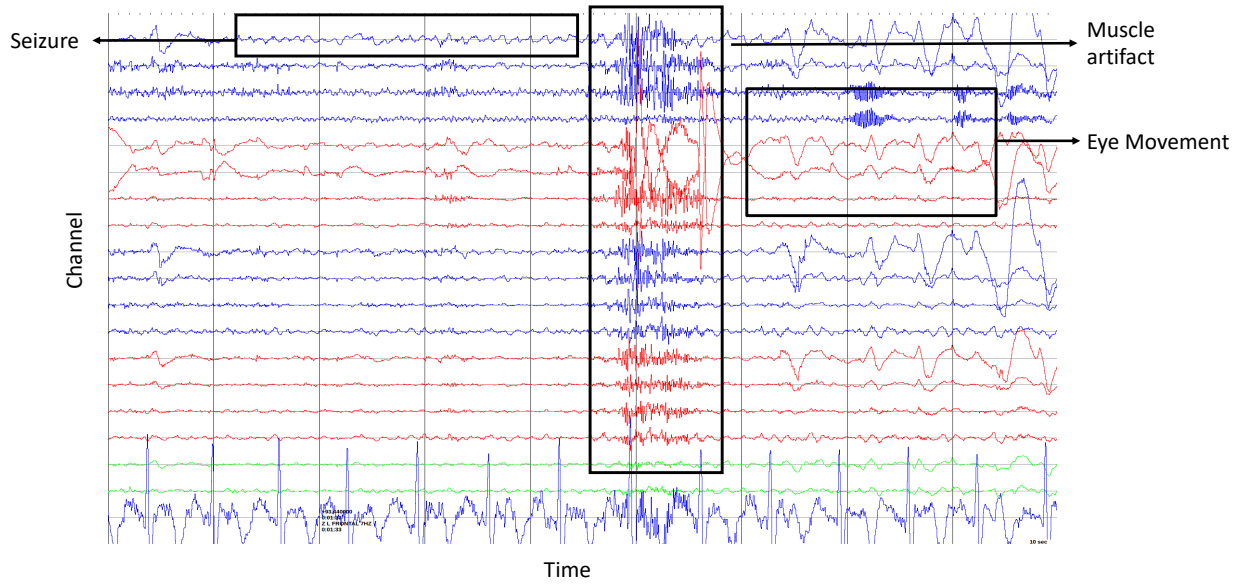


Figure 2-2. Seizure and arifact signals. EEG channels from the longitudinal bipolar montage are shown. Rhythmic 6–8 Hz seizure activity can be observed originating in the top channels. This rhythmic activity is followed immediately by muscle artifact throughout all channels and by eye movement artifact in the frontal channels.

to be localized. In Figure 2-1, arrows between channels indicate the positive and negative components of difference channels used in the bipolar longitudinal montage. For example, in this montage channel F7 would be subtracted from Fp1 to form the difference channel Fp1-F7. While alternative montage are used in clinical localization, in this thesis EEG signals will be used in either the average reference or longitudinal bipolar montage.

Complicating the analysis of EEG signals is the presence of high amplitude artifacts which obscure the underlying brain activity. This is illustrated in Figure 2-2, where muscle and eye artifacts of much greater signal amplitude immediately follow seizure onset. The original seizure signal is subsumed beneath these artifacts making tracing the propagation of seizure activity difficult. Though ocular and muscle artifacts are the most commonly occurring, other sources of EEG artifacts include cardiac activity, tongue movement, electrical leakage, changes in impedance due to sweat, and sources of noise stemming from faulty electrodes.

While many artifact removal techniques have been developed [16], these methods have more acceptance in research applications and are not as often employed in the clinic.

2.1.2 Epilepsy and the Clinical Role of EEG

Epilepsy is a heterogeneous neurological disorder characterized by spontaneous bursts of neuronal synchrony in the brain that manifest as seizures [17]. Nearly 3.4 million people in the United States, or 1.2% of the population, are believed to have active cases of epilepsy [18]. 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 [19].

The underlying causes of epilepsy are manifold, resulting in a wide variety of patient presentations. At a high level, epilepsy can be divided into two categories, generalized and focal. While diverse in their presentations, generalized seizures are marked by epileptic activity manifesting throughout the cortex concurrently. By contrast, focal seizures originate from a single location in the brain. Onset in focal epilepsy is typically marked by local rhythmic activity, such as in the theta or alpha band. This epileptic activity can to spread to adjacent regions of the brain and potentially involve the entire brain, resulting in secondarily generalized seizures [13]. Focal epilepsy affects approximately 60% of all patients with epilepsy [20].

While many patients have seizures controlled with medication, roughly 30% of patients with epilepsy have medically refractory epilepsy [2] and do not achieve seizure freedom with anti-epileptic drugs [2]. Alternative treatments for focal epilepsy rely on our ability to detect, and localize seizure activity in their brains. Namely, if we can determine that the seizures originate from a discrete SOZ, then the most effective treatment may be to surgically remove this region [20].

Scalp electroencephalography (EEG) plays a critical role in determining the course of

Figure 2-3. A patient undergoing continuous EEG monitoring. Image downloaded from https://consultqd.clevelandclinic.org/ in November 2021

treatment for medically refractory epilepsy patients. 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 [3]. 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 [15], 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. A patient undergoing clinical continuous EEG monitoring is shown in Figure 2-3

Identification of the seizure and its temporal evolution in scalp EEG is key for epilepsy diagnosis. However, due to the rarity of epileptic activity, the length of recorded EEG, and the subtlety of epileptic activity, analyzing continuous scalp EEG recordings is time consuming and requires extensive training. As a result of these difficulties, inter-rater agreement among clinicians can be low [21, 22] requiring labor-intensive review and discussion. Thus accurate automatic seizure detection and localization has the potential to save clinician time and improve the diagnosis and management of epilepsy.

2.1.3 Additional Non-Invasive Modalities for SOZ Localization

While scalp EEG is of primary importance for epilepsy diagnosis and SOZ localization, physicians rely on concordance between multiple sources of localization information to reliably treat the condition. As different locations of seizure onset result in differing clinical manifestations, seizure semiology offers clinicians a way to coarsely localize SOZ to lobes and hemispheres of the brain. Imaging modalities such as Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and Single Photon Emission Computed Tomography (SPECT) allow for the identification of lesions potentially responsible for epileptic activity. Similar to EEG, magnetoencephalograpy (MEG) records magnetic fields from different locations in the brain. Finally, source localization in MEG and EEG shows potential in providing clinically useful localization information. After acquisition, concordance between evidence for SOZ location from these sources is discussed in clinical conferences and treatment plans are agreed upon by teams of clinicians.

During inpatient EMU admission, video monitoring alongside EEG acquisition has become standard. Synchronous monitoring of electrographic and paroxysmal activity allows clinicians to assess seizure types and generate hypotheses of seizure localization [23]. Seizure semiologies, i.e. the outward clinical signs accompanying epileptic events including auras, changes in consciousness, and motor activity, are diverse and often indicative of potential seizure foci [24]. Video-EEG allows clinicians to assess this semiology, increasing the effectiveness of inpatient monitoring [25].

High quality structural MRI has become standard in the clinical evaluation of epilepsy patients [26] and is considered mandatory by many clinicians for a wide variety of etiologies [27]. If a lesion is found in concordance with EEG based localization, resective surgery may be planned provided the region is not near the eloquent cortex. Studies have shown that the identification of a lesion in MRI greatly improves the odds of seizure freedom after resective surgery [28]. However, 15–30% of focal epilepsy patients are non-lesional and require

alternative methods for discovering additional localizing information [5].

When no structural lesion is found, PET can be used to assess candidate areas for SOZ. By visualizing radioactive tracers within the body, PET can be used to image in vivo processes such as metabolic processes and blood perfusion. Interictal areas of hypometabolism in the onset zone were among the earliest clinical findings for PET imaging [29], though the exact mechanism is not entirely understood. While areas of hypometabolism revealed by FDG-PET can indicate SOZ, these areas are typically wider than the seizure focus itself, complicating surgical planning from PET imaging alone. However, due to its wide usage in oncology centers, PET remains a commonly available and popular imaging technique in addition to MRI [5].

Ictal SPECT imaging can provide information regarding dynamic changes in cerebral perfusion during an epileptic seizure. Similar to PET imaging, SPECT uses radiotracers injected into the blood stream to visualize in vivo processes occurring in the body. Using SPECT imaging, changes in blood flow before, during, and after a seizure can be recorded. When a seizure is captured, an epileptogenic region can be indicated by an area of hyperperfusion, indicating increased brain activity, surrounded by an area of hypoperfusion. However, ictal SPECT imaging is difficult to achieve, as the radiotracer takes roughly 40 seconds to reach the brain. Thus the timing of the radiotracer injection complicates the acquisition of true ictal imaging [5].

Where EEG measures electric potentials from the scalp using a standardized system of electrode placement, MEG uses an array of Superconducting QUantum Interference Devices (SQUIDs) to measure magnetic fields produced within the brain. As magnetic fields are less attenuated by the brain, skull, and scalp, MEG signals can be more accurately localized within the brain. However, where EEG electrodes are applied directly to the scalp, SQUIDs are housed in large apparatuses within magnetically shielded rooms and require the patient to remain as still as possible while recording. Thus while shown to provide complimentary localization information to EEG [30], MEG has yet to become as widely used in epilepsy

diagnosis due to the difficulties in ictal MEG acquisition and limited availability of devices [31].

Source imaging from EEG signals is a method of localizing activity recorded in scalp EEG to locations inside the brain. Research in electric source imaging has typically focused on improving the EEG spatial resolution (typically 20–40 sensors) by deconvolving the signals into current dipoles [32, 33] or distributed sources [34–36] at the millimeter scale. However, these *inverse solvers* are sensitive to physiological noise, the number of EEG channels, and the underlying head model [37, 38]. Hence while studies have noted diagnostic value added, source imaging has yet to be widely adopted in the clinic due to difficulties in its interpretation and the lack of available automated methods [39].

2.2 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 [40]. 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 [41] 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 [42]. Below we detail common approaches in feature extraction and classification.

2.2.1 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 [14]. 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 [43] and [44] use the FFT to compute power in the 2.5–12 Hz band of each EEG channel. Thresholding techniques developed in [44] were applied to find periods of seizure activity within long-term recordings. A more sophisticated approach uses filter banks to compute the spectral power in different frequency bands. In [45], 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 [46], 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 [47], energy and spectral features were calculated for each wavelet subband and used for classification in an array of classifiers. [48] follows a similar approach, extracting amplitude features for classification after performing a multichannel empirical wavelet transform to the original EEG signal. Similarly, [49] 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 [14] but cannot be captured by methods that focus on just the power spectrum.

2.2.2 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 [50]. Non-linear signal processing techniques quantify the predictability of the system. For example, [50] uses

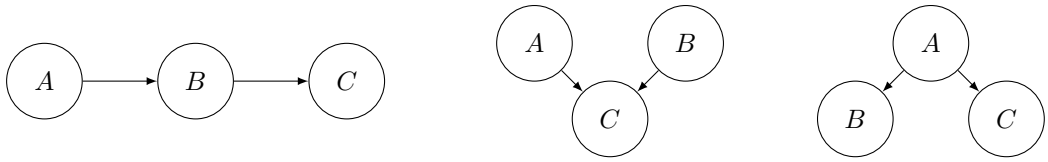
approximate entropy, sample entropy, and phase entropy to measure the similarity of the EEG to its past behavior. It is noted that approximate and sample entropy are lower for non-seizure intervals, indicating a more predictable signal than that of ictal EEG.

Finally, many studies have combined time-frequency and time domain features to leverage advantages from each representation for gains in detection performance. For example, [51] 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 [52, 53] and [54] extract non-linear features following the decomposition of the original EEG signal into separate wavelet bands. Both [52] and [53] compute the correlation dimension and largest Lyapunov exponent for each subband after application of the wavelet transform. Similarly, [54] 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.

2.2.3 Classification

Many classification strategies have been used for seizure detection. SVMs are perhaps the most popular classifier, finding use in [45, 55, 56]. 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 [57].

The random forest classifier is another popular choice [51, 58]. 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



(a) $P(A, B, C) = P(A)P(B \mid A)P(C|B)$

(b) $P(A, B, C) = P(A)P(B)P(C \mid A, B)$

(c) $P(A, B, C) = P(A)P(B \mid A)P(C \mid A)$

Figure 2-4. Example graphical models and factorizations. Directed connections indicate the conditional independencies between the random variables of the network.

[59]. Other classifiers for seizure detection include adaptive thresholding [43], which updates a thresholding parameter based on previous intervals, and k-nearest neighbors [58], 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.

2.3 Bayesian Networks and Applications in EEG Signals

2.3.1 Directed Acyclic Graphical Models

To efficiently perform inference in ensembles of jointly distributed variables, conditional independencies between variables must be taken into account. Directed Acyclic Graphs (DAGs), or Bayesian networks, offer an intuitive means of visually representing statements about conditional independence. If a directed edge exists between two nodes, the originating node of the edge is referred to a s the parent while the downstream node is referred to as the child node. Let the children of a node X be defined as $\text{ch}(X)$ while the parents are defined as $\text{pa}(X)$. The ancestors of a node are defined as the nodes which can be reach through successive parent relationships while the descendants are defined as the nodes reachable through child relationships. Let the ancestors of a node X be defined as $\text{anc}(X)$ while the descendants are defined as $\text{desc}(X)$. Crucially, DAGs allow Markov properties encoding the

conditional independencies of jointly distributed random variables to be visualized easily. A node is conditionally independent of any ancestors higher than its parents given its parents, or mathematically $X \perp \text{anc}X \backslash \text{pa}X \mid \text{pa}X$. Using the Markov property, convenient factorizations of complicated joint distributions can be derived.

Consider the simple DAG examples shown below in Figure 2-4 with random variables A, B, and C. In Figure 2-4 (a), by the Markov property C is independent of A given its parent B. This yields the convenient factorization of the joint probability $P(A, B, C) = P(A)P(B \mid A)P(C|B)$. Similarly, Figure 2-4 (b) depicts the case where $A \perp B$ and C is conditionally dependent on both A and B yielding, $P(A, B, C) = P(A)P(B)P(C \mid A, B)$. Finally, the example in Figure 2-4 (c) A is the exclusive parent of B and C leading to the factorization $P(A, B, C) = P(A)P(B \mid A)P(C \mid A)$. While the simple examples shown here represent small joint distributions with only three variables, the basic structures illustrated appear in more complex distributions with much greater numbers of random variables. Thus understanding the conditional independence statements implied by these simple graphs will be useful as more complicated DAGs are introduced.

2.3.2 Hidden Markov Models for Sequence Modeling

With usage in speech, natural language processing, and bioinformatics [57, 60], the Hidden Markov Model (HMM) is a widely popular Bayesian Network for modeling the evolution of sequences. A graphical model depicting the HMM is shown in Figure 2-5. In an HMM, there is an underlying sequence of unobserved states which evolve over steps of the sequence. At each step, an observation is made which depends on the underlying state. In this thesis, the basic HMM structure will be used to model the observed EEG signal given the occurence of seizure, modeled by the evolving underlying states of the model.

The model consists of observed nodes $X[t]$ and hidden nodes $Y[t]$ for times $t = 0, \ldots T$. Nodes $Y[t]$ form a Markov chain of discrete states. At every timestep, an observed emission $X[t]$ is generated from an emission likelihood $P(X[t] \mid Y[t])$, i.e. $X[t]$ is conditionally

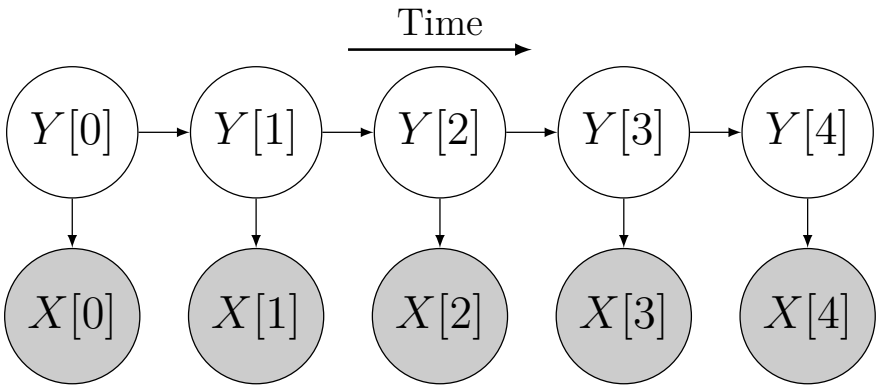


Figure 2-5. Graphical model depicting an HMM. Observed variables, $X[t]$, are shown shaded. Hidden variables, $Y[t]$, are unshaded.

independent of all other variables in the model given the latent state of the Markov chain at t. Given a distribution over the initial state $Y[0]$, the joint probability distribution of ensemble variables $\mathbf{Y}$ and $\mathbf{X}$ factorizes as:

$$P(\mathbf{Y}, \mathbf{X}) = P(X[0] \mid Y[0])P(Y[0]) \prod_{t=1}^{T} P(X[t] \mid Y[t])P(Y[t] \mid Y[t-1]).$$

The conditional distribution $P(Y[t] \mid Y[t-1])$ governing transitions between latent states is often expressed by a stochastic transition matrix A, where $P(Y[t] = j \mid Y[t-1] = i) = a_{i,j}$.

The forward-backward algorithm [57, 60] is used for exact marginal inference in the HMM. At a high level, this algorithm uses dynamic programming to propagate information back and forth along the HMM chain using forward and backward message passing. More details about this procedure can be found in [57, 60]. The forward messages $\alpha_i[t](k)$ and backward messages $\beta_i[t](k)$ are computed via the following recursions:

$$\begin{aligned}
\alpha[t](i) &\coloneqq P(X[0], X[1], \ldots, X[t], Y[t] = i) \\
&= \sum_j P(X[t] \mid Y[t] = i) a_{j,i} \alpha[t-1](j) \\
\beta[t](i) &\coloneqq P(X[t+1], X[t+2], \ldots, X[T] \mid Y[t] = i) \\
&= \sum_j \beta[t+1](j) P(X[t+1] \mid Y[t+1]) a_{i,j}.
\end{aligned}$$

The data likelihood can be easily calculated by noting that $P(X[0], X[1], \ldots, X[t]]) = \sum_j \alpha[t](j)$. The singleton and pairwise marginals, $\gamma[t](i)$ and $\xi[t](i,j)$ respectively, are

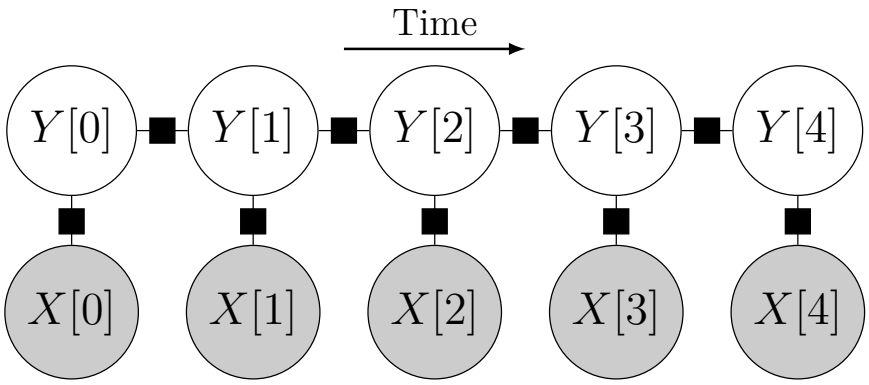


Figure 2-6. Factor graph model depicting an HMM. Observed variables, $X[t]$, are shown shaded. Hidden variables, $Y[t]$, are unshaded. Factors are shown connection random variables, indicating the conditional probability relationships governing the HMM.

obtained by normalizing the following expressions:

$$\gamma[t](i) = P(Y[t] = i \mid \mathbf{X}) \propto \alpha[t](i)\beta[t](i)$$
$$\xi[t](i,j) = P(Y[t] = i, Y[t+1] = j \mid \mathbf{X})$$
$$\propto \alpha[t](k)a_{i,j}P(X[t+1] \mid Y[t+1] = j)\beta[t+1](j).$$

2.3.3 Factor Graphs and Message Passing

Factor graphs are useful generalizations of many types of graphical models. In a factor graph, random variables are shown as circle nodes. The relationships between these variables are illustrated through connected factors, shown as black boxes. These factors can take many forms, e.g. conditional distributions as in Bayesian networks or factor potentials as in Markov random fields. A factor graph representation of the HMM covered in the previous subsection is shown in Figure 2-6. The structure of the model is almost identical, however arrows representing conditional probabilities have been replaced with factors connecting related random variables.

Inference in a factor graph can be performed, either exactly or approximately, via the sum-product algorithm. In tree structured factor graphs, those without cycles such as the HMM, the sum-product algorithm is capable of performing exact inference. Indeed, it can be shown that the forward-backward algorithm detailed in the previous subsection is a special

case of the more general sum-product algorithm. In this algorithm, messages are passed between variables and factors which update the current belief state of the graph. Messages take different forms depending on whether they originate at a variable or a factor. Messages from a generic variable X to a generic factor f take the form

$$\mu_{X\to f}(x) = \prod_{h\in ne(X)\setminus f} \mu_{h\to X}(x) \quad . \tag{2.1}$$

All neighboring factor messages h into X except for the one from f are multiplied together to generate the message $\mu_{X\to f}(x)$. Messages from factors to variables require summarization over the variables V sharing the factor f.

$$\mu_{f\to X}(x) = \sum_{V_{ne(f)\setminus X}} f(V_{ne(f)}) \prod_{V\in ne(f)\setminus X} \mu_{V\to f}(v) \tag{2.2}$$

This procedure is akin to marginalization, as sums over all possible settings of the neighbors of f are taken to generate the message to X. To take posteriors over a single variable, all messages into that variable are multiplied and normalized.

$$P(X = x) \propto \prod_{h\in ne(X)} \mu_{h\to X}(x) \tag{2.3}$$

Pairwise marginals can be computed similarly. Let X and Y be variables that share factor $f(x, y)$.

$$P(X = x, Y = y) \propto f(x, y) \prod_{h\in ne(X)\setminus f} \mu_{h\to X}(x) \prod_{g\in ne(Y)\setminus f} \mu_{g\to Y}(y) \tag{2.4}$$

By passing messages completely from one end of a tree structured factor graph and back, exact inference can be performed.

2.3.4 Bayesian Models for EEG Analysis

Many approaches to seizure detection relying on the basic HMM structure have been proposed in the seizure detection literature. In [61], wavelet features were used to train an HMM model for detecting seizures in epileptic rats. Similarly, [62] trains HMMs to classify seizures from the CHB-MIT dataset using features derived from time-domain and chaos theory signal

processing. Using topographic maps of activity in individual electrodes, [63] develops an HMM based seizure detector trained on propagations of epileptic activity in the brain. Though this approach considers seizure propagation in seizure detection, this propagation must be provided to the HMM a priori, as a single HMM chain applied across multi-channel features is incapable of learning this propagation. These models are capable of analyzing the temporal evolution of seizure activity but due to their reliance on a single HMM chain for detection, fail to capture spatial activity.

Extensions to the HMM have been applied to several problems in EEG analysis. For example, the work of [64] develops an Autoregressive HMM (AR-HMM) to model the changing correlation structure in raw EEG data as an unsupervised way to discover different dynamical regimes in the EEG signal. However, no labels were used in training the model, so expert labeling was required to apply the learned states to related problems. For example, to apply the outcome of this model to seizure prediction in canine EEG, the pre-seizure states had to be manually identified from the AR-HMM output.

The Coupled Hidden Markov Model (CHMM) extends the original HMM formulation to include multiple latent chains. Within the context of EEG, each latent chain corresponds to a single EEG electrode. Coupling is defined such that each latent chain may be affected by and may influence the states of other chains. CHMMs made their debut in modeling audio-visual relationships [65]. In another domain, the work of [66] proposed a CHMM to model the spread of infectious disease by defining a coupling structure based on the physical proximity of individual people.

Small two channel CHMMs have been used in behavioral EEG experiments [67]. In addition, the work of [68] developed a distance coupled HMM to model EEG signals in both alcoholics and their healthy peers. Classification was performed between groups by assigning a test sequence to the model class under which its likelihood was maximum. While capable of discerning patients from controls, these models were designed to classify entire EEG sequences and not to label pathological activity within a single recording. In addition,

these works simplify the analysis to just two EEG channels, for which exact inference using the forward-backward algorithm is tractable. However, the methodology outlined in [68] does not generalize to more dense recordings.

2.4 Deep Learning, A New Direction for Seizure Detection

Though the origins of artificial neural networks (ANNs) date as early as the 1950s [69], in the last two decades deep learning has emerged as a dominant trend in machine learning. This trend was fueled by advances in image classification [70] and [71] speech applications. Where previous approaches to machine learning pair hand crafted feature extraction techniques with statistical classifiers, deep learning seeks to learn classification functions from large models composed of a succession of relatively simple computational layers. By learning these layers directly from the data, the neural network learns hierarchical representations of the data itself. This process removes the need for feature crafting, allowing machine learning systems to learn more expressive functions capable of classification performance exceeding hand crafted approaches.

In deep learning, the goal is to approximate some optimal function $y = f^*(x)$ with a succession simple of layers. After the application of each layer, higher order representations of the original data are learned. These layers are each parameterized by a set of weights. For example, a three layer neural network is composed of the application of three layers, $f(x;\theta) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$, where each layer operates on the output of the previous layer. Typically, three main types of layers, feedforward or linear, convolutional, or recurrent, are used. The choice of layer type is dependent on the structure of the data and the task at hand. In the following subsections, each layer type and its parameterization will be described. A relatively new feed forward architecture, the transformer, will then be described. A brief discussion of loss functions and training via backpropagation will be provided. Finally, recent work in deep learning for seizure detection and other EEG applications is discussed.

2.4.1 Fully Connected Neural Networks

The simplest type of neural network is the fully connected neural network. Also known as the ANN or multilayer perceptron, these networks consist of successive linear layers with non-linear activation functions applied. Let $h^{(l-1)}$ represent the hidden representation after the application of layer $l-1$ in the network. The hidden representation for the next layer, $h^{(l)}$, is given by

$$h^{(l)} = g(h^{(l-1)}A^T + b) \tag{2.5}$$

where A is a matrix of learnable weights, b is a learnable bias, and g is a non-linear activation function.

The nonlinear activation function $g(\cdot)$ plays a critical role in the fully connected network. Without the nonlinearity, fully connected networks are capable of learning affine transformations of the input data. However, it has been shown that with the addition of nonlinearities, fully connected neural networks are universal approximators capable of representing arbitrary functions. Thus these nonlinearities are essential for the expressive power of neural networks.

Five common nonlinearities are shown in (2.6). The sigmoid function $\sigma(x)$ was an early choice. By restricting outputs to the range of 0 to 1, the sigmoid function saturates at high input values, mimicking the spiking action of biological neurons. Similarly, the $\tanh(x)$ function saturates at high values. However, the range of the tanh function extends from $(-1, 1)$, allowing the output of each layer to take negative values. Noting that the slope of the sigmoid and tanh function is close to 1 near 0, the ReLU activation offers a simple alternative to these more complicated functions. Due to its simplicity, the ReLU function reduces computation time while maintaining the universal approximation property of multilayer neural networks. However, the ReLU takes a value of 0 for inputs less that 0, potentially nullifying of these nodes in the network. To avoid this problem, the LeakyReLU modifies the ReLU by scaling inputs less than 0 by a slope factor α, allowing negative input values to result in varying output values. Finally the PReLU activitation takes the same form as the

LeakyReLU but allows the parameter α to be learned.

$$\begin{aligned}
\text{Sigmoid}(x) = \sigma(x) &= \frac{1}{1+\exp(-x)} \\
\tanh(x) &= \frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)} \\
\text{ReLU}(x) &= \max(0,x) \\
\text{LeakyReLU}(x) &= \max(0,x) + \alpha \min(0,x) \\
\text{PReLU}(x) &= \max(0,x) + \alpha \min(0,x),\ \alpha \text{ learnable}
\end{aligned} \tag{2.6}$$

2.4.2 Convolutional Neural Networks

Two dimensional convolutional neural networks as approximations of the way the brain interprets images Popular for image applications. LeNet [72] was developed for handwritten digit recognition. Showed the efficacy of CNNs trained with backpropagation. AlexNet [70] greatly surpassed previous performance on the ImageNet challenge [73]. Trained on GPUs, this breakthough led to more deep learning. VGG Net showed that increasing depth leads to better performance [74]. Resnet introduced residual connections in order to trained much deeper networks [75]. Following ideas from computer vision, one dimensional CNNs have found applications to time-series signals as well [76].

Let $h_c^{(l)}$ represent the hidden representation at layer l in channel c. Let $w_{c,k}$ be the learnable convolution kernel for output channel c depending on input channel k. $h_c^{(l)}$ depends on the hidden representation in the previous layer, $h^{(l-1)}$, through the summation of convolutions over each channel k.

$$h_c^{(l)} = b_c + \sum_{k=0}^{C_{in}-1} w_{c,k} \star h_k^{(l-1)} \tag{2.7}$$

Through this formulation, CNNs learn decision functions which are symmetric to spatial and temporal transformations of the original inputs.

2.4.3 Recurrent Neural Networks

While multilayer perceptrons and CNNs have proven themselves powerful tools in AI applications, these feedforward architectures assume data samples are independent identically distributed (IID) instances with no sequential relationship. In both MLPs and CNNs input and output dimensions must be fixed and constant across the data. In contrast, Recurrent Neural Networks (RNNs) are popular architectures for sequence analysis capable of analyzing variable length inputs and producing variable length outputs. For each element of the sequence, the RNN maintains a hidden representation of the signal. This hidden representation is continually updated based on its past value, thus allowing information from previous timepoints to inform the output of the RNN at each time. Analogous to the universal approximation theorem for MLPs, it has been shown that RNNs are capable of learning arbitrary mappings from input to output sequences [77]. As RNNs operate on a sequence of inputs in single temporal direction, a natural extension is the bidirectional RNN [78]. By jointly training two networks, one in the forward temporal direction and one in the reverse, the combined network is capable of analyzing information from the entire sequence as opposed to just previous inputs.

A popular RNN architecture is the long short-term memory (LSTM) network introduced in [79]. The LSTM network has been used in speech [80] and natural language processing applications [81] and as well as EEG [82]. The LSTM cell incorporates an input, output, and forgetting gate to allow the network to learn effectively from longer sequences. Equations governing the LSTM are given in 2.8. Each LSTM cell receives the current data x_t and the previous hidden state h_{t-1} as input. Along with the previous value of the internal cell state c_{t-1}, these inputs govern the behavior of cell. Values of the input, output, and forgetting gate, i_t, o_t, and f_t are computed using learnable weights $W_{..}$ and biases $b_{..}$. For these weights and biases the first subscript denotes the input source while the second denotes the effected output. An update to the internal cell state g_t is computed using a similar set of weights and biases. The cell state c_t is updated using a combination of the previous cell state c_{t-1}

and g_t weighted respectively by the values of the forgetting gate f_t and the input gate i_t. Finally, the cell state c_t is passed through a tanh non-linearity and multiplied by the value of the output gate o_t to form the final hidden state of the cell h_t. As with RNNs, pairs of LSTM networks can be trained in both the forward and backward temporal directions to form bidirectional LSTM (BLSTM) networks.

$$\begin{aligned} i_t &= \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \\ f_t &= \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \\ g_t &= \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \\ o_t &= \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{2.8}$$

Gated Recurrent Units (GRUs) are a more recently introduced recurrent neural network [83] and have become a widely adopted alternative to LSTMs. GRUs simplify LSTM cells by removing a gate and removing the distinction between in hidden and memory states. Despite this simplification, they have been shown to perform comparably to LSTMs in some cases [84]. In addition, due to their relative simplicity it has been hypothesized that GRUs may have some advantages in generalization performance when compared to LSTMs when applied to small datasets.

In the GRU cell, reset and update gates, r_t and z_t, are computed based on the previous value of the hidden state, $h_{(t-1)}$, and the input x_t. By using the sigmoid function, these gate values range between 0 and 1. The values of these gates controls how the GRU weights incoming information from the previous hidden state and the value of the new observed datapoint. An update to the hidden state n_t is computed based on the input data x_t and the previous state $h_{(t-1)}$ weighted by the value of the reset gate. If the reset gate is near zero, only information from the the new datapoint will be considered in the update. Finally a linear combination of the previous hidden state and the update n_t are computed according

to the value of the update gate z_t. Mathematically,

$$\begin{aligned} r_t &= \sigma\left(W_{ir}x_t + b_{ir} + W_{hr}h_{(t-1)} + b_{hr}\right) \\ z_t &= \sigma\left(W_{iz}x_t + b_{iz} + W_{hz}h_{(t-1)} + b_{hz}\right) \\ n_t &= \tanh\left(W_{in}x_t + b_{in} + r_t\left(W_{hn}h_{(t-1)} + b_{hn}\right)\right) \\ h_t &= (1 - z_t) * n_t + z_t * h_{(t-1)} \end{aligned} \tag{2.9}$$

Thus the GRU unit weights incoming data according to the previous values of the hidden state to produce a new hidden representation.

2.4.4 Transformers

The transformer architecture is a recently developed architecture introduced in [85]. The transformer is able to encode sequential relationships in data using completely feedforward networks. Thus transformer networks reduce the computational complexity required during the forward and backward passes when compared to recurrent networks. Furthermore, as recurrent networks rely on an autoregressive structure for propagating information in sequential data, transformers are capable of analyzing entire sequences at once. This improvement mitigates some of the difficulties involved in training recurrent networks, as gradients must retain information along multiple timesteps during training.

In natural language processing, BERT has become an incredibly popular architecture for language modeling [86]. In computer vision applications, the image transformer has shown capabilities in image generation, inpainting, and other imaging applications [87]. In multi-modal vision and text applications, ImageBERT [88] and VideoBERT have shown efficacy in image and video captioning, respectively.

Fundamentally, the transformer architecture relies on scaled dot-product attention. Scaled dot-product attention computes similarity weights between a set of queries Q and keys K. These similarities are scaled using a softmax operation such that they sum to 1 and multiplied with a set of values V. Thus the output of the attention layer is a linear combination of the

values, as weighted by the similarity between the keys and queries.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.10}$$

Multi-head attention extends scaled dot-product attention to allow relationships to be learned between the keys, queries, and values. For each head i, Q, K, and V are projected using learnable weight matrices W_i^Q, W_i^K, and W_i^V. Attention is then computed based on the projected values QW_i^Q, KW_i^K, and VW_i^V. The outputs of each attention head are then concatenated and combined using the weight matrix W^O.

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \\ \text{where head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2.11}$$

The transformer architecture employs multi-head attention in a series of encoder layers, followed by a similar series of decoder layers. In the encoder layers, input features are projected into a new representation capable of informing the downstream decoder. Each encoder layer consists of two sublayers, a multi-head attention layer and a feed forward network. Residual connections are added around each of these sublayers followed by application of layer normalization [89]. Let x be the input and e represent the encoder output.

$$\begin{aligned} h &= \text{LayerNorm}\left(x + \text{MultiHead}(x, x, x)\right) \\ e &= \text{LayerNorm}\left(x + g(hA^T + b)\right) \end{aligned} \tag{2.12}$$

In this formulation, it is common to set attention weight matrices W_i^Q, W_i^K, W_i^V and feedforward parameters such that input, hidden, and output dimensions remain a consistent size d_{model}.

The decoder layer follows a similar structure but adds an intermediate sub-layer which compares decoder inputs to outputs from the encoder layers. Decoder inputs are first passed through a multihead attention layer where the input is used as the queries, keys, and values. Next an additional multi-head attention is applied to combine encoder outputs with hidden decoder representations. Outputs from the encoder serve as queries and keys while the hidden state of the previous layer serves as the values. Finally a feed forward network is applied.

As in the encoder layer, each of these sub-layers is outfitted with a residual connection and subsequent LayerNorm operation. Again let x be inputs to the decoder, e be the outputs from the previous decoder layer, and o be outputs from the decoder. Mathematically,

$$\begin{aligned} h_1 &= \text{LayerNorm}\left(x + \text{MultiHead}(x, x, x)\right) \\ h_2 &= \text{LayerNorm}\left(h_1 + \text{MultiHead}(e, e, x)\right) \\ o &= \text{LayerNorm}\Big(x + g(hA^T + b)\Big) \end{aligned} \tag{2.13}$$

By combining a series of encoder and decoder layers, information from one source can be encoded such that it may be used to inform the behavior of the neural network given a different source of information. While this architecture first found success in machine translation and other NLP applications, it will be used in this thesis to compare multi-channel EEG information.

2.4.5 Training

Due to the large number of parameters and relative lack of structure, network training is a challenging problem of its own. While many approaches have been proposed, the backpropagation algorithm [90] allows the gradient of such large networks to be computed with a single backward pass through the network. The efficient computation of the gradient allows for gradient based approaches such as gradient descent or stochastic gradient descent. Recently, the Adam optimization algorithm [91] has become a popular for training networks. By scaling the gradient of each parameter according to a decayed estimate of the gradient mean and variance, the Adam algorithm stabilizes and speeds up training. In general, the performance of these training algorithms depends on the size of minibatches and learning rates used, further adding to the search space of possible network training configurations.

Orthogonally, regularization and dataset augmentation can help large neural networks generalize to unseen data. By randomly setting hidden representations by 0, dropout [92] forces networks to learn robust representations of the input data less susceptible to noise. Dataset augmentation techniques [93] seek to add variability to the training data, thus forcing

the network to learn hidden representations invariant to shifts in the data. Examples from computer vision include rotations, mirror imaging, zooms, and translations. Finally, weight decay has been shown to improve generalization in some applications by applying an ℓ_2 penalty to the weights of the network during training.

2.4.6 Deep Learning for EEG Applications

Driven by successes in domains such as computer vision and natural language processing, deep neural networks have come to dominate the machine learning field [94]. This interest in deep learning has extended into EEG analysis, finding applications in brain computer interfaces, sleep state analysis, and seizure detection [95]. 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 [58], 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 [96] 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.

To eliminate the need for hand designed feature extraction, CNNs have been applied to the EEG signal. 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, [97] constructed a CNN to operate on Short-Time Fourier Transform (STFT) spectrograms for seizure detection. Similarly, in [98] 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 [99] where a wavelet decomposition was used to build a spectrogram for the CNN input. [100] 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. In [101], single channels of intercranial EEG are classified using a one dimensional CNN. In [102] 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 [103], 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.

Similarly, RNNs have been found to be effective architectures for EEG applications. In [104], 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 [82] 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 [105] and [106]. In [106], a CNN and LSTM layer were combined to perform seizure detection. Specifically, long windows (101 seconds) 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 second sequence.

In contrast the authors of [105] create STFT images that span 30 seconds 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 second segment. [107] 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 second 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 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.

2.5 Preliminaries

2.5.1 Epilepsy Datasets

Here we introduce the three EEG datasets used throughout this thesis. Experiments reported in the following chapters will use these datasets and as such they are introduced here. Each dataset was recorded at a separate location and contains patients with different demographic and clinical characteristics. The main dataset used throughout this thesis is the Johns Hopkins Hospital (JHH) dataset, which contains adult patients with focal epilepsy. Similarly, the University of Wisconsin-Madison (UWM) dataset contains EEG recorded from pediatric focal epilepsy patients. The third dataset contains recordings from pediatric patients at the Children's Hospital in Boston collected in conjunction with MIT (CHB-MIT). In the following subsections, the datasets are described in more detail.

2.5.1.1 JHH Dataset

Our primary EEG dataset consists of 201 seizure recordings obtained from 34 focal epilepsy patients undergoing presurgical evaluation in the Johns Hopkins Hospital between 2016–2019.

Table 2-I. Patient demographics and clinical attributes for our JHH evaluation dataset (N=34).

	JHH Dataset
Seizure Type	Focal Epilepsy
Has localizations?	Yes
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 per Patient	1.8 ± 1.8 hours
Average Seizure Duration	112 seconds
Minimum/Maximum Seizure Duration	13/979 seconds
Indicated Conditions	Cavernoma, MTS, perventricular heterotopia, stroke, encephalocele, FCD, low grade glioma
Lesional/Non-Lesional/Unspecified	24/3/7
Temporal/Extra-Temporal	26/8
Posterior/Anterior	27/7
Right/Left	18/16

As this dataset was compiled over the duration of the work presented in this thesis, differing numbers of patients are used in earlier experiments. Table 2-I contains a summary of this dataset while Table 2-II contains patient level demographic and clinical information. Where available, we have included imaging notes from the patient medical record. Patients ranged from 6–77 years with a mean age of 35.7 ± 16.8 years. The dataset contains 18 females and 16 males. Inclusion criteria were that the patient was a candidate for a focal resection with planned intracranial monitoring in the future. As part of this criteria, all patients had a well-characterized seizure onset zone based on the available clinical data. Exclusion

Table 2-II. Demographic information and localization notes from the JHH dataset. Where available, other notes regarding imaging results or other clinically relevant information is provided as well.

Pt ID	Sex	Age	Localization Notes	Other Notes
1	M	50	Right temporal epilepsy, though hard to tell onsets	Right temporal lobe cavernoma
2	F	32	Complex partial seizures with right frontocentral onset	
3	F	52	Maximal over left temporal head region	Left MTS, left frontal and temporal injuries
4	F	25	Left frontocentral or left temporal, has periventricular nodular heterotopia	Left periventricular heterotopia
5	M	77	Right temporal	Right temporal ovoid focus
6	M	52	Posterior left parietal	White matter disease
7	M	38	At first difficult lateralization, later thought to be left fronto-temporal	Possible left MTS
8	F	46	Right temporal	History of stroke
9	M	39	Left frontal	Normal brain MRI
10	F	41	Left frontal	
11	F	29	Right fronto-temporal	Nonlesional MRI and PET
12	F	10	Left antero-temporal	Normal brain MRI
13	M	45	Right or left temporal onset	Early evidence of left MTS
14	F	12	Right temporal	Right MTS, right frontal white matter lesion
15	M	20	Left temporal	Abnormality in left amygdala, small middle fossa encephalocele
16	F	18	Right paracentral anterior frontal	Right frontal FCD
17	F	45	Left mid/posterior temporal	
18	M	74	Right temporal	MRI suggestive of right MTS
19	F	33	Right temporal	MRI suggestive of right MTS
20	F	19	Right temporal	MRI compatible with right MTS
21	M	51	Left temporal	Left MTS, left frontal and temporal injuries
22	M	49	Left temporal	
23	F	22	Left posterior parietal-temporal region (P9, P7)	Cortical thickening in left inferior occipitotemporal junction suggestive of FCD
24	M	21	Right temporal	
25	F	36	left anterior and mid temporal	Left lateral temporal low grade glioma
26	F	25	Likely left temporal, however had prior R surgery, probably has bilateral seizures	Right MTS
27	M	17	Right anterior to mid temporal lobe	Right MTS
28	F	44	Right temporal (had right temporal tip encephalocele)	Small right inferior temporal encephalocele
29	M	22	Right temporal	Right MTS, multiple cavernous malformations
30	F	41	Right temporal	Right MTS
31	F	35	Left temporal	
32	M	56	Right fronto-temporal, thought to be temporal due to imaging	MRI consistent with right MTS
33	M	6	Right temporal lobe	
34	M	33	Bilateral temporal, thought to be left due to imaging	Left temporal abnormality, possibly cortical dysplasia or low grade glioma

criteria included non-epileptic seizures, generalized epilepsy, and patients who were not deemed to be surgical candidates. Many patients in the dataset separately underwent MRI or PET imaging. These scans revealed a range of structural abnormalities, including but not limited to focal cortical dysplasia (FCD), mesial temporal sclerosis (MTS), white matter disease, encephalocele, and gliomas.

All EEG were recorded on Nihon Kohden EEG1200 recorders with 70 Hz low pass filter, and 0.016 high pass filter. The data was recorded at 200 Hz using the 10-20 electrode placement system [15]. For our analysis, the EEG recordings have been clipped to include roughly 10 minutes of pre-seizure and post-seizure activity. EEG data was collected with IRB approval during routine clinical care and was anonymized prior to analysis.

2.5.1.2 UWM Dataset

The UWM dataset consists of 53 seizure recordings from 15 pediatric patients admitted to University of Wisconsin-Madison (UWM) from February 2018 to December 2019. Patients ranged from 8–17 years with a mean age of 13 ± 3.1 years. The dataset contains 5 females and 10 males. Inclusion criteria included a suspected focal onset of the epileptic seizures, as characterized by expert review of the medical record. As the UWM dataset was drawn from a larger study of multimodal neuroimaging, inclusion criteria also included that the patient underwent MRI scanning with available T1 MRI and resting-state fMRI. However, the imaging data was not used in the present study. Of the 19 patients at UWM that met the inclusion criteria, 4 patients were excluded, 1 due to prior resection, 1 patient whose EEG recordings contained only auras, and 2 patients with indeterminate seizure onset zone. Similar to the JHH dataset, many patients had structural brain abnormalities visible on MRI and PET, such as FCD, encephalocele, gliosis, MTS, and encephalitis. Table 2-III summarizes the patient characteristics. Table 2-IV contains patient demographics, localization information, and where available, other clinical notes.

The EEG were recorded on on a Natus Xltek EMU40EX system with built-in high and

Table 2-III. Patient demographics and clinical attributes for our UWM evaluation dataset (N=15).

	UWM Dataset
Seizure Type	Pediatric Focal Epilepsy
Has localizations?	Yes
Number of patients	15
Average Age	12.8 ± 3.1 years
Minimum/Maximum Age	8/17 years
Number of Males/Females	10/5
Seizures per Patient	6.6 ± 7.9
Minimum/Maximum Seizures per Patient	1/33
Average EEG per Patient	2.1 ± 0.9 hours
Average Seizure Duration	60 seconds
Minimum/Maximum Seizure Duration	13/212 seconds
Indicated Conditions	Focal cortical dysplasia, gliosis, encephalitis, encephalocele
Lesional/Non-lesional/Unspecified	14/0/1
Temporal/Extra-Temporal	3/12
Posterior/Anterior	10/5
Right/Left	6/9

low pass filters of 0.1 and 40 Hz. Further notch filtering was done after acquisition. The data was recorded at 256 Hz using the 10-20 common reference and was resampled to 200 Hz to be consistent with the JHH dataset. All EEG data was collected during routine clinical care and was anonymized prior to analysis under an approved IRB protocol.

2.5.1.3 CHB Dataset

We also tested our algorithm on a publicly available dataset recorded at Children's Hospital in Boston (CHB) [45],[108]. Table 2-V summarizes statistics regarding the dataset. While this dataset contains many more hours of baseline EEG than the JHH dataset, the average number of seizures per patient is comparable. Furthermore, this dataset contains no clinical notes regarding seizure type or potential localization, rendering it unsuitable for localization experiments. As such, we employ this dataset only in seizure detection applications presented in this thesis. The dataset contains scalp EEG recordings from pediatric patients and one adult ranging from age 1.5 to age 22. In total, 185 seizures from 24 cases were used. The

Table 2-IV. Demographic information and localization notes from the UWM dataset. Where available, clinician provided notes regarding underlying etiologies are provided as well.

Patient	Sex	Age	Localization Notes	Other notes
A	M	14	Left anterior and medial parietal region, post central gyrus	FCD
B	M	8	Left posterior frontal and pre central gyrus, focal at C3	FCD
C	M	17	Right temporal lobe, focus at T4	FCD
D	M	15	Left anterior parietal region, max at C3	FCD
E	M	17	Left anterior temporal pole, focus at T1	Encephalocele
F	F	10	Left anterior frontal pole, focus at Fp1	Left frontal FCD
G	F	13	Left frontal lobe	Gliosis, left frontal encephalomalacia due to prenatal stroke
H	M	11	Right temporal lobe	Gliosis, right frontotemporal due to right MTS. Seizure-free after right anterior temporal resection
J	F	17	Diffuse left onset, best developed in P3	Encephalitis
K	M	14	Left inferior frontal operculum and insula	Unknown
L	M	15	Right middle frontal region	Diffuse onset. Right frontal FCD, resected with good outcome.
M	F	11	Right middle frontal region, EEG max Fp2/F4	Multifocal FCDs, tuberous sclerosis, start from R frontal FCD
O	F	14	Hard to localize, Fz-Cz-Fp2-F4-C4 +/- P4?	FCD
Q	M	9	Right frontal maximum.	FCD
S	M	8	Focus at C3	FCD

Table 2-V. Patient demographics and clinical attributes for the CHB-MIT dataset (N=24).

	CHB Dataset
Seizure Type	Unspecified
Has localizations?	No
Number of patients	23
Average Age	9.9 ± 5.6 years
Minimum/Maximum Age	1.5/22 years
Number of Males/Females	5/17
Seizures per Patient	7.7
Minimum/Maximum Seizures per Patient	3/27
Average EEG per Patient	40.8 ± 28.4 hours
Average Seizure Duration	60 seconds
Minimum/Maximum Seizure Duration	6/752 seconds
Indicated Conditions	Unspecified
Lesional/Non-Lesional/Unspecified	Unspecified
Temporal/Extra-Temporal	Unspecified
Posterior/Anterior	Unspecified
Right/Left	Unspecified

data has been released in the longitudinal bipolar montage, requiring some alterations to methods presented in the coming chapters. EEG was recorded at 256 Hz and resampled to 200 Hz for consitency with the JHH dataset.

2.5.1.4 Pre-Processing

While specific preprocessing settings may vary between experiments, a set of standard filtering and normalizing operations are performed on the raw EEG signal. First, low pass and high pass filtering is performed on each channel individually. High pass filtering is typically performed at frequencies below 1.6 Hz to remove DC trend lines. Low pass filtering between 30 and 60 Hz is applied to remove contributions from high frequency artifacts. Notch filtering may also be applied at 60 Hz to remove any artifact due to contamination from power line signals. Following filtering, EEG in the experiments presented here are often normalized to have mean 0 and standard deviation 1. Amplitudes above a certain number of standard deviations may be clipped as well, as these regions in general consist of high amplitude artifact. Presence of extreme amplitude artifacts potentially complicates training, thus we

eliminate these signals so the remaining EEG signals are within a standardized range.

2.5.2 Notation

Throughout this thesis, I will adopt a consistent standardized notation. Where necessary, deviations from this notation will be noted in the effected chapters. Observed variables from EEG channels will be represented by X while underlying seizure states for each EEG channel will be represented by the variable Y. Individual channels will indicated by a subscript, e.g. X_i will represent the sequence of EEG signals or extracted features from electrode channel i. To refer to the observation made for window t, bracket notation will be employed. For example, $X_i[t]$ will refer to the observation made for channel i at time t and $X[t]$ will denote observations in all channels at time t. Similarly, $Y_i[t]$ and $Y[t]$ will denote hidden seizure states for channel i and for all channels at time t, respectively. The variable S will be used to refer to a global seizure state occurring across all channels and will likewise be indexed using bracket notation. Undirected graphs will be indicated using caligraphic notation. Sets of neighbors of a node in a graph will be denoted with the function ne, e.g. $ne_{\mathcal{S}}(i)$ represents the set of nodes connected to node i in graph $\mathcal{S}$. Groups of EEG channels may also be indicated using sets as subscripts. Specifically, $Y_{ne_{\mathcal{S}}(i)}$ will indicate channels whose indices are neighbors of node i in graph $\mathcal{S}$. In addition, the subscript $-i$ may be used to refer to all channels excepting i, e.g. Y_{-i}.

Chapter 3

Coupled Hidden Markov Model for Seizure Detection

3.1 Introduction

3.1.1 Chapter Contributions

In this chapter, we present a high-dimensional multichannel Coupled Hidden Markov Model (CHMM) for seizure detection from scalp EEG. This CHMM fuses information from the individual EEG channels via a spatio-temporal model of seizure spreading. Evaluating the CHMM on the JHH and CHB clinical datasets demonstrates its superior performance over standard pipelined machine learning approaches for seizure detection. Seizure tracking and localization efficacy using a CHMM approach is demonstrated for example seizures. A preliminary version of this work was introduced in [6]. The version presented here draws from [9] and provides a more complete description of inference and learning with further algorithmic refinements which improve performance. Additional real-world and simulated experiments are provided as well.

The latent chains in our CHMM correspond to the standard electrode placement locations and represent the key areas of interest on the scalp. Interactions between EEG channels are coupled such that as channels enter the seizure state, their neighboring and contralateral electrodes are more likely to also enter a seizure state. Due to the high dimensional state space, exact inference in this model is intractable. Thus a structured mean field variational

inference algorithm is developed to compute the latent posterior distributions of seizure activity and infer the seizure spreading characteristics.

We evaluate the performance of our model on simulated data and on two clinical EEG datasets from inpatient monitoring. The first testbed consists of a 15 focal epilepsy patients from an early version of the Johns Hopkins Hospital dataset. The second is drawn from the publicly available Children's Hospital of Boston (CHB) dataset [45], which contains pediatric patients with both focal and general epilepsy. Spectral power and line length are utilized as observed features in our CHMM. These features are simple yet robust variants of popular features in the seizure detection literature, which we have observed to perform well.

The generative framework is discussed in Section 3.2. Section 3.3 contains an overview of the variational inference and Expectation-Maximization (EM) procedure for fitting the model to data. Sections 3.4 and 3.5 present our results using simulated and real-world EEG data, respectively. In Section 3.6 the implications of our results are discussed and future directions for our model are suggested. Section 3.7 reviews the findings of our experimentation and concludes the chapter.

3.2 Generative Model of Seizure Propagation

This section details the generative process governing our CHMM based seizure detection algorithm. Figure 3-1 illustrates the graphical model for a generic three chain CHMM. Here, observed emissions $X_i[t]$ for chain i at time t remain conditionally independent given latent states $Y_i[t]$. The transition prior factorizes such that $P(\mathbf{Y}[t] \mid \mathbf{Y}[t-1]) = \prod_{i=1}^{N} P(Y_i[t] \mid \mathbf{Y}[t-1])$ where N denotes the number of chains. Accompanying these latent states are observed emission variables $\mathbf{X}_i$ which are conditionally independent given the latent variables.

Unlike past work, we design a transition prior $P(\mathbf{Y}[t] \mid \mathbf{Y}[t-1])$ capable of tracking the spread of seizures in focal epilepsy. In this prior, each electrode channel is represented by a chain of latent states $\mathbf{Y}_i$ where i indexes the EEG electrode. These states represent

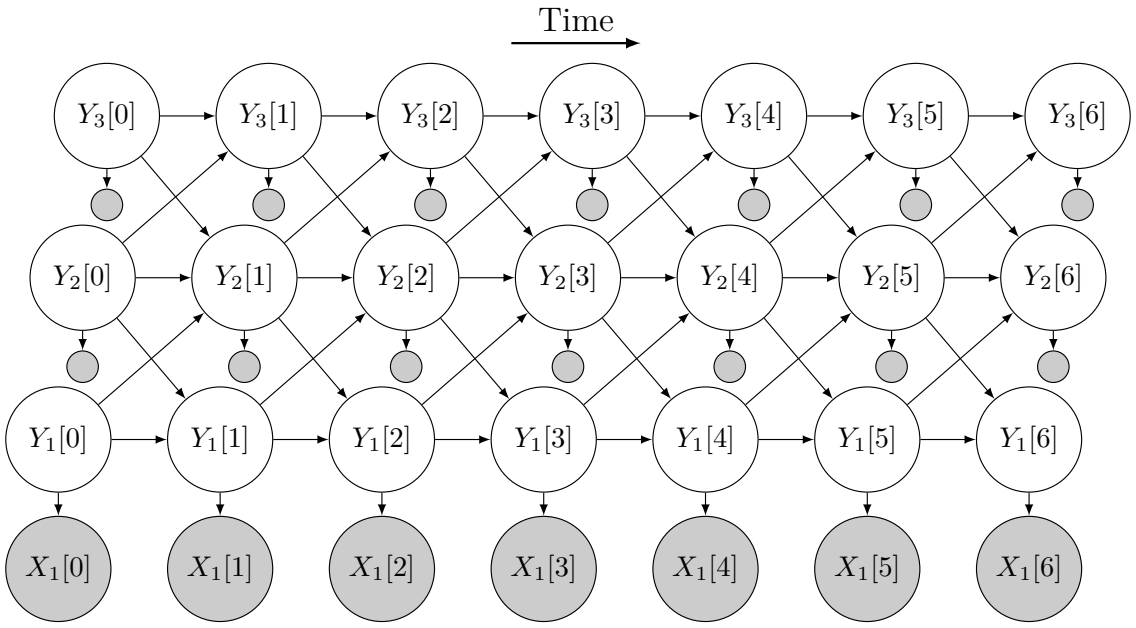


Figure 3-1. Graphical model depicting a CHMM. Observed variables, $X_i[t]$, are shown shaded. Hidden variables, $Y_i[t]$, are unshaded.

the current seizure vs. baseline label of the signal at each electrode. The observations $\mathbf{X}_i$ represent the features calculated from channel i of the EEG data.

Mathematically, our transition prior includes a contribution from all other chains via the above factorization. The observed emissions $X_i[t]$ for chain i at time t remain conditionally independent given the latent states $Y_i[t]$. In the rest of this paper, $Y_i[t]$ and $X_i[t]$ will denote latent and observed random variables, respectively, for a single chain i at timestep t. Variable ensembles are bolded, with the corresponding subscript or superscript dropped. Thus $\mathbf{Y}_i = \{Y_i[0], Y_i[1], \ldots, Y_i[T]\}$ refers to the ensemble of latent states from chain i, $\mathbf{Y}[t] = \{Y_1[t], Y_2[t], \ldots, Y_N[t]\}$ refers to the latent states for all chains at time t, and $\mathbf{Y}$ refers to all latent variables. Similarly, the subscript $-i$, as in $\mathbf{Y}_{-i}$, will indicate a collection of random variables taken over all chains excluding those of chain i. For convenience, Table 3-I defines symbols for random variables and non-random parameters used in our model.

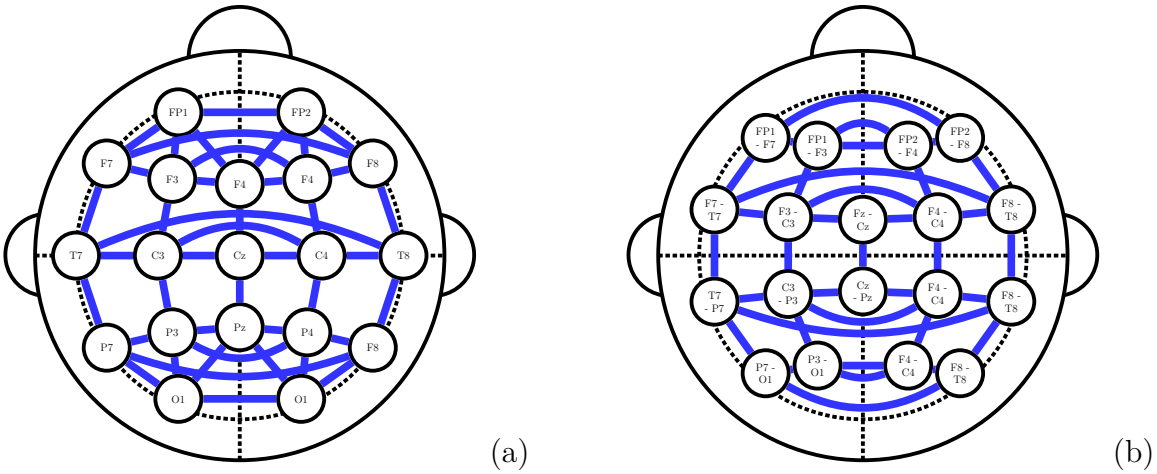
Table 3-I. Random variables (top) and non-random parameters (bottom) in our graphical model shown in Figures 3-1 and 3-2

Symbol	Description
$Y_i[t]$	Latent state in chain i at time t
$X_i[t]$	Observed variables for chain i at time t
$\eta_i[t]$	Sum of the neighbor nodes for channel i at time t
$A_i[t]$	Transition matrix for chain i at time t
$g_i[t]$	Probability of seizure onset in chain i at time t
$h_i[t]$	Probability of seizure offset in chain i at time t
$\mathcal{S}$	Connectivity graph between electrode channels
ρ_0	Seizure onset parameter
ρ_1	Seizure onset spread parameter
ϕ_0	Seizure offset parameter
ϕ_1	Seizure offset spread parameter
π_{im}^k	Emission mixture weight for mixture m for chain i for state $Y_i[t] = k$
μ_{im}^k	Emission mean for mixture m for chain i for state $Y_i[t] = k$
Σ_{im}^k	Emission covariance for mixture m for chain i for state $Y_i[t] = k$

3.2.1 Transition Prior

At the heart of our model is a graph which encodes the clinically informed spreading of a focal seizure. Connections between the latent chains of the CHMM, as illustrated in Figure 3-1, are constructed according to this propagation graph. This graph is defined in the sensor space, using the common average and bipolar montage from the 10/20 international system [15]. The bipolar montage is popular with neurologists for tracking phase changes in the raw EEG signal. For both the common reference signals and the longitudinal bipolar montage, we define a network $\mathcal{S}$ of seizure propagation by connecting neighboring and contralateral EEG channels. These graphs are shown in Figure 3-2. Neighboring connections capture local seizure spreading between adjacent EEG channels. Contralateral connections account for seizure activity that appears to manifest simultaneously on each hemisphere. An example of the seizure spreading our model encodes is shown in Figure 3-3.

Let $ne_{\mathcal{S}}(i)$ be the set of neighbors of node i as the neighbors in the graph $\mathcal{S}$. Our transition



(a) (b)

Figure 3-2. Electrode placement in the 10/20 international system [15] with seizure propagation pathways shown in blue. The edges in the graph indicate conditional independences in between nodes in consecutive timesteps of our model. (a) Graph defined on the common average montage. (b) Graph defined for the longitudinal montage.

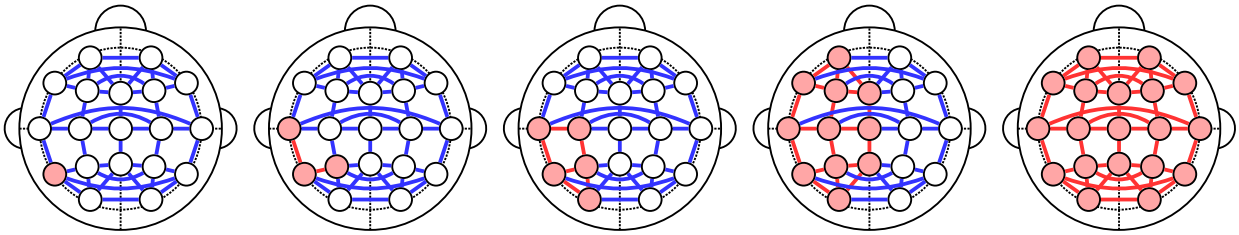


Figure 3-3. Hypothetical spreading of a focal seizure. (a) A seizure originates in a single channel. (b) The seizure propagates to neighboring EEG channels. (c) Further spreading progresses to involve more EEG channels. (d) The left hemisphere is involved. (e) The seizure becomes generalized to the entire scalp.

prior now simplifies to

$$P(Y_i[t] \mid \mathbf{Y}[t-1]) = P\left(Y_i[t] \mid Y[t-1]_i, \mathbf{Y}_{ne_{\mathcal{S}}(i)}[t-1]\right). \tag{3.1}$$

As seen in (3.1), transitions in channel i depend only on the previous state of the chain and the previous states of chains $ne_{\mathcal{S}}(i)$.

A three state left-to-right time imhomogenous transition matrix is used to encode the probability of transitions between latent states. States 0 and 2 represent pre- and post-seizure

baseline while state 1 represents a seizure as shown below.

$$A_i[t] = \begin{bmatrix} 1 - g_i[t] & g_i[t] & 0 \\ 0 & 1 - h_i[t] & h_i[t] \\ 0 & 0 & 1 \end{bmatrix} \tag{3.2}$$

$$\begin{aligned} \log\left(\frac{g_i[t]}{1 - g_i[t]}\right) &= \rho_0 + \rho_1 \eta_i[t] \\ \log\left(\frac{h_i[t]}{1 - h_i[t]}\right) &= \phi_0 + \phi_1 \eta_i[t] \end{aligned} \tag{3.3}$$

State 2 is the final state and, once entered, the chain remains there for the duration of the recording. The transition probabilities in (3.2) are computed via the logistic functions in (3.3) based on the the neighboring state assignments. Here, let $\eta_i[t]$ be defined as the number of neighbor nodes in the seizure state in the previous timestep, i.e. $\eta_i[t] \coloneqq \sum_{j \in ne_{\mathcal{S}}(i)} \mathbf{1}(Y_j[t-1] = 1)$. The parameters ρ_0 and ρ_1 represent the base parameter and neighbor influence, respectively, for transitions from pre-seizure to seizure. Namely, at any timestep, there is a small base probability that a channel that has not transitioned into a seizure state may enter one. We expect to learn a positive value for ρ_1 indicating more neighbors in a seizure state will encourage a transition into seizure. Similarly, ϕ_0 and ϕ_1 represent corresponding parameters for the transition out of the seizure state into the post-seizure baseline.

3.2.2 Emission Likelihood

Emission likelihoods $P(X_i[t] \mid Y_i[t])$ are modeled using Gaussian Mixture Models (GMMs). Let M be the number of mixtures. The parameter π_{ij}^k is the weight of mixture m in chain i when $Y_i[t] = k$. Likewise, μ_{im}^k and Σ_{im}^k are the mean and covariance, respectively, for mixture m and chain i when $Y_i[t] = k$, i.e.,

$$P(X_i[t] \mid Y_i[t] = k) = \sum_{m=1}^{M} \pi_{im}^k \mathcal{N}\left(X_i[t]; \mu_{im}^k, \Sigma_{im}^k\right). \tag{3.4}$$

As seen, the likelihood of the emission variable $X_i[t]$ is the weighted sum of Gaussian densities with weights π_{im}^k. For simplicity, we tie the parameters for the pre- and post-seizure baseline states, i.e. $\pi_{im}^0 = \pi_{im}^2$, $\mu_{im}^0 = \mu_{im}^2$, and $\Sigma_{im}^0 = \Sigma_{im}^2$.

3.3 Inference and Learning

The joint distribution of our CHMM can be written as

$$P(\mathbf{Y}, \mathbf{X}) = \prod_{i=1}^{N} P(X_i[0] \mid Y_i[0]) P\left(Y_i^0\right) \prod_{t=1}^{T} P(X_i[t] \mid Y_i[t]) P\left(Y_i[t] \mid Y_i[t-1], \mathbf{Y}_{ne_{\mathcal{S}}(i)}[t-1]\right). \tag{3.5}$$

Notice that our transition prior allows all possible latent configurations, which amounts to 3^{19} states under the 19 channel 10/20 system. Due to the high dimensionality of this latent space, exact inference is intractable. However, the structure of our model lends itself well to approximation by variational inference. We develop a structured mean field algorithm that approximates the latent posterior probability $P(\mathbf{Y} \mid \mathbf{X})$ using separate independent HMM chains for each channel.

A variational EM algorithm [109] is used to fit our model to the observed data. This algorithm alternates between an Expectation (E) step that computes current posterior beliefs of the latent seizure states given fixed values of the likelihood and transition parameters. The Maximization (M) step updates the model parameters according to these beliefs. The E- and M-steps are iterated until convergence, to obtain both the model parameters and the marginal posterior beliefs. The following subsections outline the E-step, M-step, initialization, and training of the model.

3.3.1 E-step: Variational Inference

Structured mean field variational inference is performed by defining an analytically tractable family of approximating distributions $\mathcal{Q}$ and minimizing an upper bound on the data negative log-likelihood, known as the variational free energy:

$$\mathcal{FE} \coloneqq -E_Q\left[\log P(\mathbf{Y}, \mathbf{X})\right] + E_Q\left[\log Q(\mathbf{Y})\right] \geq -\log P(\mathbf{X}). \tag{3.6}$$

The bound in (3.6) is derived via Jensen's inequality. Notice that the distribution Q that minimizes the free energy also minimizes the KL divergence between the approximating

distribution and the true posterior distribution, i.e., $D(Q(\mathbf{Y}) \parallel P(\mathbf{Y} \mid \mathbf{X}))$. Said another way, this variational inference process finds the closest distribution $Q \in \mathcal{Q}$ to the posterior $P(\mathbf{Y} \mid \mathbf{X})$ in an information theoretic sense.

Let the family of approximating distributions $\mathcal{Q}$ for the CHMM be the product of N independent HMM chains across the dynamic latent states $\mathbf{Y}_i$ for each of the N EEG channels:

$$\begin{aligned} Q(\mathbf{Y}) &= \prod_{i=1}^{N} \frac{1}{Z_{Q_i}} Q_i(\mathbf{Y}_i) \\ &= \prod_{i=1}^{N} \frac{1}{Z_{Q_i}} \prod_{t=1}^{T} \psi_i^t(Y_i[t], Y_i[t-1]) \omega_i[t](Y_i[t]). \end{aligned} \tag{3.7}$$

The distribution of each chain in (3.7) is defined by singleton factors $\omega_i[t](Y_i[t])$ and pairwise factors $\psi_i^t(Y_i[t], Y_i[t-1])$. Substituting this approximating distribution into (3.6) allows us to decompose $\mathcal{FE}$ into terms dependent on chain i and those dependent on all the other chains, denoted $-i$.

$$\begin{aligned} \mathcal{FE} =& - E_{Q_i}\left[E_{Q_{-i}}\left[\log p(\mathbf{Y}_i, \mathbf{X}_i \mid \mathbf{Y}_{-i}, \mathbf{X}_{-i})\right]\right] + E_{Q_i}\left[\log Q_i(\mathbf{Y}_i)\right] \\ & - E_{Q_{-i}}\left[\log p(\mathbf{Y}_{-i}, \mathbf{X}_{-i})\right] + E_{Q_{-i}}\left[\log Q_{-i}(\mathbf{Y}_{-i})\right] \\ =& - E_{Q_i}\left[E_{Q_{-i}}\left[\log p(\mathbf{Y}_i, \mathbf{X}_i \mid \mathbf{Y}_{ne_{\mathcal{S}}(i)})\right]\right] \\ & + E_{Q_i}\left[\log Q_i(\mathbf{Y}_i)\right] + \text{constant} \end{aligned} \tag{3.8}$$

In this substitution, we have used the factorization $p(\mathbf{Y}, \mathbf{X}) = p(\mathbf{Y}_{-i}, \mathbf{X}_{-i})\, p(\mathbf{Y}_i, \mathbf{X}_i \mid \mathbf{Y}_{ne_{\mathcal{S}}(i)})$ to isolate terms pertaining to chain i as in the last two lines of (3.8). Notice that this factorization leads to a natural coordinate descent algorithm. Namely, by holding chains $-i$ constant and minimizing the free energy with respect to chain i, the upper bound on the negative log-likelihood can be iteratively refined. Since $\mathcal{FE}$ is bounded from below (i.e., it cannot diverge to $-\infty$), this coordinate descent procedure is guaranteed to converge to a

local optima of the free-energy objective.

$$\begin{aligned}\arg\min_{Q_i} \mathcal{FE} &= \arg\min_{Q_i} -E_{Q_i}\left[E_{Q_{-i}}\left[\log p(\mathbf{Y}_i, \mathbf{X}_i \mid \mathbf{Y}_{ne_{\mathcal{S}}(i)})\right]\right] + E_{Q_i}\left[\log Q_i(\mathbf{Y}_i)\right] \\ &\quad + \text{constant} \\ &= \arg\min_{Q_i} -E_{Q_i}\left[E_{Q_{-i}}\left[\log p(\mathbf{Y}_i, \mathbf{X}_i \mid \mathbf{Y}_{ne_{\mathcal{S}}(i)})\right]\right] + E_{Q_i}\left[\log Q_i(\mathbf{Y}_i)\right] \\ &= \arg\min_{Q_i} D(Q(\mathbf{Y}) \parallel p(\mathbf{Y}_i \mid \mathbf{Y}_{ne_{\mathcal{S}}(i)}, \mathbf{X}))\end{aligned} \tag{3.9}$$

Hence, we perform inference, i.e. optimize $Q_i(\mathbf{Y}_i)$, over the individual chains in a coordinate descent procedure until $\mathcal{FE}$ converges. From the last line in (3.9), note that at optimality, the approximating distribution Q_i is related to the expected value of the neighbor chains as follows:

$$Q_i \propto \exp\left\{E_{Q_{ne_{\mathcal{S}}(i)}}\left[\log p(\mathbf{Y}_i, \mathbf{X}_i \mid \mathbf{Y}_{ne_{\mathcal{S}}(i)})\right]\right\}. \tag{3.10}$$

Effectively, the approximating distribution Q_i incorporates information from neighboring chains via the $p(\mathbf{Y}_i, \mathbf{X}_i \mid \mathbf{Y}_{ne_{\mathcal{S}}(i)})$ terms in (3.10). Notice that the exponent of (3.10) factors into $p(\mathbf{Y}_i, \mathbf{X}_i \mid \mathbf{Y}_{ne_{\mathcal{S}}(i)}) = p(\mathbf{Y}_i \mid \mathbf{Y}_{ne_{\mathcal{S}}(i)})\, p(\mathbf{X}_i \mid \mathbf{Y}_i)$. These two factors can be matched to the pairwise and singleton terms, respectively, of the approximating distribution $Q_i(\mathbf{Y}_i)$ in (3.7). In addition, we approximate the contribution of future Y_j^{t+1} for chains $j \in ne_{\mathcal{S}}(i)$ via a linearized approximation that we incorporate into the singleton terms of (3.7). This approximation is further described in Section 3.3.1.2.

The expectations in (3.8) are computed by iteratively applying the forward-backward algorithm to a single chain, while fixing the approximate posterior probabilities of the remaining chains. Let the singleton and pairwise marginals in the approximating distribution for chain i at time t be defined as $\tilde{\gamma}_i[t](j) \coloneqq E_{Q_i}\left[\mathbf{1}(Y_i[t] = j)\right]$ and $\tilde{\xi}_i[t](j,k) \coloneqq E_{Q_i}\left[\mathbf{1}(Y_i[t] = j, Y_i[t+1] = k)\right]$ where $\mathbf{1}(\cdot)$ is the indicator function.

3.3.1.1 Pairwise factors

The pairwise factors of the approximating distribution $\psi_i[t](Y_i[t], Y_i[t-1])$ mimic the transition parameters of the original distribution, i.e. $\mathcal{FE}$ is minimized via a left-to-right time

imhomogenous structure

$$\tilde{A}_i[t] = \begin{bmatrix} 1-\tilde{g}_i[t] & \tilde{g}_i[t] & 0 \\ 0 & 1-\tilde{h}_i[t] & \tilde{h}_i[t] \\ 0 & 0 & 1 \end{bmatrix}$$

where $\psi_i[t](Y_i^t = k, Y_i[t-1] = j) = (\tilde{A}_i[t])_{jk}$, where $(\cdot)_{jk}$ corresponds to the entry in row j and column k of the matrix argument. From (3.10) the pairwise term becomes

$$\psi_i[t](Y_i[t], Y_i[t-1]) \propto \exp\left\{E_{Q_{ne_{\mathcal{S}}(i)}}\left[p(Y_i[t] \mid Y_i[t-1], \mathbf{Y}_{ne_{\mathcal{S}}(i)}[t-1])\right]\right\} \tag{3.11}$$

Substituting in the parameters of our distribution, this relationship implies that the onset probability satisfies

$$\tilde{g}_i^t \propto \exp\{E_{Q_{ne_{\mathcal{S}}(i)}} \log g_i[t]\} \tag{3.12}$$

with

$$(1-\tilde{g}_i[t]) \propto \exp\{E_{Q_{ne_{\mathcal{S}}(i)}} \log(1-g_i[t])\}. \tag{3.13}$$

A similar relationship is true for the variational offset parameter $\tilde{h}_i[t]$. Dividing these terms and taking the logarithm, the variational transition terms are given by the expected value of the logits in the original transition prior.

$$\log\left(\frac{\tilde{g}_i[t]}{1-\tilde{g}_i[t]}\right) = \rho_0 + \rho_1 E_{Q_{ne_{\mathcal{S}}(i)}}[\eta_i[t]]$$

$$\log\left(\frac{\tilde{h}_i[t]}{1-\tilde{h}_i[t]}\right) = \phi_0 + \phi_1 E_{Q_{ne_{\mathcal{S}}(i)}}[\eta_i[t]]$$

$$E_{Q_{ne_{\mathcal{S}}(i)}}[\eta_i[t]] = \sum_{j \in ne_{\mathcal{S}}(i)} \tilde{\gamma}_j[t-1]$$

These equations bare a strong resemblance to the original transition terms presented in (3.2) and (3.3) and incorporate cross-channel information via the $E_{Q_{ne_{\mathcal{S}}(i)}}[\eta_i[t]]$ terms.

3.3.1.2 Singleton Factors

The singleton factors in our approximating distribution $\omega_i[t](Y_i[t])$ mimic the emission likelihood. However, these terms also absorb information from the neighbor chains in the subsequent timestep. This information is captured by the multiplicative factor $\exp\{\alpha_i[t](z)\}$.

We use a linearized approximation of this term, shown in (3.16), to easily fold it into the singleton factors. Namely the expectation in these equations is easily computed as the sum of the neighbor's marginals in the previous timestep $\nu_i[t+1] = \sum_{j \in ne_{\mathcal{S}}(i)} \tilde{\gamma}_j[t]$:

$$\omega_i[t](Y_i[t] = 0, 2) = P(X_i[t] \mid Y_i[t] = 0, 2) \exp\{\alpha_i[t](0)\} \tag{3.14}$$

$$\omega_i[t](Y_i[t] = 1) = P(X_i[t] \mid Y_i[t] = 1) \exp\{\alpha_i[t](1)\} \tag{3.15}$$

$$\begin{aligned} \alpha_i[t](z) \approx \sum_{j \in ne_{\mathcal{S}}(i)} \Big[& \tilde{\xi}_j[t](0,0) \left(-\rho_0 - \rho_1 \left(\nu_j[t+1] + z\right)\right) \\ & - \tilde{\gamma}_j[t](0) \log\left(1 + e^{-\rho_0 - \rho_1(\nu_j[t+1]+z)}\right) \\ & + \tilde{\xi}_j[t](1,1) \left(-\phi_0 - \phi_1 \left(\nu_j[t+1] + z\right)\right) \\ & - \tilde{\gamma}_j[t](1) \log\left(1 + e^{-\phi_0 - \phi_1(\nu_j[t+1]+z)}\right) \Big]. \end{aligned} \tag{3.16}$$

3.3.2 M-Step: Update Model Parameters

In the M-step, the parameters of the generating distributions are updated using the marginals computed in the E-step.

3.3.2.1 Emission Parameters

Updating each GMM emission likelihood requires a nested EM update. The inner E-step in (3.17) computes the expected latent state and mixture combination for every emission.

$$\tilde{\tau}_{i,j}[t](k) \coloneqq E_{Q_i}\left[Y_i[t] = k, X_i[t] \text{ drawn from mixture } j\right] \tag{3.17}$$

The inner M-step in (3.23) updates the parameters of the GMM via standard mean and variance updates [57].

$$\mu_{ij}^{0,2} = \frac{\sum_{k\in\{0,2\}} \sum_{t=0}^{T} \tilde{\tau}_{i,j}[t](k) X_i[t]}{\sum_{k\in\{0,2\}} \sum_{t=0}^{T} \tau_{i,j}[t](k)} \tag{3.18}$$

$$\mu_{ij}^{1} = \frac{\sum_{t=0}^{T} \tilde{\tau}_{i,j}[t](k) X_i[t]}{\sum_{t=0}^{T} \tilde{\tau}_{i,j}[t](k)} \tag{3.19}$$

$$\Sigma_{ij}^{0,2} = \frac{\sum_{k\in\{0,2\}} \sum_{t=0}^{T} \tilde{\tau}_{i,j}[t](k) \left(X_i[t] - \mu_{ij}^{k}\right)^2}{\sum_{k\in\{0,2\}} \sum_{t=0}^{T} \tilde{\tau}_{i,j}[t](k)} \tag{3.20}$$

$$\Sigma_{ij}^{1} = \frac{\sum_{t=0}^{T} \tilde{\tau}_{i,j}[t](k) \left(X_i[t] - \mu_{ij}^{k}\right)^2}{\sum_{t=0}^{T} \tilde{\tau}_{i,j}[t](k)} \tag{3.21}$$

$$\pi_{ij}^{0} = \pi_{ij}^{2} = \frac{\sum_{t=0}^{T} \tilde{\tau}_{ij}[t](0) + \tilde{\tau}_{ij}[t](2)}{\sum_{j'} \sum_{t=0}^{T} \tilde{\tau}_{ij'}[t](0) + \tilde{\tau}_{ij'}[t](2)} \tag{3.22}$$

$$\pi_{ij}^{1} = \frac{\sum_{t=0}^{T} \tilde{\tau}_{ij}[t](1)}{\sum_{j'} \sum_{t=0}^{T} \tilde{\tau}_{ij'}[t](1)} \tag{3.23}$$

The nested E-step and nested M-step are repeated until the algorithm converges. This iterative procedure is initialized using the previous settings for the emission likelihoods.

3.3.2.2 Transition Parameters

The transition parameters form a logistic regression onto the expected transition posteriors $\tilde{\xi}_i[t](j,k)$. Here we provide the update equations for the onset parameters ρ_0 and ρ_1. Equations for offset parameters ϕ_0 and ϕ_1 are almost identical and are omitted for space. Newton's method is used to minimize the $\mathcal{FE}$. Let $\nabla_\rho \mathcal{FE}_k$ and $\nabla_\rho^2 \mathcal{FE}_k$ be the gradient and Hessian of the free energy with respect to the vector of onset parameters $\rho = (\rho_0, \rho_1)$. A single iteration of the Newton's method algorithm is

$$p_k = -\left(\nabla_\rho^2 \mathcal{FE}_k\right)^{-1} \nabla_\rho \mathcal{FE}_k \qquad \rho_{k+1} = \rho_k + \alpha_k p_k$$

where the subscript k indicates the iteration number and α_k is the step size. Newton's method is prone to oscillation in logistic regression in some cases. Therefore we employ backtracking to ensure our updates remain within a stable region around the minimum. Specifically, we require our step size to fulfill the second strong Wolfe condition $|\nabla f(\rho_k + \alpha_k p_k)^T p_k| \leq |\nabla f(\rho_k)^T p_k|$

[110]. This ensures that each step approaches a stationary point. Defining the logistic sigmoid function as $\sigma(x) \coloneqq \frac{1}{1+e^{-x}}$, the first and second derivatives making up the gradient and Hessian are shown below.

$$
\begin{aligned}
\frac{\partial}{\partial \rho_0} E_Q\left[\log P(\mathbf{Y})\right] &= \sum_{t=1}^{T}\sum_{i=1}^{N}\left(\tilde{\xi}_i[t](0,1) - \tilde{\gamma}_i[t](0) E_{Q_{ne_{\mathcal{S}^{(i)}}}}\Big[\sigma(\rho_0 + \rho_1\eta_i[t])\Big]\right) - 2\lambda\rho_0 \\
\frac{\partial}{\partial \rho_1} E_Q\left[\log P(\mathbf{Y})\right] &= \sum_{t=1}^{T}\sum_{i=1}^{N}\tilde{\xi}_i[t](0,1)\Bigg(E_{Q_{ne_{\mathcal{S}^{(i)}}}}\left[\eta_i[t]\right] \\
&\qquad - \tilde{\gamma}_i[t](0) E_{Q_{ne_{\mathcal{S}^{(i)}}}}\Big[\eta_i[t]\sigma(\rho_0 + \rho_1\eta_i[t])\Big]\Bigg) - 2\lambda\rho_1 \\
\frac{\partial^2}{\partial \rho_0^2} E_Q\left[\log P(\mathbf{Y})\right] &= \sum_{t=1}^{T}\sum_{i=1}^{N} - \tilde{\gamma}_i[t](0) E_{Q_{ne_{\mathcal{S}^{(i)}}}}\Big[\sigma(\rho_0 + \rho_1\eta_i[t])\left(1 - \sigma(\rho_0 + \rho_1\eta_i[t])\right)\Big] \\
&\qquad - 2\lambda \\
\frac{\partial^2}{\partial \rho_1^2} E_Q\left[\log P(\mathbf{Y})\right] &= \sum_{t=1}^{T}\sum_{i=1}^{N} - \tilde{\gamma}_i[t](0) \\
&\qquad \cdot E_{Q_{ne_{\mathcal{S}^{(i)}}}}\Big[(\eta_i[t])^2\sigma(\rho_0 + \rho_1\eta_i[t]) \cdot (1 - \sigma(\rho_0 + \rho_1\eta_i[t]))\Big] \\
&\qquad - 2\lambda \\
\frac{\partial^2}{\partial \rho_0 \partial \rho_1} E_Q\left[\log P(\mathbf{Y})\right] &= \sum_{t=1}^{T}\sum_{i=1}^{N} - \tilde{\gamma}_i[t](0) E_{Q_{ne_{\mathcal{S}^{(i)}}}}\Big[\eta_i[t]\sigma(\rho_0 + \rho_1\eta_i[t]) \\
&\qquad \cdot (1 - \sigma(\rho_0 + \rho_1\eta_i[t]))\Big]
\end{aligned}
\tag{3.24}
$$

Here ℓ_2 norm regularization with weight λ is used to stabilize the learning and deal with identifiability issues.

3.3.3 CHMM Initialization and Semi-Supervised Training

Our model is trained on multichannel EEG snippets in which an expert has annotated the approximate start and end of a single seizure. We emphasize that we do not use localization information about where the seizure originates and how it spreads. Rather, our model automatically learns this information from the data. Pre-seizure is fixed as state $Y_i[t] = 0$ and post-seizure is fixed as $Y_i[t] = 2$ throughout the course of training. Inference is performed over the seizure interval with a required transition into the seizure state. This semi-supervised

strategy overcomes both the lack of exact onset and offset labels and the lack of spreading labels.

We initialize the emission distribution based on the seizure interval annotations. A GMM for the seizure state is trained on all data from the seizure interval while non-seizure state GMM is trained on data from the rest of the recording. Transition prior parameters are initialized to $\rho_0 = -7$, $\rho_1 = 2$, $\phi_0 = -3$, and $\phi_1 = 0$. These settings correspond to one expected seizure every 13 minutes lasting an expected length of 15 seconds. An neighbor in a seizure state raises the probability of seizure onset in a given channel by a multiple of roughly 7, with no change in offset probability due to neighbors in a seizure state. *We emphasize that these settings are just for initialization. The model updates these parameters through the variational EM algorithm.* In fact, we observe convergence to a stable set of parameter values regardless of initialization.

3.3.4 Comparison to Machine Learning Baselines

We compared the performance of our model to classifiers from the machine learning literature, performing classification using a Deep Neural Network (DNN) and a Random Forests classifier (RF) [57]. The neural network includes 2 hidden layers with a discriminative output layer. In contrast, RFs are an ensemble of simple decision trees that perform classification using majority vote over the decisions of the ensemble. By combining relatively simple classifiers, random forests can create complicated decision functions while remaining robust to overfitting.

In addition, a GMM based Likelihood Ratio Test (LRT) was used to perform posterior inference. The GMM-LRT is analogous to our original model with no prior over the hidden states. One GMM is trained for all non-seizure intervals to model $P(Y \mid X = 0)$ and another is trained on the seizure intervals to model $P(Y \mid X = 1)$. $\delta \coloneqq P(X = 1)$ is fixed to the proportion of seizure in the dataset. The posterior probability of a test frame belonging to

the seizure class is

$$P(X = 1 \mid Y) = \frac{\delta P(Y \mid X = 1)}{\delta P(Y \mid X = 1) + (1 - \delta)\, P(Y \mid X = 0)}.$$

For each classifier, DNN, RF, or GMM, two approaches to seizure detection are evaluated. In the first approach, features from each EEG channel are concatenated to form a single stacked feature vector used for detection. The second approach trains classifiers on each channel independently to evaluate the performance from a channel-wise perspective. These comparisons allow us to quantify the gain from fusing information across channels. When presenting our baseline results, the prefixes S and I are used to represent stacked feature vectors and independent channel-wise classification, respectively, e.g. SGMM for stacked feature Gaussian mixture model or IRF for independent channel random forest.

3.4 Evaluation on Synthetic Data

Synthetic data is generated by simulating different seizure propagation patterns. The CHMM and baseline algorithms are then used to infer the underlying spatio-temporal dynamics. The latent seizure states are sampled from a modified version of the transition prior outlined in Section II-A governed by onset parameters $\{\rho_0, \rho_1\}$, and offset parameters $\{\phi_0, \phi_1\}$. Each seizure recording begins in a non-seizure state. Prior to a seizure occurring, the probability of a channel entering the seizure state depends on its neighbors via $\sigma(\rho_0)$. After onset, the probability a channel enters the seizure state is given by $\sigma(\rho_0 + \rho_1 \eta_i[t])$. Thus, ρ_1 controls the speed of seizure spreading, where higher values cause faster spreading. Departing from our prior, we enforce that all channels must enter the seizure state before offset is allowed. Once all channels enter the seizure state, the probability a channel returns to the normal state is given by $\sigma(\phi_0 + \phi_1 \eta_i[t])$.

Paremeters are fixed $\rho_0 = -9.0$, $\phi_0 = -3.0$, and $\phi_1 = 0.0$ to simulate seizures occurring after roughly 425 timesteps. These parameters control the likelihood of seizure onset, the base rate of offset, and the between channel influence during offset. As these parameters are

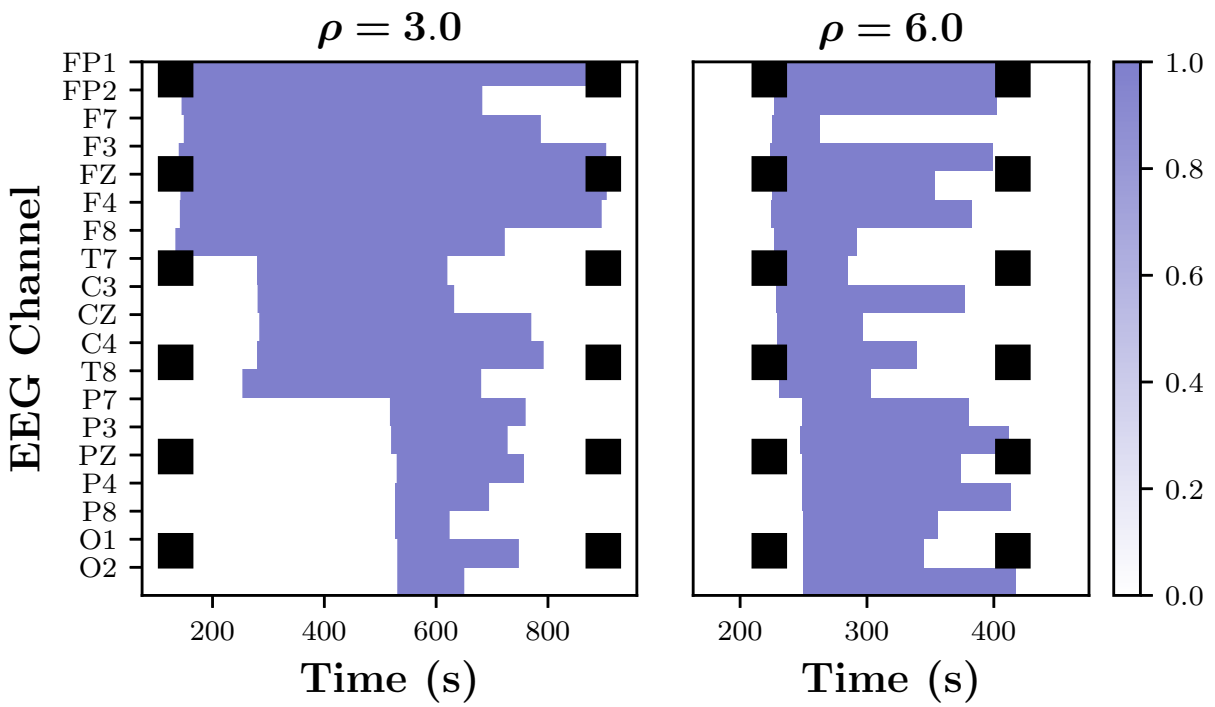


Figure 3-4. Simulated underlying seizure stats for (left) slow and (right) fast propagation. Seizure onset and offset are shown with dashed black vertical lines. Seizure is shown in blue while non-seizure is shown in white.

less clinically relevant than the speed of seizure spreading, we do not vary them between tests. Instead, we vary the neighborhood influence ρ_1, which indirectly controls the spreading rate. In our experiments, $\rho_1 \in \{3.0, 4.0, 5.0, 6.0\}$ to explore a range of seizure spreading speeds, where higher values of ρ_1 correspond to quicker spreading. While recording length in the real-world datasets varies, this variation is not clinically meaningful. Thus the length of the simulations is fixed to 1600 samples, corresponding to 20 minute recordings. Figure 3-4 shows two simulated seizures for $\rho_1 = 3.0$ (slow) and $\rho_1 = 6.0$ (fast).

Emissions are sampled from a univariate normal distributions with mean 0 for non-seizure and mean 1 for seizure. The intra-class variance parameter is swept in $[0.1, 1]$ to evaluate the model performance under different degrees of separability between the seizure and non-seizure classes. If the variance is small, the two classes remain easily separable. However, as the variance increases, the data distributions have a higher degree of overlap, making

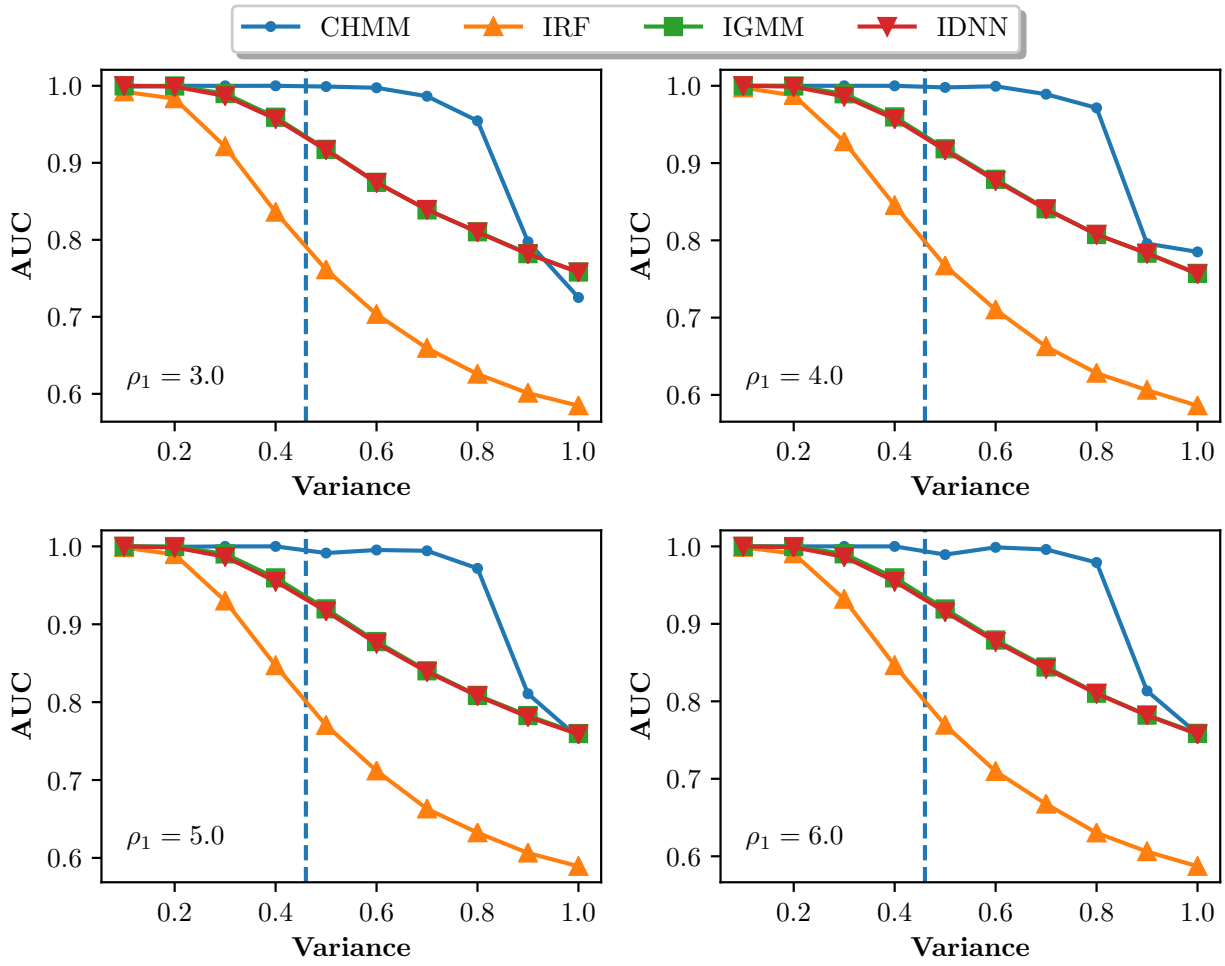


Figure 3-5. AUC results for the CHMM and channel-independent baseline methods across a range intra-class variance values. Class separability estimated from the real-world EEG data is shown by the vertical dashed blue line. Values of ρ correspond to rate of seizure spread.

classification more difficult. For each setting, 10 sets of simulated training and testing data Y are generated, each containing 100 seizures. Classifiers are trained using the training sets and test performance is reported for each classifier using the average across all folds.

Figures 3-5 and 3-6 show the results of the simulated experiment. For both the IDNN and SDNN we use a network with two hidden layers of 10 nodes. The Area Under the Curve (AUC) for each test is shown on the y-axis while the intra-class variance parameter is shown on the x-axis. Figure 3-5 shows the CHMM models and independent baselines for the full range of intra-class variances. Because the channels are evaluated individually,

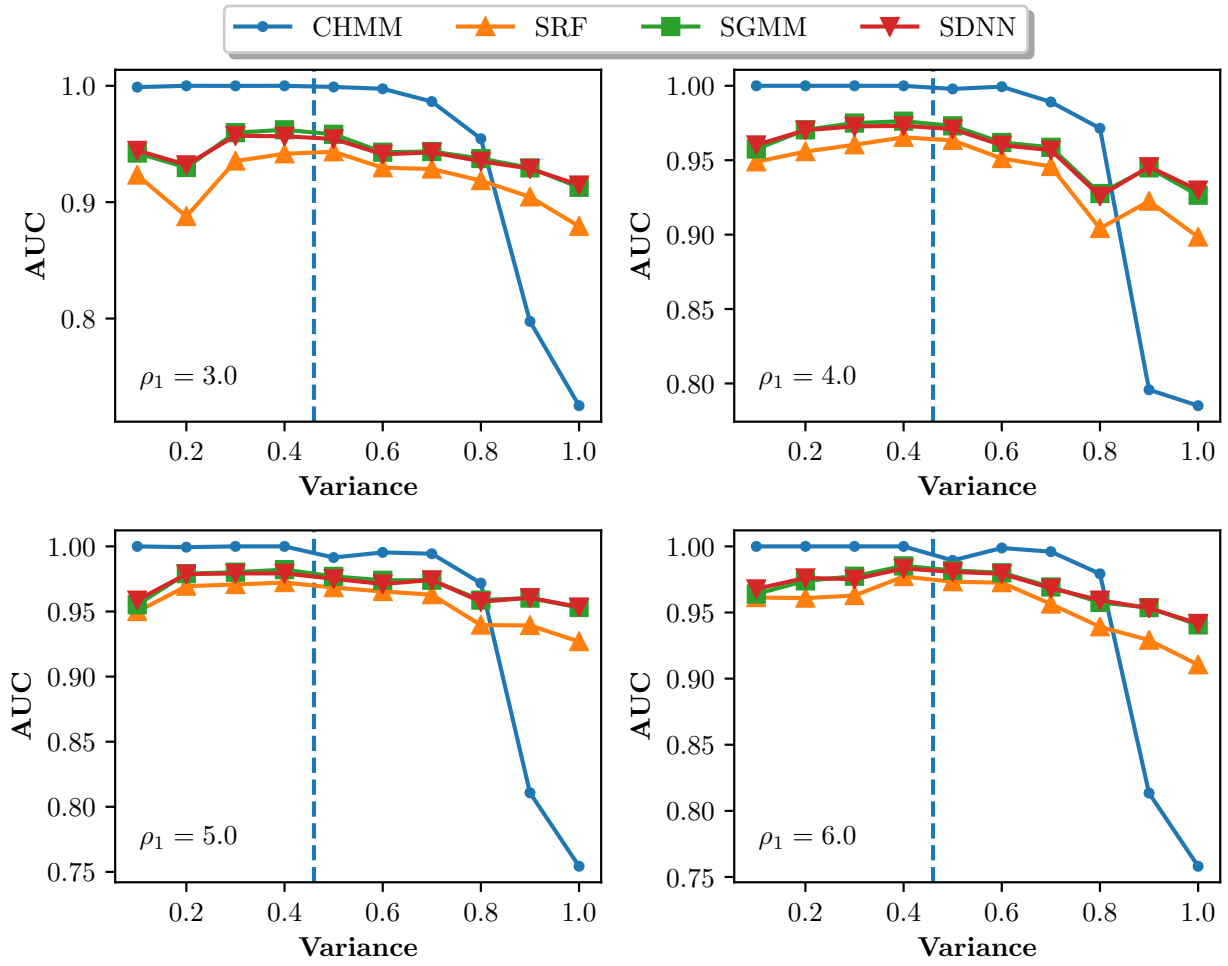


Figure 3-6. AUC results for the CHMM and stacked-channel baseline methods across a range intra-class variance values. Class separability estimated from the real-world EEG data is shown by the vertical dashed blue line. Values of ρ correspond to rate of seizure spread.

baselines in this figure achieve extremely similar performance for all values of ρ_1. At higher noise levels we observe that performance of the CHMM model degrades faster under slower spreading seizures. Intuitively, this makes sense as seizures originating close to simultaneously in each channel should result in easier cross channel information fusion, leading to increased performance. Figure 3-6 shows the performance for the CHMM and stacked baselines. In general we observe that the performance of the stacked baselines increases as ρ_1 increases. This makes intuitive sense, as faster spreading seizure activity will be present in more channels concurrently and is thus easier to classify.

Finally, we have estimated a lower bound on the class separability of our real-world EEG data by computing the Hellinger distance between the seizure and non-seizure classes in the focal JHH dataset. Specifically, a multivariate normal distribution with full covariance is fit to the temporal features extracted for each class (see Section V for details on the data preprocessing). The Hellinger distance between these two distributions for each EEG channel is then computed. These distances were then averaged across all EEG channels and recordings. Based on the average Hellinger distance, we computed the real-world data to have an approximate emission variance of 0.46. This value is marked with a dashed vertical line in Figures 6–7.

3.5 Evaluation on Clinical Data

We evaluate our model on clinical EEG data recorded in two different hospitals. Details of the datasets, preprocessing, and feature extraction are given below. In our experiments on real data, the two hidden layers of the IDNN and SDNN contain 10 and 50 units, respectively.

JHH Dataset: We validate our experiments using patients from an early iteration of the JHH dataset. In total this dataset includes 90 seizures from 15 patients. Due to the liberal annotation procedures followed at this hospital, many of our annotated seizure intervals are overly generous and contain periods of baseline before and after the seizure event. Each of these seizure recordings contains up to 10 minutes of baseline EEG before and after the seizure. Recordings were sampled at 200 Hz in 10/20 reference space using the common average montage. In early experimentation, we evaluated our methods after applying the bipolar montage but found no sizable change in performance. Though not used during training, a subset of recordings contain clinical annotations of the likely seizure onset. When possible, we validate the seizure spread information generated by our model using these annotations.

CHB Dataset: We also tested our algorithm on a publicly available dataset recorded at

Children's Hospital in Boston (CHB) [45],[108]. Seizure regions from the CHB dataset were trimmed with a random amount of pre- and post-seizure baseline not exceeding 10 minutes in each case. In total, 185 seizures from 24 patients were used. This dataset contains both focal and generalized seizures with more accurately annotated seizure intervals. The data has been released in the bipolar montage; hence the network depicted in Figure 3-2b for was used inference.

3.5.1 Preprocessing and Feature Extraction

Each EEG recording is minimally preprocessed using a high-pass filter at 1.6 Hz and a low-pass filter at 50 Hz to remove DC trends and high frequency noise. In addition, a second order notch filter at 60 Hz with $Q = 20$ was used to remove any remaining interference from the power supply.

We extracted features based on 1 second Tukey windows with shape parameter 0.25 and a 750 ms advance. First, short time Fourier transform coefficients for each window are computed. The coefficient magnitudes were summed according to the standard EEG frequency bands: theta (1–4 Hz), delta (4–8 Hz), alpha (8–13 Hz), and beta (13–30 Hz). A logarithm was applied to the summed features. These features track activity in each brain wave band, which has been noted to change during seizures and closely resembles feature extraction techniques in [55, 56]. Log line length features [111], computed as $\log L = \log\left(\sum_{t=1}^{T} |s[t] - s[t-1]|\right)$ where s is a time series, were also included. Line length captures the smoothness of a signal. The features for each channel were normalized to mean 0, standard deviation 1 for each recording. This combination of features echoes those cited as optimal in [112]. Our prior experimentation with different feature extraction methods verified that spectral power and line length outperformed more sophisticated EEG features in the literature such as wavelet and entropy measures. In addition, similar features have been employed for use with implanted EEG sensors [113].

3.5.2 Evaluation

Five-fold cross validation was used to evaluate the methods. Each recording was randomly assigned to a fold irrespective of patient. This approach stands in contrast to prior work, which trains patient specific classifiers. Rather, we evaluate our model in a general setting where we may not have prior data from any given patient. Training was performed on four folds while testing was performed on the remaining fold. Reported metrics are averaged across all test folds.

We quantitatively evaluate each model based on the amount of detected seizure activity within the annotated region. Here, each channel detection within the seizure interval is counted as True Positive (TP). Each channel detection outside the interval counted as a False Positive (FP). Let $t = 0, \ldots, T$ index the one-second time windows within a single recording and let t_s and t_e denote the starting and ending time of the annotated seizure interval. Mathematically let the TP, FP, True Negatives (TN), and False Negatives (FN) for channel i be defined as:

$$TP_i = \sum_{t=t_e}^{t_s} \mathbf{1}(Y_i[t] = 1) \qquad FP_i = \sum_{t=1}^{T} \mathbf{1}(Y_i[t] = 1) - TP_i$$
$$FN_i = \sum_{t=t_e}^{t_s} \mathbf{1}(Y_i[t] \neq 1) \qquad TN_i = \sum_{t=1}^{T} \mathbf{1}(Y_i[t] \neq 1) - FN_i \ .$$

Detections are aggregated across channels to yield True Positive Rate (TPR) and True Negative Rate (TNR):

$$TPR = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{t_e - t_s}$$
$$TNR = \frac{1}{N} \sum_{i=1}^{N} \frac{TN_i}{T - (t_e - t_s)}.$$

As we lack onset annotations for each individual channel we calculate these statistics based on the single clinician provided onset annotation. This strategy is based on the assumption that given liberal onset annotations, any positive classification within the annotation are likely to be correct.

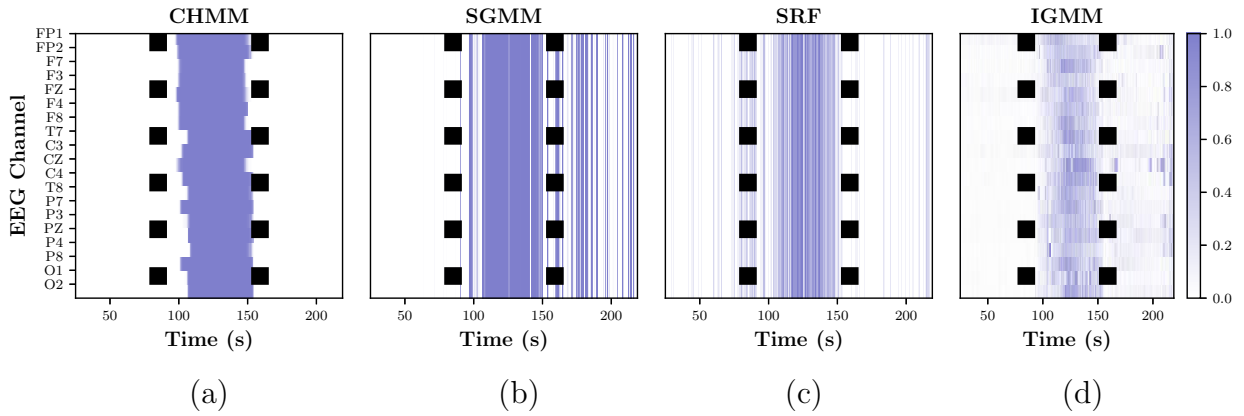


Figure 3-7. CHMM and selected baseline classification posteriors for a representative JHH patient. EEG channels are arranged on the y-axis with time along the x-axis. Seizure onset and offset are indicated by the vertical dashed lines. (a) Classification results using our CHMM model. Stacked features are used in conjunction with GMM and RF classifiers in (b) and (c), respectively. (d) Classification performed on each channel with a GMM. Posterior beliefs are shown in blue where intensity depicts the strength of the belief.

Our reported performance is averaged over (potentially generous) seizure regions. Notice that our evaluation criterion is more stringent than the metrics reported in prior work as we report *percentages of correctly classified activity* rather than a single correct detections within the seizure interval. Hence, lower overall TPR than is presented in other seizure detection papers is expected. Precision (P) and Recall (R) averaged across channels are reported along with the AUC and F1 score to evaluate overall performance of each detector:

$$P = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i}$$
$$R = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i}.$$

3.5.3 Experimental Results

Figure 3-7 depicts the output of our model for a single recording from the JHH dataset. Figure 3-7 (a) shows the posterior beliefs of our model in blue. EEG channels are arranged along the y-axis of the image while time progresses horizontally. The dashed black lines indicate the annotated onset and offset of the seizure. Once again, these annotations serve as a rough guide, rather than a precise demarcation of onset. Figure 3-7 (b–d) shows baseline

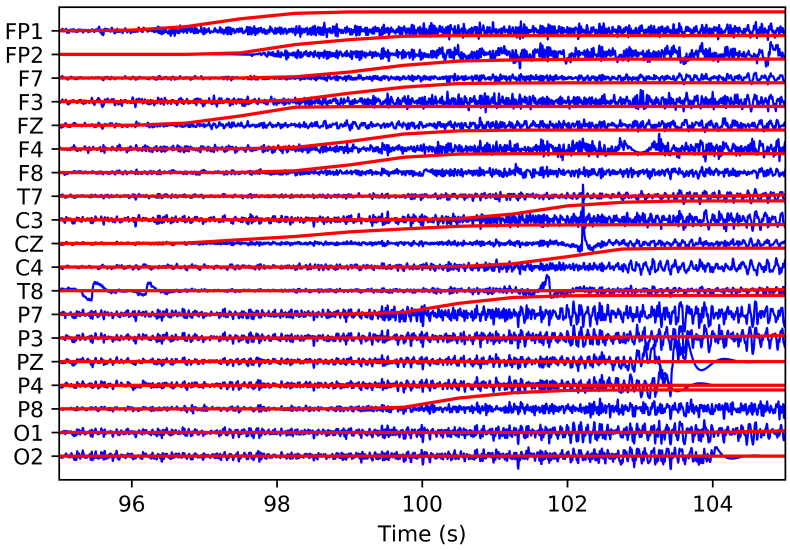


Figure 3-8. CHMM posteriors superimposed on EEG for the recording in Figure 3-7. Raw EEG signal is shown in blue while CHMM posteriors are shown in red. EEG channels are organized on the y-axis, while time progresses along the x-axis.

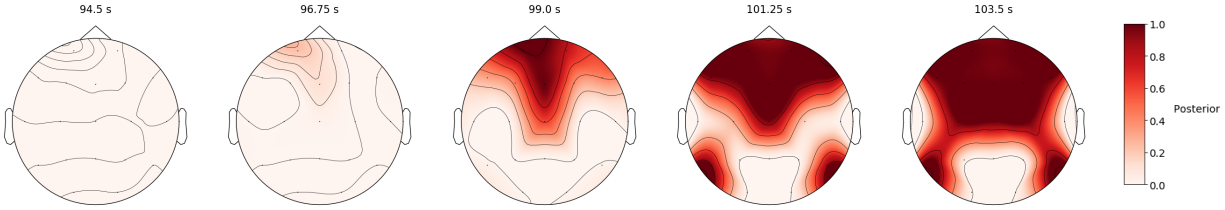


Figure 3-9. Spread of the seizure depicted in Figure 3-7, as computed by the CHMM. The CHMM classifies the earliest ictal activity occurring in the left frontal channels in agreement with clinical annotations.

classifications for the same recording shown in Figure 3-7 (a). Figure 3-8 shows the same posterior distributions superimposed on the raw EEG signal while Figure 3-9 shows them topographically on the scalp as the seizure progresses.

These results illustrate the ability of our CHMM to correctly label seizure intervals. Clinical annotations for the seizure in Figure 3-7 note rhythmic theta activity in the left frontal area at the onset of the seizure, which correlates with the earliest CHMM detection in Figure 3-7 (a). Likewise, the superimposed posteriors in Figure 3-8 and topographic detail in Figure 3-9 both show the earliest response of our model occurring in the left frontal region

Table 3-II. Quantitative Results for the JHH dataset

Classifier	TPR	TNR	AUC	P	R	F1
CHMM	37.40 ± 7%	**98.29** ± 0.54%	**0.84** ± 0.05	**0.66** ± 0.11	**0.44** ± 0.09	**0.49** ± 0.09
SDNN	**38.10** ± 4 %	93.76 ± 1.08 %	0.82 ± 0.03	0.35 ± 0.04	0.41 ± 0.05	0.35 ± 0.04
SGMM	33.42 ± 5 %	95.19 ± 1.45 %	0.72 ± 0.04	0.44 ± 0.08	0.37 ± 0.06	0.36 ± 0.00
SRF	29.35 ± 3 %	92.53 ± 1.20 %	0.79 ± 0.03	0.29 ± 0.05	0.34 ± 0.04	0.28 ± 0.06
IDNN	22.31 ± 2 %	92.93 ± 1.02 %	0.80 ± 0.03	0.24 ± 0.03	0.25 ± 0.02	0.22 ± 0.02
IGMM	26.20 ± 3 %	92.79 ± 1.36 %	0.79 ± 0.03	0.27 ± 0.04	0.30 ± 0.03	0.25 ± 0.04
IRF	24.11 ± 3 %	92.51 ± 1.04 %	0.74 ± 0.03	0.24 ± 0.03	0.28 ± 0.03	0.23 ± 0.03

and spreading through the rest of the channels.

Figures 3-7 (b) and 3-7 (c) illustrate the behavior when using the concatenated spectral power and line length features to make a single framewise classification. Notice that both models place higher weight on the seizure interval but lack contiguity. In addition, the GMM model in Figure 3-7 (b) reacts strongly to activity after the seizure that is likely muscle artifact. Conversely, Figure 3-7 (d) shows the results when channels were classified independently using a GMM classifier. This strategy allows us to isolate seizure activity in different channels but does not impose spatial or temporal contiguity in classification.

Mean and standard deviations of evaluation metrics are shown for the JHH dataset in Table 3-II. As seen, our method outperforms all machine learning baselines. The clinical annotations demarcating seizure intervals have a tendency to extend beyond onset and offset. This coarse labeling and our stringent evaluation explains the low TPR across all models relative to results in the literature. The effect of the spatio-temporal transition prior is clear by comparing the CHMM with the IGMM. The prior allows our model to correctly place more posterior confidence in seizure, resulting in a higher TPR, while ignoring non-seizure baseline behavior that resembles seizure activity, simultaneously improving TNR. Of the baselines evaluated, only the SDNN surpassed the CHMM in TPR. The CHMM model surpasses the baseline methods in the summary scores AUC and F1, standing as much as a standard deviation above the best performing baselines.

Similar to the JHH dataset, Figure 3-10 depicts posterior beliefs for a representative

Table 3-III. Quantitative Results for the CHB dataset

Classifier	TPR	TNR	AUC	P	R	F1
CHMM	$57.43 \pm 5.64\%$	$\mathbf{98.67\% \pm 0.38}$	0.86 ± 0.03	$\mathbf{0.65 \pm 0.05}$	0.54 ± 0.06	$\mathbf{0.56 \pm 0.04}$
SDNN	$52.31 \pm 8.44\%$	$97.79 \pm 0.41\%$	$\mathbf{0.91 \pm 0.01}$	0.51 ± 0.05	0.48 ± 0.09	0.47 ± 0.07
SGMM	$\mathbf{59.40 \pm 4.28\%}$	$97.36 \pm 0.67\%$	0.87 ± 0.02	0.53 ± 0.05	$\mathbf{0.55 \pm 0.02}$	0.51 ± 0.03
SRF	$46.89 \pm 4.63\%$	$96.55 \pm 0.29\%$	0.88 ± 0.02	0.41 ± 0.03	0.45 ± 0.03	0.40 ± 0.03
IDNN	$27.73 \pm 2.09\%$	$95.62 \pm 0.22\%$	0.83 ± 0.02	0.27 ± 0.03	0.27 ± 0.02	0.24 ± 0.02
IGMM	$30.82 \pm 2.05\%$	$95.56 \pm 0.22\%$	0.83 ± 0.02	0.28 ± 0.03	0.30 ± 0.02	0.26 ± 0.02
IRF	$29.66 \pm 2.52\%$	$95.45 \pm 0.20\%$	0.76 ± 0.02	0.27 ± 0.03	0.29 ± 0.02	0.25 ± 0.02

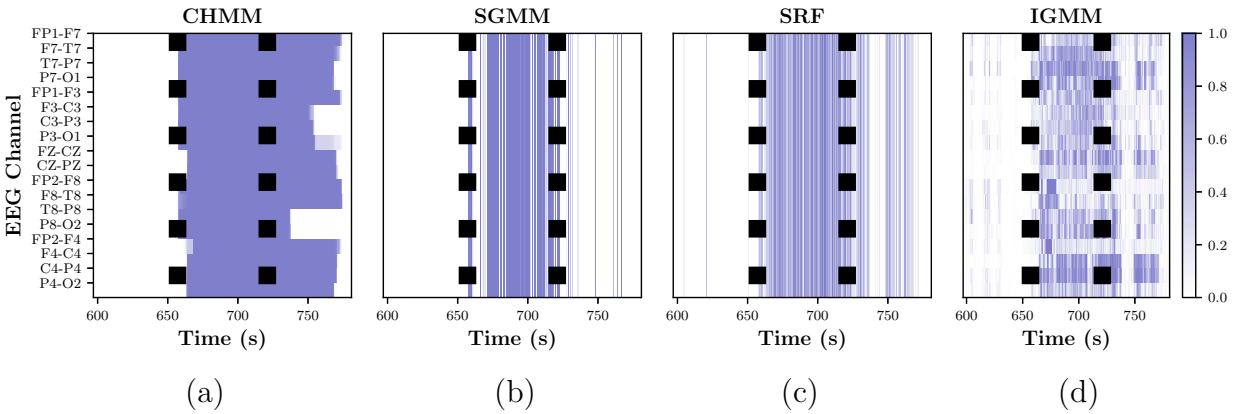


Figure 3-10. CHMM and selected baseline classification posteriors for a representative CHB patient. EEG channels are arranged on the y-axis with time along the x-axis. Seizure onset and offset are indicated by the vertical dashed lines. (a) Classification results using our CHMM model. Stacked features are used in conjunction with GMM and RF classifiers in (b) and (c), respectively. (d) Classification performed on each channel with a GMM. Posterior beliefs are shown in blue where intensity depicts the strength of the belief.

patient from the CHB dataset. Again, black dashed lines indicate annotated seizure onset and offset. Figure 3-10 (a) shows posteriors from the CHMM. Since the dataset does not include annotations of seizure type or focus, images of posteriors superimposed on the raw EEG signal and topographic details have been omitted, Once again, posteriors from both GMM baselines and the SRF are included.

Notice that in Figure 3-10 (a) the CHMM correctly labels the region containing the seizure but its detection extends slightly beyond the annotated offset. This is behavior can be attributed to the high degree of artifact present in the EEG signal post-seizure. While we lack clinical annotations regarding seizure types for this dataset, the presence of rhythmic

activity occurring simultaneously in all channels at the annotated seizure onset indicates that this is likely a generalized seizure. Our CHMM inference readily captures this phenomenon by turning all channels on simultaneously. Hence even though our CHMM prior assumes a assumes a focal spreading pattern, our method is flexible enough to capture multiple seizure types. Once again, the baseline classifiers suffer from the same drawbacks. The models trained on stacked features place higher beliefs in seizure regions but allow for many spurious onsets and offsets while models trained on each channel individually place lower posterior beliefs in seizure regions.

Quantitative results for the CHB dataset are shown in Table 3-III. The best performance under each metric is bolded. We have underlined when our model achieved the second highest performance in any metric. Our model outperforms baselines in most metrics while remaining within a standard deviation of the best performing baselines. In general, the GMM and RF classifiers using stacked features are slightly biased towards positive classifications, resulting in higher TPR and recall, but lower TNR and precision.

In summary, our Bayesian model outperforms all baseline methods in the JHH dataset. The focal epileptic seizures present in this population are most accurately classified using the CHMM with a transition prior designed for this task. Baseline methods fared poorly in part due to the heterogeneity of focal seizure presentations. In general, the CHB dataset contains better annotations, resulting in higher performance across the board. Furthermore, stacked feature vector based classification in the CHB dataset is better due to the presence of generalized seizures. Despite being tailored to capture focal spreading patterns, our CHMM maintains robust performance across both datasets. This cross hospital evaluation is the first of its kind and demonstrates our model's ability to generalize to diverse seizure types and patient populations.

3.6 Discussion

In this chapter a novel CHMM framework that captures the spatio-temporal propagation of a seizure for robust seizure detection was descriped. Using a variational approximation, we are able to efficiently perform inference and learn the model parameters despite its high dimensional state space. The CHMM is compared to baseline classifiers based on both individual and concatenated EEG features trained across patients. The framework is evaluated on EEG data acquired at two different hospitals, which was not previously been reported in the seizure detection literature. Our CHMM model outperformed or performed comparably to the best machine learning baselines in both the JHH dataset of focal epilepsy and the publicly available CHB dataset of pediatric epilepsy recordings.

Performance of our model in the JHH dataset exceeded that of all the baselines in all but one statistic. This improvement demonstrates our models efficacy in patient-agnostic seizure detection in a heterogeneous focal epilepsy dataset. In contrast, for the CHB dataset, our CHMM performed within a standard deviation of the best baseline approaches. We believe these differences arise from two clinically-relevant factors. First, our modeling choices regarding the spread of focal seizures mirror that of the patient cohort, as every patient in the JHH dataset has focal epilepsy. Thus our model more closely models the data than any of the baselines, leading to increased performance. In the CHB dataset, seizures appear to spread faster, indicating the presence of patients with generalized seizures. The simulations demonstrate that stacked feature baselines perform better in these conditions, thus explaining their better performance in this more homogeneous patient cohort. Second, the CHB dataset has been well-curated prior to its release, allowing better training of baseline models. In contrast, the JHH dataset has undergone minimal pre-preprocessing to better reflect clinical conditions.

Furthermore our model provided onset localization information for several patients in the JHH dataset, which highlights its potential use in localizing seizure foci. This coarse

localization in the EEG sensor space mirrors the early stages of clinical diagnosis, where EEG provides localization to a channel or lobe. As EEG is cheap to acquire, this coarse localization provides diagnostic information to guide expensive or invasive modalities, such as PET, MRI, or ECoG. Findings from EEG can be used in conjunction with these modalities for more comprehensive localization and treatment planning.

Interestingly, the CHB dataset includes generalized seizures which do not fit the assumptions of our spreading prior. Correspondingly, our simulated experiments show that the stacked baselines achieve better performance with faster spreading seizures. This behavior is mirrored in the CHB dataset where the CHMM performs on par with the baselines, particularly for the stacked feature representations. The advantage of our model stems from the ability of the CHMM transition prior to isolate the highest probability seizure interval. Despite the lack simultaneous onset in generalized seizures, this temporal data fusion still increases the CHMM performance relative to the individual channel baselines.

3.7 Conclusion

In this chapter we developed a spatio-temporal propagation model for epileptic seizures based on a CHMM architecture. We demonstrated our model on a dataset of focal epilepsy recordings and on a publicly available dataset of pediatric EEG recordings. By specifically modeling the spread of focal seizures, our model outperforms baseline classifiers in the dataset containing focal epilepsy recordings. In the dataset comprised of pediatric seizure recordings, our model performs the best or comparably to our baselines. While commercial seizure detection algorithms exist, they have yet to supplant manual annotation. Accurate and reliable seizure detection remains a clinical necessity. Our experimentation here indicates that direct modeling of cross-channel interactions present in EEG signals can improve the performance of seizure detection algorithms. This modeling shows the ability to provide information capable of aiding the localization process for diagnosis and treatment planning of focal epilepsy.

While the cross-channel information fusion in the CHMM yields gains in seizure detection, further modeling additions have the capability to improve the performance of the model. The CHMM analyzes seizure activity in each channel individually, sharing information with neighboring channels. This spatio-temporal information fusion encourages the model to track seizure activity as it spreads through the scalp electrodes. However, the model places no restrictions over the number of onsets in individual channels, allowing the CHMM to predict seizure onsets in discontiguous regions of the brain. Similarly, the model allows channel offsets to occur at separate points in time. While this behavior does not affect the utility of the model, restricting the model to enforce synchronous channel offset would result in more accurate predictions of seizure activity. Most importantly, localization from the CHMM model is heuristic and requires post hoc analysis of the model outputs. Incorporating localization directly into the model would improve the clinical utility further, allowing automated prediction of the seizure onset zone. Extensions to the CHMM model addressing these shortcomings will be explored in Chapter 5.

In exploratory analysis prior to the developoment of the CHMM model presented here, performance of many features from the seizure detection literature was evaluated. For the work presented in this chapter, spectral bands and line length were identified as robust and simple features. This finding has been validated in the literature by [112], where both line-length and spectral band features were recommended after the evaluation of many features from the seizure detection literature. Though spectral band and line-length features are simple and effective, deep neural networks offer the potential to learn powerful feature representations directly from raw data. Incorporating neural network likelihoods for EEG analysis will be explored further in Chapter 4.

Chapter 4

CHMM-CNN

4.1 Introduction

4.1.1 Chapter Contributions

In the previous chapter, work from [9] that used a Coupled Hidden Markov Model (CHMM) to track the propagation of a seizure across the scalp was presented. This method relied on carefully selected features in order to learn a highly strutured likelihood function. Despite leading to good performance, feature extraction focused specifically on a small number of spectral features combined with a line-length feature. These features, combined with likelihood scoring using a restricted set of functions, likely missed relevant seizure information. Here we explore data-driven strategies using deep architectures to directly learn more effective representations and analysis functions.

This chapter presents an integrated framework for epileptic seizure detection that blends the interpretability of the CHMM with advancements in deep learning and was originally published in [8]. The CHMM presented in [9] and detailed in the previous Chapter is augmented with a Convolutional Neural Network (CNN) likelihood model. This CNN replaces the feature extraction and GMM likelihood of the original CHMM formulation with a deep network capable of learning features relevant for detection from limited amounts of training data. We demonstrate the combined CHMM-CNN framework on multichannel EEG data acquired from the JHH and CHB datasets. Our CHMM-CNN framework correctly

identifies more of the annotated seizure activity in both datasets than comparable baseline methods. This performance suggests that by incorporating neural networks, the performance of graphical model based seizure detection can be improved.

4.1.2 Feature Engineering for EEG Analysis

As detailed in Section 2.2, automated seizure detection has been an active field of research for the past three decades. Most algorithms follow a two-stage machine learning pipeline consisting of (1) feature extraction from the EEG signal over short time windows, followed by (2) a binary classifier to identify seizure versus non-seizure intervals [42]. This end-to-end pipeline was exemplified by Shoeb et al. [45] where power features in different frequency bands were used in conjuction with a support vector machine classifier.

Prior work in the seizure detection community has focused largely on the feature extraction step. A diverse set of feature extraction methods for seizure detection has been developed, drawing from a range of signal processing areas. By combining features from these different sources, prior work in seizure detection has sought to identify effective ensembles of features. For example, [51] computes a diverse feature set consisting of time domain and frequency domain features. Work such as [52–54] performs feature extraction based on chaos theory after applying frequency based signal decompositions.

While combining diverse features shows performance gains over classification using features from a single domain, the generalization power is fundamentally limited by the chosen features. Hand designed feature engineering relies on signal processing to extract statistics hypothesized to be relevant to a given task. Shortcomings both in the generation of these hypotheses and the ability of signal processing techniques to adequately capture these phenomena contribute to limitations in the performance of feature engineering. Alternatively, deep networks learn representations that capture facets of the data directly applicable to the task at hand [114]. This improved analytical power can come at the cost of interpretability, as these features may lack inuitive explanation.

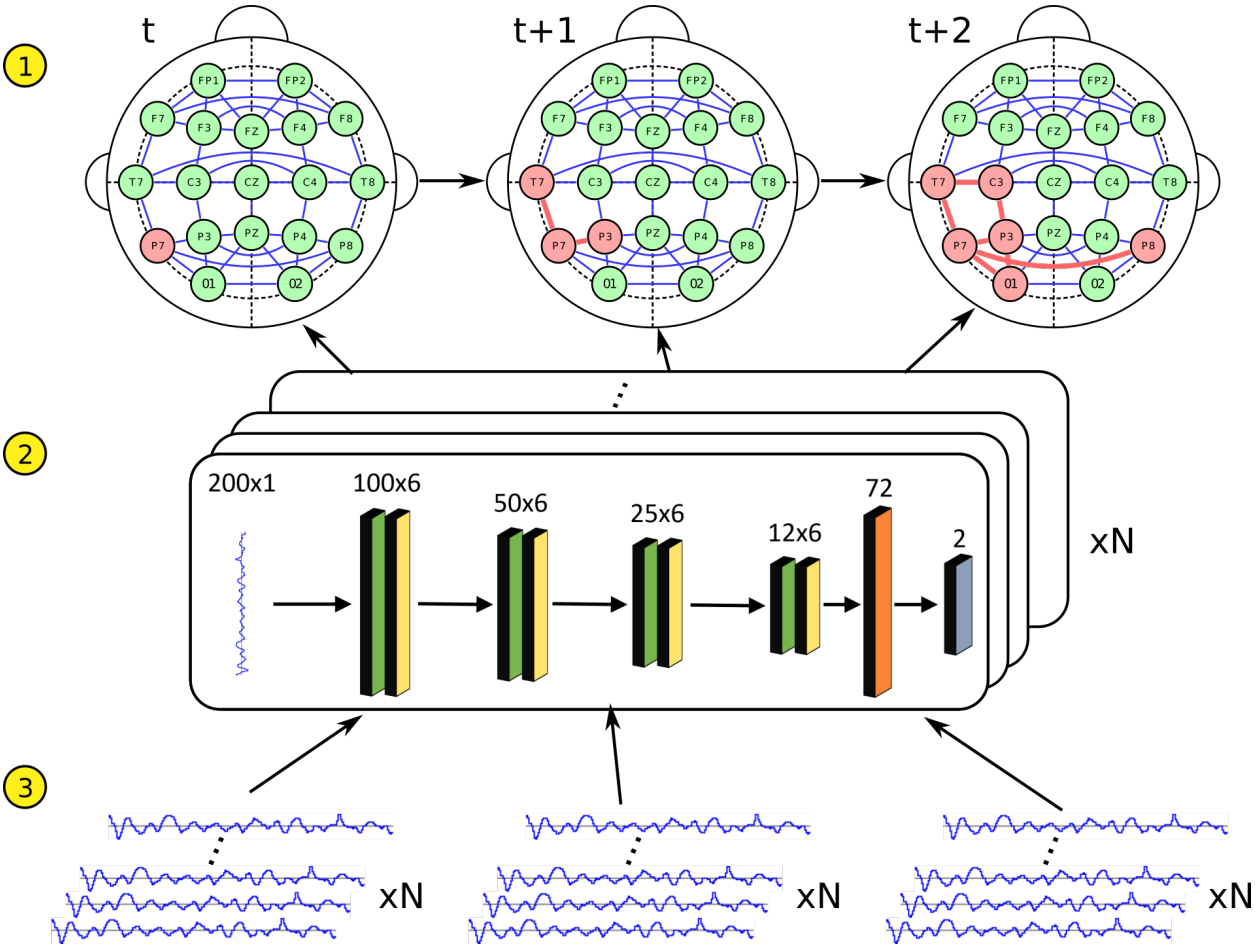


Figure 4-1. Detail of the inference procedure. Time flows to the right while information flows upwards. In the third row, we depict the raw EEG signal. The signal from each channel is fed into a dedicated CNN for scoring in the second row. The first row depicts a hypothetical seizure spreading through the propagation network of the CHMM.

4.2 Integrating CNNs in the CHMM

Figure 4-1 outlines our modeling strategy. Raw EEG signal from each channel in row three is fed directly into the CNNs in row two, where one CNN is trained for each channel. The CNNs score the signal for seizure activity and feed this information into the CHMM prior shown in row 1. The CHMM fuses these scores across the scalp and through time to perform posterior inference for seizure activity. Below, we formalize the adjustments to the original CNN allowing us to incorporate deep likelihoods.

4.2.1 Nonparameteric Likelihood via Convolutional Neural Networks

CNNs have become standard in computer vision due to their ability to learn spatially invariant features across multiple scales [114]. At a high level, the early layers learn simple features, such as edge detectors, while subsequent layers learn more and more complicated features. CNNs are also becoming popular for one-dimensional and time series data, where they provide a valuable alternative to the standard Recurrent Neural Network (RNN). While RNNs have been particularly effective in analyzing short sequences, CNNs with large receptive fields can be trained much faster than RNNs for long sequences. In addition, CNNs are restricted to learning highly structured functions composed of convolutions, which reduces their ability to overfit when training data is limited [114]. While CNNs are powerful tools for data analysis, they suffer from a lack of interpretability. However, from a clinical standpoint, we are primarily concerned with the seizure propagation patterns, as opposed to the underlying feature representation. Our hybrid approach captures the clinically relevant information by using a directly interpretable CHMM prior for capturing seizure spreading while giving the CNN free rein over the data likelihood to improve EEG signal analysis, resulting in gains in detection performance.

One important caveat to integrating a CNN data likelihood is that, by default, a CNN is trained for posterior inference. Namely, given the input data $X_i[t]$, the CNN will output a soft class assignment of seizure versus baseline, i.e. $P(Y_i[t] \mid X_i[t])$. In contrast, the joint distribution in (3.5) relies on the data likelihood, $P(X_i[t] \mid Y_i[t])$. We can obtain this factor by applying Bayes' rule:

$$P(X_i[t] \mid Y_i[t]) = \frac{P(Y_i[t] \mid X_i[t])P(X_i[t])}{P(Y_i[t])} \propto \frac{P(Y_i[t] \mid X_i[t])}{P(Y_i[t])} \quad . \tag{4.1}$$

Notice that we ignore the marginal probability $P(X_i[t])$, as this term is the same regardless of the class label, and we only require data likelihoods up to a constant factor for posterior inference. Hence, we can rescale the CNN output by $P(Y_i[t])$ to arrive at a surrogate likelihood

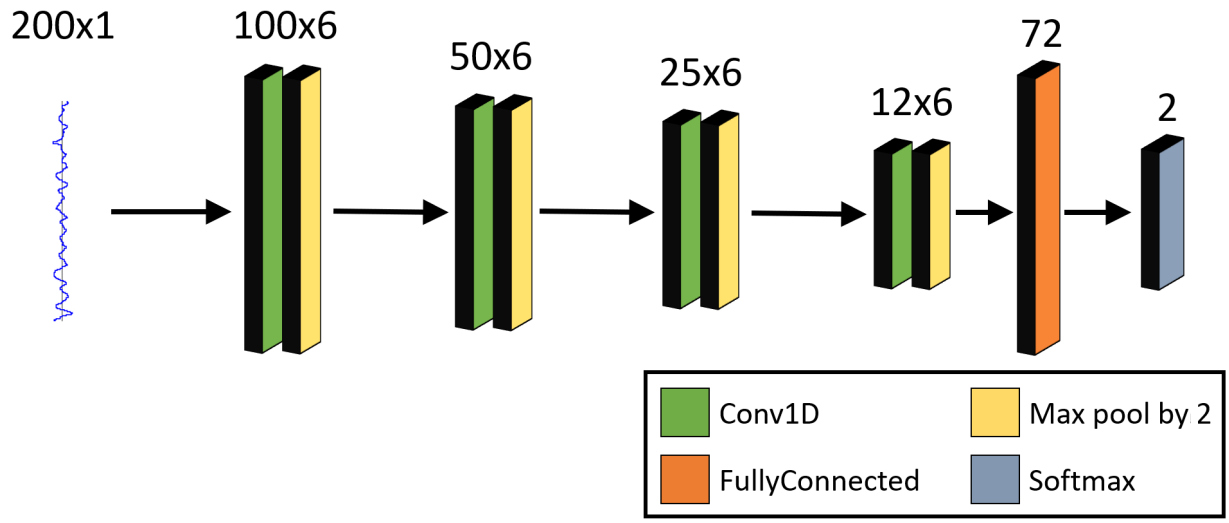


Figure 4-2. Convolutional neural network architecture used in this work

term [71]. We approximate $P(Y_i[t])$ by the proportion of seizure versus baseline in the dataset, i.e.

$$\hat{P}(Y = 1) = \frac{\#\text{ seizure windows}}{\#\text{ windows}}, \qquad \hat{P}(Y = 0) = 1 - \hat{P}(Y = 1) \ . \tag{4.2}$$

The rescaling of the discriminative posterior in (4.1) using the approximate prior over states in (4.2) will serve as the likelihood in our CHMM-CNN model.

4.2.2 Fitting the CHMM-CNN Model

We fit the PGN-CNN using the variational algorithm presented in Chapter 3. We approximate the latent posterior as the product of independent HMM chains.

$$P(\mathbf{Y} \mid \mathbf{X}) \approx Q(\mathbf{Y}) = \prod_{i=1}^{N} \frac{1}{Z_{Q_i}} Q_i(\mathbf{Y}_i) = \prod_{i=1}^{N} \frac{1}{Z_{Q_i}} \prod_{t=1}^{T} \psi_i[t](Y_i[t] \mid Y[t-1]_i)\omega_i^t(Y[t]_i) \ . \tag{4.3}$$

The factors $\omega_i^t(Y[t]_i)$ and $\psi[t]_i(Y[t]_i \mid Y[t-1]_i)$ encode the emission and transition terms of the approximating HMMs, respectively. We infer the latent posterior distribution by iteratively running the forward-backward algorithm [57] over each of the individual chains, while holding the remaining chains constant. The forward-backward algorithm calculates the following posterior statistics under the distribution Q.

4.2.3 Neural Network Implementation

We implemented the CNN in PyTorch. The CNN consists of 4 convolution and pool layers as shown in Figure 4-2. Each layer uses 6 channels with a kernel size of 5 samples and 2 sample zero padding to maintain a constant size. A LeakyReLU activation, where $\text{LeakyReLU}(x) = \max(0, x) + 0.01 \cdot \min(0, x)$, is applied at each layer. A max pooling operation with a kernel size of 2 and a stride of 2 is applied, halving the size of the representation at each layer. After the final convolution, the hidden units are concatenated and passed to a single linear layer for classification using a softmax activation.

During experimentation in the design of this network, we investigated similar architectures of varying depths, numbers of channels, and activation functions. Networks with saturating activations failed to train in some cases, perhaps due to the presence of artifact with extreme amplitudes. We found the LeakyReLU to be the most robust, likely due to the fact that it does not saturate.

The CNNs were trained discriminatively with a cross entropy loss function prior to posterior inference. We trained separate CNN classifiers for each EEG channel to capture behavior specific to different parts of the scalp. Stochastic gradient descent was performed using the Adam optimizer with a batch size of 32 samples and a learning rate of 0.5. We trained each CNN for 60 epochs, which was sufficient to achieve reliable performance without overfitting.

4.3 Evaluation

4.3.1 Baseline Comparisons

We compare our CHMM-CNN detection performance to baseline methods ranging from simple classifiers on hand selected features to a fully CNN strategy. The features used in Chapter 3 (sum of spectral components and line-length) were used for all non-CNN baselines. Recordings were randomly assigned to 5-folds for cross validation. Training was performed on

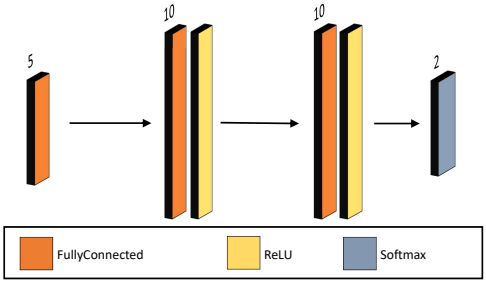


Figure 4-3. Artificial neural network used for seizure detection in this work.

4 folds while the remaining fold was used for testing. The baseline methods are summarized below.

4.3.1.1 CNN

We implement an end-to-end deep learning pipeline based on the CNN classification architecture described in Section 4.2.2. This comparison evaluates the predictive value of the CNN without the smoothing in the CHMM prior.

4.3.1.2 CHMM

We implement the original CHMM model proposed in Chapter 3 which assumes a Gaussian Mixture Model (GMM) likelihood using the suggested parameter settings. This comparison will evaluate the performance gain in using a non-parametric likelihood with data driven learning from the raw EEG signal.

4.3.1.3 ANN

Similar to the CNN baseline, we evaluate the performance of the predefined features as inputs for an Artificial Neural Network (ANN) classifier. Due to the relatively small feature space we opted for a small ANN shown in Figure 4-3 to avoid overfitting. Our networks are composed of two hidden layers with 10 units each. Each layer is fully connected with Rectified Linear

Unit (ReLU) activations. The final output layer contains two nodes with a softmax activation applied. Thus the final layer represents the posterior probability of a hidden state given the associated feature vector.

4.3.1.4 GMM

Finally, we implement a simple GMM classifier based on the precomputed EEG features. The inclusion of this baseline allows us to evaluate the relative performance of the CNN, ANN, and parametric GMM likelihoods directly without the inclusion of the latent seizure spreading prior.

4.3.2 Performance Metrics

Our performance metrics are based on the maximum a posterior (MAP) estimate of baseline versus seizure for each method and are presented as averages across test folds. Since the clinical seizure annotations tend to be overly generous and do not contain spatial information about onset, we aggregate the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) across all windows, channels, and all seizure recordings. In general, the recordings contain muscle artifact directly following the seizure which confounds the offset for all methods. Therefore, we count any seizure classification occuring within the annotated seizure region as TP. However any contiguous classifications continuing past the annotated offset is not counted in our evaluation statistics.

Below we detail the summary statistics computed for each model. True Positive Rate (TPR), also known as recall, represents the total rate of correct classification. False Positive Rate (FPR) represents the rate of incorrect classification of baseline regions as seizure after excluding classifications beginning within the seizure region. We calculate a lower bound on the Area Under the Curve (AUC) using these two metrics. Precision (P) details the ratio of correct seizure classifications to the total number of seizure classifications. In addition to AUC, the F1 score offers a similar summary by computing the harmonic mean of P and TPR.

Mathematically, these statistics are given by:

$$TPR = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} (TP_i + FN_i)} \qquad FPR = \frac{\sum_{i=1}^{N} FP_i}{\sum_{i=1}^{N} (TN_i + FP_i)}$$

$$P = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} (TP_i + FP_i)} \qquad F1 = 2\frac{P \cdot R}{P + TPR}$$

$$AUC = FPR \cdot TPR/2 + (1 - FPR)(1 + TPR)/2 \ .$$

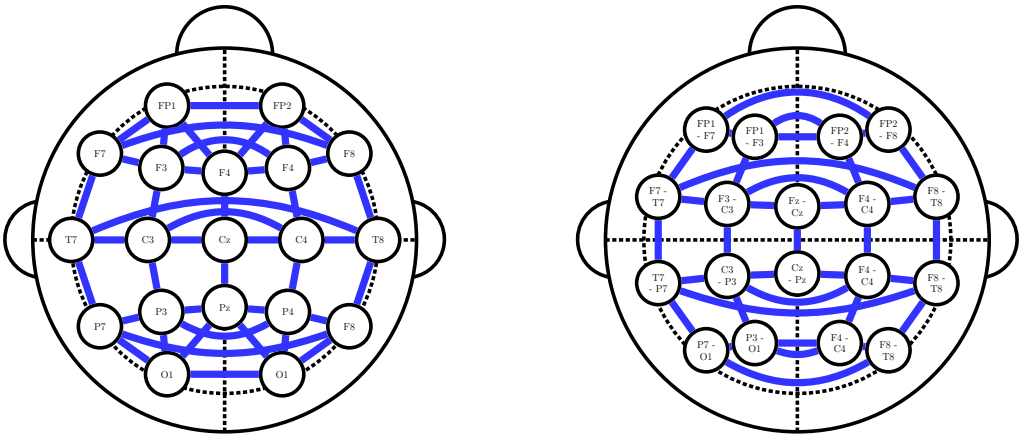
4.4 Experimental Results

4.4.1 Data and Preprocessing

Epileptic seizures are extremely heterogeneous. For example, generalized seizures manifest across the entire cortex at once, whereas focal seizures originate from a single area and may spread to other regions of the cortex. Given this heterogeneity, we again evaluate our algorithm on the JHH and CHB datasets. In this experiment, 90 seizures from 15 adult patients in the JHH dataset were used. All 185 recordings from 24 pediatric patients from the CHB dataset were used as well.

Besides the patient populations, another difference between the two datasets is the acquisition protocol. The JHH dataset contains the original recordings of the 10/20 international system in common reference. As such, we employ the original coupling graph presented in the previous chapter and depicted again in Figure 4-4 (a). In contrast, the CHB dataset uses the longitudinal montage, which forms difference channels by subtracting the signals in neighboring electrodes. We specify a propagation network appropriate for this montage as shown in Figure 4-4 (b). This coupling preserves neighboring and contralateral relationships on the scalp from the original prior.

Our recordings contain one seizure and up to ten minutes of pre- and post-seizure baseline. For the CHB data, these segments were clipped from the original release. EEG channels were low and highpass filtered at 50 and 1.6 Hz, respectively. A notch filter at 60 Hz was applied to remove any remaining power supply artifact. As in [6], 4 spectral features and one



(a) Common Reference Coupling

(b) Longitudinal Montage Coupling

Figure 4-4. Propagation paths for the (a) common reference and (b) longitudinal montage.

line-length feature were extracted from 1 second windows with a 250 ms overlap. The CNN model was trained directly on the raw EEG signal from the 1 second windows.

4.4.2 Detection Performance

Table 4-I. Results for the JHH dataset

Trial	TPR	FPR	AUC	P	F1
CHMM-CNN	**0.45**	0.010	**0.72**	0.79	**0.57**
CHMM	0.37	0.0083	0.68	**0.80**	0.50
CNN	0.19	0.010	0.59	0.62	0.28
DNN	0.11	**0.0070**	0.55	0.58	0.18
GMM	0.18	0.015	0.58	0.52	0.27

Table 4-II. Results for the CHB dataset

Trial TPR	FPR	AUC	P	F1	
CHMM-CNN	**0.61**	0.013	**0.80**	0.74	**0.67**
CHMM	0.571	**0.0067**	0.78	**0.83**	**0.67**
CNN	0.27	0.0071	0.63	0.70	0.39
DNN	0.23	0.0071	0.61	0.66	0.34
GMM	0.26	0.010	0.62	0.61	0.37

Tables 4-I and 4-II report the seizure detection performance averaged across the testing folds for both the JHH and CHB datasets, respectively. We have reported True Positive Rate

(TPR), False Positive Rate (FPR), Area Under the Curve (AUC), Precision (P), and F1, as described in Section 4.3.2.

Our CHMM-CNN dramatically outperforms all of the baseline methods on the JHH dataset. The only drawback is a slightly higher FPR, since our CNN shows more sensitivity to seizure activity, and classifies slightly more baseline as seizure. Despite the numerical increase in FPR, the increased sensitivity is valuable in the clinic, particularly when augmenting the expert manual inspections. Moreover, these spurious detections are compensated by more accurate true detections, which are reflected in the AUC, precision, and F1 measures. We emphasize that our evaluation metrics are much more conservative than in prior studies, which is why the TPR seems uniformly low. Instead of measuring singular detections within the annotated seizure period, we aggregate over channels and windows. This allows us to evaluate not only correct detections of seizures but *how much seizure activity* our algorithms are capable of discerning.

Interestingly, the same detection trends are seen in the CHB data, despite our CHMM spreading prior being designed for focal and not generalized seizures. The CHMM-CNN achieves the best TPR and AUC as well as a comparable F1 score. In short our flexible data likelihood based on the CNN allows us to learn complex data representations that better separate seizure from baseline. This leads to better detection rates, which is valuable for clinical planning.

Finally, we note that the channel-wise baselines are uniformly bad. Detecting seizure activity is, in general, a relatively difficult problem. Both seizure and baseline contain high amplitude muscle artifact, which confound the detection over short time windows. In addition, the data distributions are highly overlapping, with seizure activity often resembling normal behavior. The effect of the prior in the CHMM-CNN and CHMM for data fusion across channels is apparent.

Figure 4-5 and Figure 4-6 show the classification posteriors of each model for an example seizure in the JHH and CHB datasets, respectively. EEG channels are presented on the

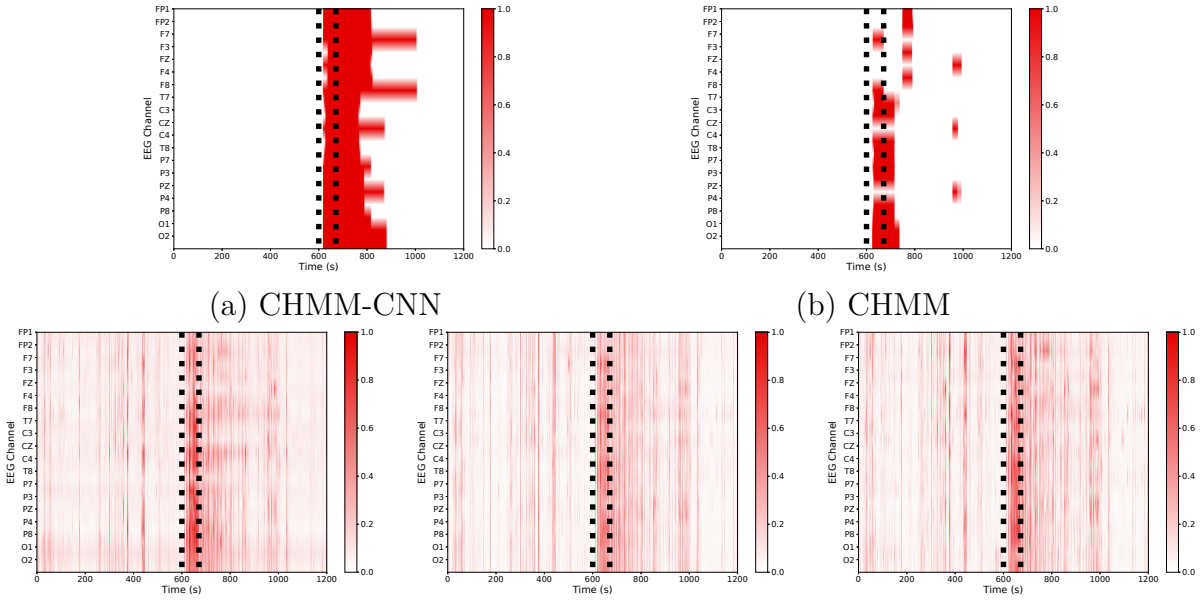
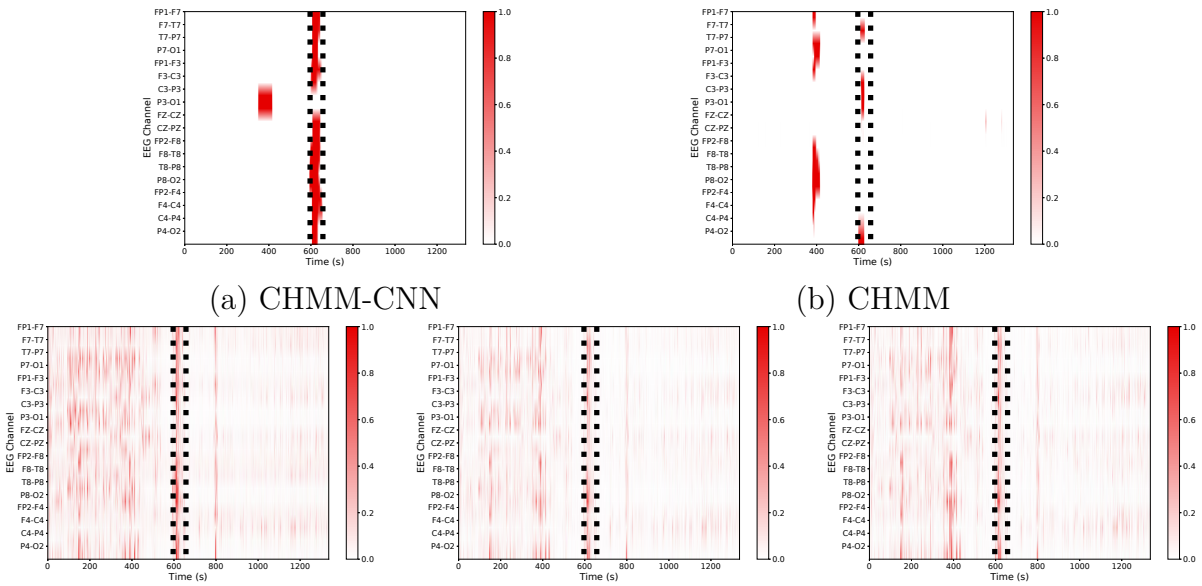


Figure 4-5. Estimated posteriors for a single seizure from the JHH dataset. EEG channels are shown on the y-axis and time proceeds in the x-direction. The first row shows models with a CHMM prior. The second row shows channel-wise classifications.

y-axis while the x-axis shows time. The dashed black lines correspond to annotations for seizure onset and offset. Red indicates the posterior probabilities of the seizure state. The CHMM-CNN correctly classifies more of the annotated seizure in Figure 4-5 than any of the other models. Each model incorrectly activates during the period immediately following the seizure, responding to the presence of artifact. However, the CNN likelihood model places more confidence in the seizure region, allowing for more correct classification. In contrast, the ANN and GMM identify strong seizure-like activity in the artifact following the actual seizure, causing an incorrect classification by the CHMM. In Figure 4-6 the CHMM-CNN correctly classifies more of the annotated seizure than the CHMM but makes a false positive, while the CHMM classifies only a small portion of the seizure and responds strongly to the artifact prior to the seizure.



(a) CHMM-CNN (b) CHMM

(c) CNN (d) ANN (e) GMM

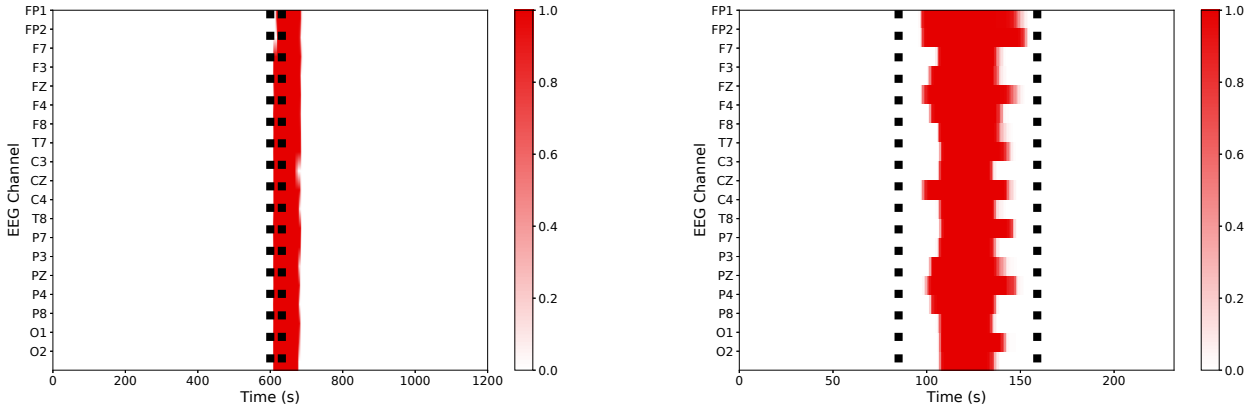
Figure 4-6. Example posteriors from the CHB dataset. CHMM and likelihood models are shown in the first and second rows, respectively.

4.4.3 Seizure Localization

Surgical resection is the standard-of-care for medically refractory focal epilepsy. The latent propagation prior of our CHMM-CNN has the potential to aid in seizure localization. Figure 4-7 shows two classifications from the CHMM-CNN. Clinical annotations for the seizure in (a) and (c) suggest an origin in the right temporal lobe and spreading left. Likewise, the annotations in (b) and (d) suggest a left frontal lobe onset. The localization information provided by our model agrees with the annotated foci. Remarkably, this spreading behavior is *learned in a completely unsupervised manner* based on the clinical hypotheses embedded in the CHMM prior. This result highlights the promise of integrating model-based and data-driven approaches for medical imaging applications.

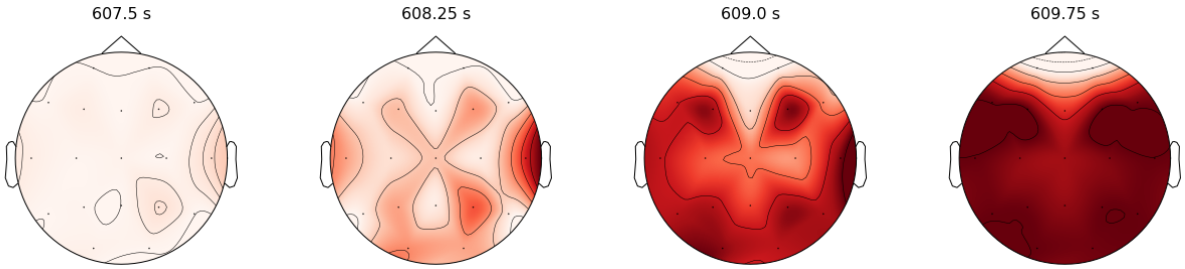
4.5 Conclusion

This chapter described the CHMM-CNN, the first generative model-deep learning hybrid for epileptic seizure detection. Our framework captures the spatio-temporal spread of a seizure

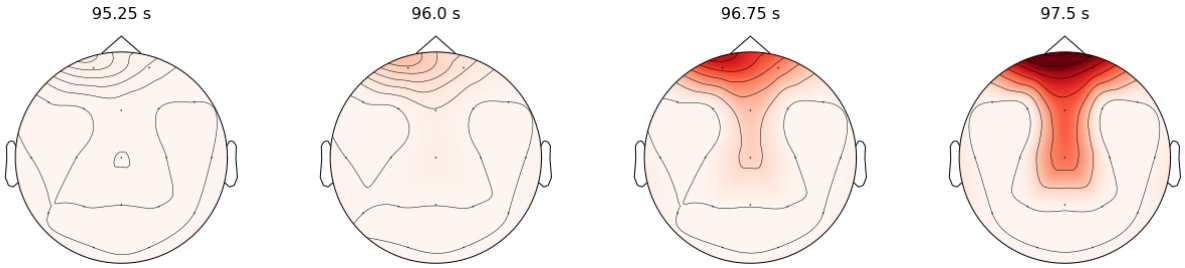


(a) CHMM-CNN posterior classifications

(b) CHMM-CNN posterior classifications



(c) Topographic detail of seizure spread in (a) originating in the right hemisphere



(d) Topographic detail of seizure spread in (b) originating in the left frontal region

Figure 4-7. Example seizure tracking from the JHH dataset. (a,b) Posteriors for all channels. (c,d) Topographic detail showing posterior onsets in clinically annotated regions.

through a structured CHMM prior, while allowing for a complex likelihood function that is implicitly learned via a CNN. This data driven approach learns representations directly from the raw EEG signal, improving upon feature extraction techniques. At the same time the CHMM preserves clinical interpretability and acts as a local smoothing process for the CNN outputs based on limited training examples.

As opposed to the fully graphical model CHMM presented in Chapter 3, the CHMM-CNN decouples learning of the likelihoods and propagation priors into two separate problems. CNN likelihoods must first be trained discriminatively on the raw EEG data. After training, the

CHMM is learned using these CNN likelihood functions. While a fully graphical modeling approach allows for the prior and likelihoods to be learned jointly, we observe that moving from hand crafted feature design to learning neural likelihoods improves the performance of the model. We evaluate our method on clinical data from two hospitals with distinct patient populations. In both cases, our CHMM-CNN achieved higher true positive detection and AUC than any of the baseline methods.

While the incorporation of a deep network likelihood resulted in marked improvements over our original approach using hand designed features and GMMs, the CHMM-CNN suffers from similar limitations as the original CHMM. Namely, without a mechanism to encode a global seizure state, individual electrode channels are allowed to enter and exit the seizure state on their own. This results in unwanted behavior such as multiple onsets for a single seizure and channels remaining in the seizure state long after most channels have stopped predicting seizure activity. In addition, the CHMM model lacks a mechanism to directly perform inference over localization areas. In the next chapter, we propose a method to address these shortcomings of the CHMM.

Chapter 5

Regime-Switching Markov Model for Detection and Localization

5.1 Introduction

In the previous chapters, we explored the CHMM for seizure detection. By allowing seizure activity to be predicted in individual EEG electrodes, this approach demonstrated potential for tracking seizure activity as it spreads through the brain. Through coupling between channels according to a biologically informed graph network, seizure activity was allowed to propagate along hypothesized seizure spreading pathways. Despite this novel approach for multi-channel information fusion, the CHMM suffered from several draw backs. Firstly, while able to provide localization information, this information was heuristic and required post hoc inspection of the algorithm's output. Secondly, while information between channels is shared, the CHMM model makes no restrictions on the number of onsets occurring whithin a given recording. Third, the CHMM lacks a mechanism to enforce that all channels turn off at the end of a seizure, which results in individual channels remaining in the seizure state long after the seizure has ended.

In this chapter, we propose a novel Regime-Switching Markov Model for Propagation and Localization (R-SMMPL) and demonstrate its effectiveness for both seizure detection and SOZ localization. This model decouples detection, propagation, and localization into three interacting sets of variables. A switching variable controls the dynamic regime of the system,

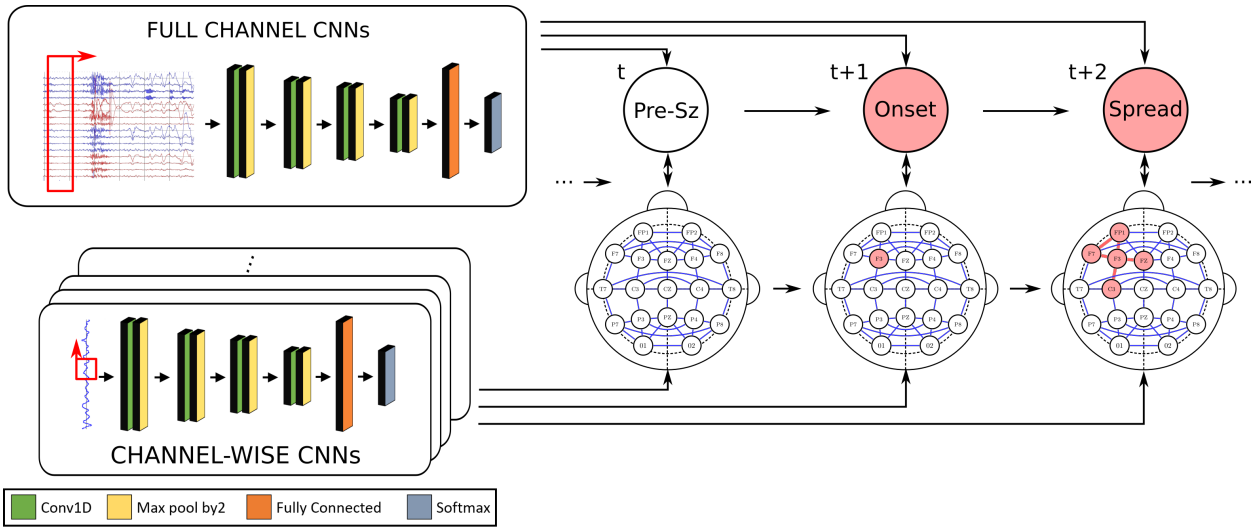


Figure 5-1. Model schematic. The left side depicts the CNNs used for likelihood scoring prior to inference. The orientations of the kernels and convolutions are shown in red. At right the system is shown at seizure onset. Channel nodes and blue connections define the propagation graph $\mathcal{S}$. The seizure switching chain is shown above, where seizure activity is shown in red while normal activity is white. During spreading, seizure propagates through $\mathbf{Y}^{nj}$ (below) along the blue propagation pathways.

acting as a seizure onset and offset detector. In response to changes in this switching variable, we use a modified CHMM, as in the previous chapters, to track the spread of seizure activity when seizures are detected. By learning SOZ distributions at the patient level, location variables allows us to tie onset location between multiple recordings.

Taken together, this combination of variables addresses the limitations of the CHMM and ensures that seizures originate at a single time point from only one location. Furthermore, all channels are required to exit the seizure state at the end of the seizure, ending the problem of lingering seizure activity in single channels. Extending beyond the heuristic localizations in the CHMM, the R-SMMPL allows us to pool information across multiple seizure recordings into an onset zone hypotheses for each patient. To our knowledge, the R-SMMPL represents the first unified framework for both seizure detection and localization from scalp EEG.

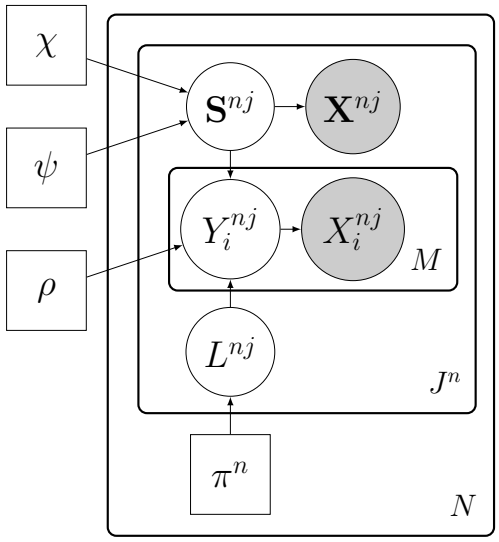


Figure 5-2. R-SMMPL plate model. Squares denote parameters while circles indicate random variables. Observed variables are shaded gray. In this model, the multi-channel EEG signal and individual channel EEG signals are considered to be separate random variables.

5.2 R-SMMPL Formulation

Figure 5-1 shows a schematic representation of the R-SMMPL. CNNs likelihoods operate on the entire EEG signal as well as the channel-wise signals from each electrode. Using a graphical model composed of a hierarchy of HMM chains, these likelihoods are used to detect, track, and localize seizure activity. Figure 5-2 shows our graphical model and Table 5-I describes the variables used in the R-SMMPL. Figure 5-1 illustrates the temporal evolution of a seizure as represented by the inner plate of Figure 5-2.

The plate notation in Figure 5-2 describes how multiple seizure recordings are aggregated for a single patient. Bold variables represent collections across time and, if applicable, EEG channel. Notably, in this chapter, we assume that the multi-channel EEG signal $\mathbf{X}$ and the individual EEG signals X_i are independent random variables generated from the overall seizure state S and individual electrode seizure states Y_i, respectively. This assumption allows us to train likelihoods for the mulit-channel and individual electrodes separately and integrate them into the R-SMMPL without added complexity. Seizure propagation pathways are defined by the graph $\mathcal{S}$, shown by the channel nodes and blue lines in Figure 5-1. Notice

Table 5-I. Variable descriptions. $\mathbf{S}^{nj} \triangleq \{S^{nj}[t]\}_{t=0}^{T}$, $\mathbf{Y}^{nj} \triangleq \{Y_i^{nj}[t]\}_{t=0,i=1}^{T,M}$, and similarly for $\mathbf{X}^{nj}$ and $\mathbf{C}^{nj}$, respectively.

Symbol	Description
$S^{nj}[t]$	Switching chain for regime-switching
$Y_i^{nj}[t]$	Seizure state of EEG channel i
L^{nj}	Seizure onset location
$X_i^{nj}[t]$	EEG observation in channel i
$X^{nj}[t]$	Full EEG observation for all channels
$\mathcal{S}$	Seizure propagation graph
π^n	Onset distribution for patient n
χ	Seizure onset probability
ψ	Seizure offset probability
ρ	Seizure propagation constant

that we have coupled neighboring and contralateral channels, as these are the most common propagation patterns observed in EEG seizure recordings.

Let N be the total number of patients, J^n be the number of recordings belonging to patient n, and let superscript nj denote recording j in patient n. M is the number of EEG channels (typically 18-20) and T is the recording duration. The switching chain $S^{nj}[t]$ tracks the overall state of the system as a seizure occurs and progresses. The chains $Y_i^{nj}[t]$ track the spread of seizure activity through EEG channel i. Each recording has an onset location $L^{nj} \in \{1, 2, \ldots, M\}$. Emission variables $X^{nj}[t]$ and $X_i^{nj}[t]$ are observed from the switching chain S^{nj} and the individual CHMM chains $Y_i^{nj}[t]$, respectively. The joint distribution is:

$$P(\mathbf{L}, \mathbf{S}, \mathbf{Y}, \mathbf{X}, \mathbf{C}) = \prod_{n=1}^{N} \prod_{j=1}^{J^n} P\left(L^{nj}\right) \prod_{t=1}^{T} P\left(S^{nj}[t] \mid S^{nj}[t-1]\right) P\left(X^{nj}[t] \mid S^{nj}[t]\right)$$
$$\prod_{i=1}^{M} P\left(Y_i^{nj}[t] \mid Y_i^{nj}[t-1], L^{nj}, S^{nj}[t], Y_{ne_{\mathcal{S}}(i)}^{nj}[t-1]\right) P\left(X_i^{nj}[t] \mid Y_i^{nj}[t]\right)$$

5.2.1 Localization

For each patient, a multinomial location parameter π^n represents the probability that a seizure from patient n will exhibit onset in a particular EEG channel. For each recording,

the onset location L^{nj} is drawn from π^n.

5.2.2 Regime-Switching and Propagation

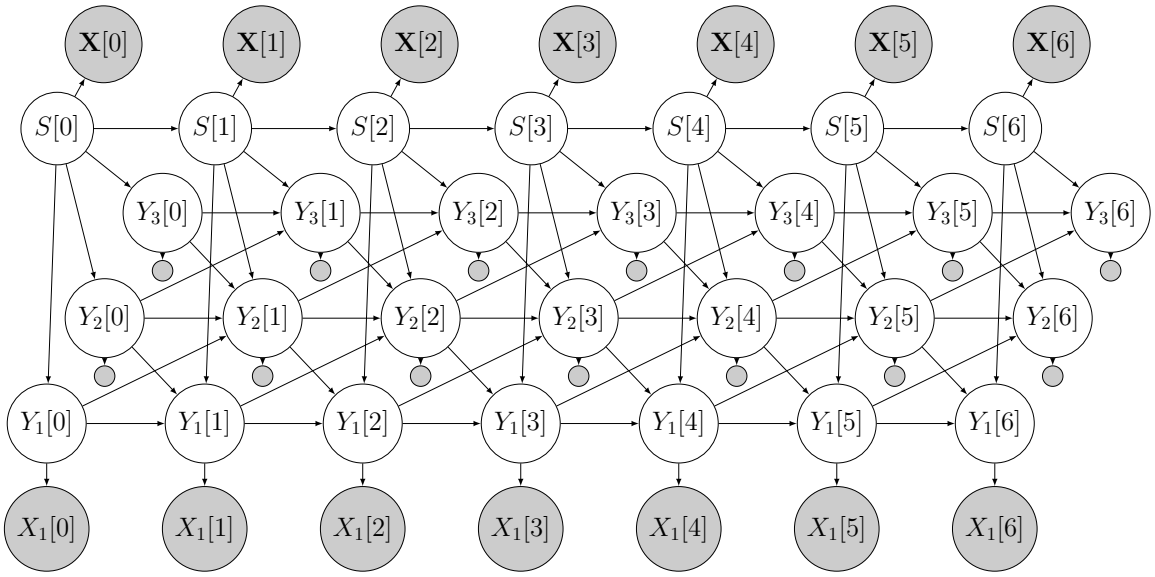


Figure 5-3. Directed acyclic fraphical model depicting the R-SMMPL. The latent S and Y chains are shown in white and observed variable F and C are shown in grey. Only three Y channels are shown and the location variable L is omitted for clarity.

An unrolled DAG depicting the interdepence between $S[t]$ and variables Y is shown in Figure 5-3. The variables $S^{nj}[t]$ progress through five states: pre-seizure baseline, seizure onset, seizure spreading, seizure offset, and post-seizure baseline. The variables $Y_i^{nj}[t]$ are binary and denote either normal ($Y_i^{nj}[t] = 0$) or seizure ($Y_i^{nj}[t] = 1$) in channel i at time t. Each recording begins in pre-seizure baseline with all channels exhibiting normal EEG activity.

Seizure onset and spread are shown on the right side of Figure 5-1. At each time step, there is probability χ of a seizure occurring, represented by the switching chain $S^{nj}[t]$ transitioning into the onset state. At onset, chain $Y_{L^{nj}}^{nj}[t]$ enters the seizure state, representing abnormal activity at the seizure onset zone. The switching chain $S^{nj}[t]$ then immediately transitions to the spreading state.

During spreading, seizure activity is allowed to spread through the seizure propagation

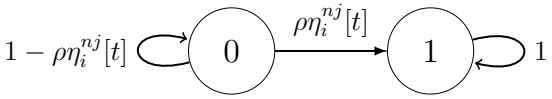


Figure 5-4. Transition diagram for $Y_i^{nj}[t]$ when $S^{nj}[t]$ is in the spreading state.

graph $\mathcal{S}$ defined in Figure 5-1. This spreading is governed by the probabilities in the transition diagram in Figure 5-4. The probability that $Y_i^{nj}[t]$ enters the seizure state (1) from the non-seizure state (0) at time t is proportionate to the number of possible ways a seizure can spread to channel i in $\mathcal{S}$. Let $\eta_i^{nj}[t] \triangleq \sum_{j \in ne_{\mathcal{S}}} Y_j^{nj}[t]$ be the number of neighbors in $\mathcal{S}$ that are in the seizure state at time t. The probability $Y_i^{nj}[t]$ enters the seizure state at time $t+1$ is $\rho\eta_i^{nj}[t]$, where ρ is the parameter that governs how quickly the seizure spreads.

During the seizure, the probability of seizure offset at any time is ψ. When the switching chain $S^{nj}[t]$ enters an offset state, all EEG channels $Y_i^{nj}[t]$ return to normal activity. This offset is immediately followed by a post-seizure baseline state for the remainder of the recording where no seizure activity is observed.

The onset and offset transitions in the S chain can be expressed as a left to right stochastic transition matrix where $P(S[t] = j \mid S[t-1] = i) = a_{ij}$ is element ij of the matrix A. The S chain transitions through five states, pre-seizure baseline, seizure onset, spreading, seizure offset, and post-seizure baseline. The system remains in the onset and offset states for exactly one timestep before transitioning.

$$A = \begin{bmatrix} 1-\chi & \chi & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1-\psi & \psi & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{5.1}$$

Similarly, the transition factors $P(Y_i[t] \mid Y_i[t-1], S, L, X_{ne_{\mathcal{S}}(i)}[t-1])$ for the Y chains in the CHMM are shown in Table 5-II. Any states not explicitly mentioned in Table 5-II are disallowed in the R-SMMPL and have probability zero. In all states except the spreading state, the states of the Y chains are fully determined by the S chain. In the pre-seizure state, all chains are in state 0, exhibiting normal EEG activity. At onset, the location of the seizure

Table 5-II. Transition factor for the CHMM chains.

S	$Y_i[t-1]$	$Y_i[t]$	L	$Y_{ne_{\mathcal{S}}(i)}[t-1]$	Transition probability
0 - Pre-seizure	0	0	$\cdot$	$\bar{0}$	1
1 - Onset	0	1	$l = i$	$\bar{0}$	1
1 - Onset	0	0	$l \neq i$	$\bar{0}$	1
2 - Spreading	0	0	$l \neq i$	$\bar{x}$	$1 - \rho \sum \bar{x}$
2 - Spreading	0	1	$l \neq i$	$\bar{x}$	$\rho \sum \bar{x}$
2 - Spreading	1	1	$\cdot$	$\cdot$	1
3 - Offset	1	0	$\cdot$	$\cdot$	1
3 - Offset	0	0	$\cdot$	$\cdot$	1
4 - Post-seizure	0	0	$\cdot$	$\bar{0}$	1

onset zone enters the seizure state. During spreading, chains may enter the seizure state with probability proportional to the number of neighbors in the seizure state in the graph $\mathcal{S}$. If a chain enters the seizure state, it remains in the seizure state until seizure offset, when all chains turn off. Chains remain in the offset state until the end of the recording.

5.2.3 CNN Likelihood

Implemented in PyTorch, each CNN contains four convolution and pool layers as shown in Figure 5-1. Convolution layers use eight kernels of five samples with two sample zero padding and LeakyReLU activation. Each convolution uses kernels of five samples with two sample zero padding. Eight kernels are used in each layer with a LeakyReLU activation. Max pooling with a kernel size of two was used to halve the size of the representation at each layer. Following the final pooling stage, softmax classification is performed on the concatenated hidden units. Softmax classification was performed on the concatenated result of the final pooling. All individual CNNs for $P(X_i^{nj}[t] \mid Y_i^{nj}[t])$ were trained for 60 epochs; those for all channels $P(X^{nj}[t] \mid S^{nj}[t])$ were trained for 100 epochs using Adam, batches of 32 samples, a learning rate of 0.5, and cross entropy loss.

By construction, the CNN outputs the posterior probability $P(Y_i^{nj}[t] \mid X_i^{nj}[t])$ and $P(S^{nj}[t] \mid X^{nj}[t])$. Therefore, as in Chapter 4, to obtain the likelihood factor, the discrimina-

tive CNN outputs are rescaled using Bayes rule, e.g.

$$P(X_i^{nj}[t] \mid Y_i^{nj}[t]) \approx \frac{P(Y_i^{nj}[t] \mid C_i[t])P(X_i^{nj}[t])}{\hat{P}(Y_i^{nj}[t])} \propto \frac{P(Y_i^{nj}[t] \mid X_i^{nj}[t])}{\hat{P}(Y_i^{nj}[t])} \,. \tag{5.2}$$

We only require the likelihood up to a constant factor for inference and thus drop the $P(C_i[t])$ term. $P(Y_i^{nj}[t])$ is approximated by the empirical distribution of seizure in the dataset, i.e. $\hat{P}(Y_i^{nj} = 1) = \frac{\text{\# seizure windows}}{\text{\# windows}}$, $\hat{P}(Y_i^{nj} = 0) = 1 - \hat{P}(Y_i^{nj} = 1)$.

5.3 Loopy Belief Propagation for Approximate Inference

The hierarchical and coupled nature of our R-SMMPL renders exact inference intractable. Therefore, we rely on loopy belief propagation [115] for approximate inference. Loopy belief propagation is a general class of algorithms where local marginal beliefs are passed as messages between neighboring random variables. These messages represent the current local beliefs. Through the sum-product algorithm we can find posterior marginal beliefs of random variables in our model. The marginals needed for learning our model are defined below. While loopy belief propagation provides no convergence guarantees, we observe this procedure to yield robust marginals, with little change after further message passing.

Our message passing schedule is detailed in Algorithm 1 and illustrated in Figure 5-5. To initialize the procedure, location messages are passed upward, as illustrated by the purple

Algorithm 1 Approximate inference using loopy belief propagation.

function APPROXIMATE INFERENCE($\mathbf{X}^{nj}, \chi, \psi, \rho, \pi^n$)
  Pass the location variable, L^{nj}, to the $\mathbf{Y}^{nj}$ chains
  for Two Iterations **do**
    Forward-backward algorithm on $\mathbf{S}^{nj}$ chain to update $\gamma_S^{nj}[t]$
    Pass detection messages from $\mathbf{S}^{nj}$ chain down to $\mathbf{Y}^{nj}$ chains
    Approximate forward-backward on $\mathbf{Y}^{nj}$ to update $\gamma_{Y_i}^{nj}[t]$, $\xi_i^{nj}[t]$, and $\phi_i^{nj}[t]$.
    Pass the $\mathbf{Y}^{nj}$ messages upward to the $\mathbf{S}^{nj}$ chain
  end for
  Pass the $\mathbf{Y}^{nj}$ messages to L^{nj} to perform localization and update τ^{nj}
  return $\gamma_S^{nj}[t]$, $\gamma_{Y_i}^{nj}[t]$, $\xi_i^{nj}[t]$, $\phi_i^{nj}[t]$, and τ^{nj}
end function

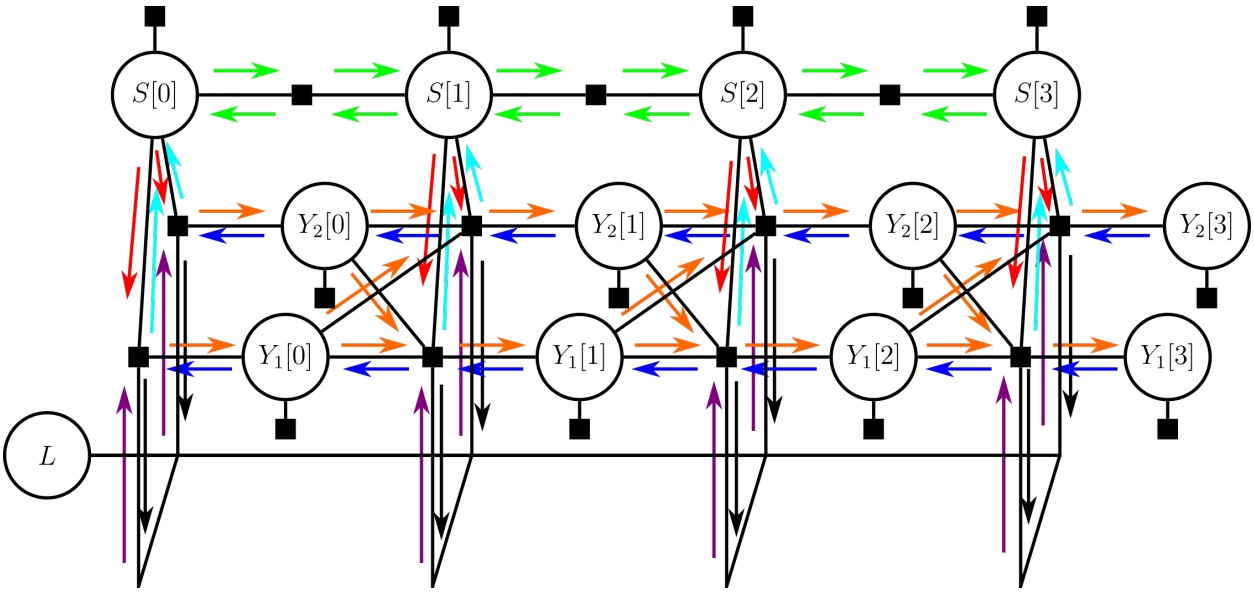


Figure 5-5. Factor graph depicting the R-SMMPL model and inference procedure detailed in Algorithm 1. Observed random variables have been omitted. Green arrows show message passing on the S chain. Red and aqua show interactions between S and $\{X_i\}_{i=1}^{M}$. Orange and blue show forward and backward messages on the CHMM. Purple and black show interactions between the location variable and the CHMM.

arrows. The approximate inference algorithm makes two loops of the following procedure. The green arrows depict the forward-backward algorithm performed on the S chain. This message passing performs rough detection over the course of the seizure recording. Messages are then passed downward to the Y chains via the red arrows. An approximate forward-backward algorithm is performed over the Y chains. The forward pass can be computed exactly, as messages for timestep $t+1$ decouple given timestep t. However, we perform the backward pass as if all chains are independent. This avoids intractability due to coupling. The messages from the Y chain are then passed upward to S and the procedure repeats. After two repeats of this procedure, messages from the Y chain are passed to L to arrive at a final localization and the approximate inference procedure terminates.

Here, we partially illustrate the message passing procedure and posterior inference in the S chain. Inference in the Y chain is analogous. Figure 5-6 shows the messages in the S chain. Figure 5-6 (a) shows messages passed towards $S[1]$, Figure 5-6 (b) shows messages passed away from $S[1]$, and (c) shows the messages passed to the connected variables $S[1]$ and $S[2]$.

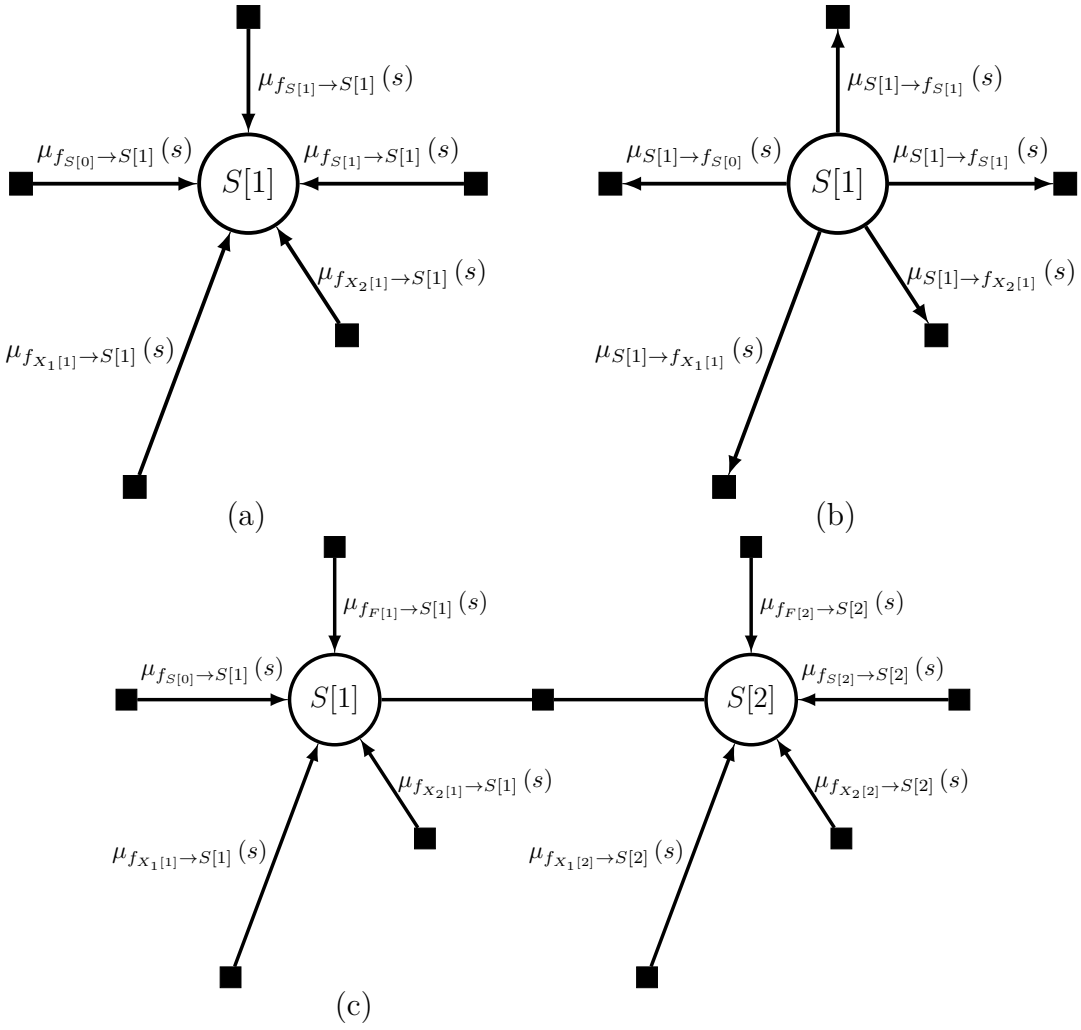


Figure 5-6. Messages on S chain. (a) shows the messages passed towards the variable $S[1]$. (b) shows the messages passed away from the variable $S[1]$. (c) shows the messages passed towards the pair of variables $S[1]$ and $S[2]$ used for pairwise inference.

The forward message from $S[1]$ can be computed by

$$\mu_{S[1] \to f_{S[1]}}(s) = \mu_{f_{F[1]} \to S[1]}(s)\, \mu_{f_{S[0]} \to S[1]}(s)\, \mu_{f_{Y_1[1]} \to S[1]}(s)\, \mu_{f_{Y_2[1]} \to S[1]}(s) \quad . \tag{5.3}$$

Continuing in the forward direction, the message from the factor $f_{S[1]}$ to $S[2]$ is

$$\mu_{f_{S[1]} \to S[2]}(s') = \sum_{s} \mu_{S[1] \to f_{S[1]}}(s)\, f_{S[1]}(s, s') \tag{5.4}$$

where $f_{S[1]}(s, s')$ is the conditional distribution $P(S[2] = s' \mid S[1] = s)$. The backward messages are computed in an analogous fashion.

Posterior inference can be performed in the singleton S variables by multiplying all incoming factors together and normalizing.

$$\begin{aligned} P(S[t]=s) \propto & \mu_{f_{F[1]}\to S[1]}(s)\,\mu_{f_{S[0]}\to S[1]}(s)\,\mu_{f_{Y_1[1]}\to S[1]}(s) \\ & \mu_{f_{Y_2[1]}\to S[1]}(s)\,\mu_{f_{S[1]}\to S[1]}(s) \end{aligned} \tag{5.5}$$

Similarly, inference can be performed for the pairwise marginal $P(S[1]=s, S[2]=s')$.

$$\begin{aligned} P(S[1]=s, S[2]=s') \propto & \mu_{f_{F[1]}\to S[1]}(s)\,\mu_{f_{S[0]}\to S[1]}(s) \\ & \mu_{f_{Y_1[1]}\to S[1]}(s)\,\mu_{f_{Y_2[1]}\to S[1]}(s) \\ & f_{S[1]}(s,s') \\ & \mu_{f_{F[2]}\to S[2]}(s')\,\mu_{f_{Y_1[2]}\to S[2]}(s') \\ & \mu_{f_{Y_2[2]}\to S[2]}(s')\,\mu_{f_{S[2]}\to S[2]}(s') \end{aligned} \tag{5.6}$$

5.4 Learning with the Expectation-Maximization Algorithm

We use an approximate expectation-maximization (EM) algorithm for fitting the R-SMMPL to data. Our model contains three unknown transition parameters: the seizure onset probability χ, the offset probability ψ, and the spreading rate ρ. In addition we learn the onset distribution π^n for each patient. In the E-step of this algorithm, we form a surrogate likelihood function using the expected marginal posteriors of our model based on the current parameter settings. We use the loopy belief propagation procedure described above to compute the required marginal posteriors. In the M-step, we maximize this surrogate likelihood with respect to the parameters of our model. We alternate between the E-step and the M-step until the surrogate likelihood function converges.

5.4.1 E-Step

Let θ denote the set of model parameters for the R-SMMPL and $\theta^{(k)}$ denote the settings of the parameters at iteration k of the EM algorithm. Let $E_{\theta^{(k)}}$ be an expectation under the

parameter settings $\theta^{(k)}$. We form the surrogate likelihood function to begin the EM algorithm.

$$\begin{aligned} Q(\theta \mid \theta^{(k)}) =& E_{\theta^{(k)}}\left[logL(\theta; L, S, Y, X) \mid X\right] \\ =& E_{\theta^{(k)}}\left[\log P(L; \pi) \mid X\right] \\ &+ E_{\theta^{(k)}}\left[\log P(S; \chi, \psi) \mid X\right] \\ &+ E_{\theta^{(k)}}\left[\log P(Y \mid L, S; \rho) \mid X\right] \\ &+ E_{\theta^{(k)}}\left[\log P(X \mid Y) \mid X\right] \\ &+ E_{\theta^{(k)}}\left[\log P(X \mid Y) \mid X\right] \end{aligned} \tag{5.7}$$

Finally, using the loopy belief propagation procedure described in Section 5.3 approximate singleton and pairwise marginal distributions are computed for each variable

5.4.2 M-Step

In the M-step we maximize the surrogate likelihood function $Q(\theta \mid \theta^{(k)})$ with respect to θ. To do this we take the first derivative of $Q(\theta \mid \theta^{(k)})$ with respect to the parameter of interest and set it to zero. Here we show the updates computed using this procedure.

5.4.2.1 Patient-Wise Location Distribution Parameter Update

Here we retain the superscripts nj as the location parameters in the R-SMMPL are patient specific. Here n will be used to a patient and j will be used to denote a recording belonging to patient n. J^n is the total number of recordings for patient n. Let $\tau^{nj}(i) \triangleq E_{\theta^{(k)}}\left[\mathbf{1}\left(L^{nj} = i\right) \mid X\right]$ be the set of location posteriors for patient n for $j \in \{1, 2, \ldots, J^n\}$. Noting that

$$E_{\theta^{(k)}}\left[\log P(L) \mid X\right] = \sum_{j=1}^{J^n} \sum_{i=1}^{M} \tau^{nj}(i) \log \pi(i) \tag{5.8}$$

we see that

$$\frac{\partial}{\partial \pi(i)} Q(\theta \mid \theta^{(k)}) = \sum_{j=1}^{J^n} \frac{\tau^{nj}(i)}{\pi(i)} \quad . \tag{5.9}$$

Setting this expression to zero and imposing that $\sum_{i=1}^{M} \pi(i) = 1$ we see that the solution takes the form $\pi(i) \propto \sum_{j=1}^{J^n} \tau^{nj}(i)$.

5.4.2.2 Regime-Switching Transition Parameter Update

Here we show the transition updates for χ and ψ. Let

$$\xi_S[t](l,m) \triangleq E_{\theta^{(k)}}\left[\mathbf{1}\left(S[t]=m, S[t-1]=l\right) \mid X\right] \tag{5.10}$$

be the pairwise posterior marginals of the S chain and

$$\gamma_S[t](l) \triangleq E_{\theta^{(k)}}\left[\mathbf{1}\left(S[t]=l\right) \mid X\right] \tag{5.11}$$

be the singleton posterior marginals.

$$\begin{aligned} E_{\theta^{(k)}}\left[\log P(S) \mid X\right] = &\sum_{t=1}^{T} \xi_S[t](0,0)\log(1-\chi) + \xi_S[t](0,1)\log\chi \\ &+ \xi_S[t](2,2)\log(1-\psi) + \xi_S[t](2,3)\log\psi \end{aligned} \tag{5.12}$$

Taking first derivatives with respect to χ and ψ, we have

$$\frac{\partial}{\partial\chi}Q(\theta \mid \theta^{(k)}) = \sum_{t=1}^{T}\left(\frac{1}{\chi}\xi_S[t](0,1) - \frac{1}{1-\chi}\xi_S[t](0,0)\right) \tag{5.13}$$

and

$$\frac{\partial}{\partial\psi}Q(\theta \mid \theta^{(k)}) = \sum_{t=1}^{T}\left(\frac{1}{\psi}\xi_S[t](2,3) - \frac{1}{1-\psi}\xi_S[t](2,2)\right) \tag{5.14}$$

By setting these derivatives for zero and solving for ξ and ψ, we arrive at the update equations for these parameters.

$$\chi = \frac{\sum_{t=1}^{T}\xi_S[t](0,1)}{\sum_{t=1}^{T}\left(\xi_S[t](0,0)+\xi_S[t](0,1)\right)} = \frac{1}{\sum_{t=1}^{T}\gamma_S[t](0)} \tag{5.15}$$

$$\psi = \frac{\sum_{t=1}^{T}\xi_S[t](2,3)}{\sum_{t=1}^{T}\left(\xi_S[t](2,2)+\xi_S[t](2,3)\right)} = \frac{1}{\sum_{t=1}^{T}\gamma_S[t](2)} \tag{5.16}$$

5.4.2.3 Seizure Spreading Parameter Update

The parameter ρ controls the speed of seizure spreading. Let

$$\xi_{Y_i}[t](l,m) \triangleq E_{\theta^{(k)}}\left[\mathbf{1}\left(Y_i[t]=m, Y_k[t-1]=l, S[t]=2\right) \mid X\right] \tag{5.17}$$

be the pairwise posterior marginals of the Y_i chain. Let the sum of expected neighbors in $\mathcal{S}$ be $\phi_i[t] \triangleq E_{\theta^{(k)}}\left[\sum_{j \in ne_{\mathcal{S}}(i)} Y_j[t] \mid X\right]$. We approximate the expected log likelihood with respect to the X chains by the expression

$$\begin{aligned} E_{\theta^{(k)}}\left[\log P(Y \mid L, S) \mid X\right] \approx & \sum_{i=1}^{M}\sum_{t=1}^{T} \xi_{Y_i}[t](0,1)\log\left(\rho\phi_i[t-1]\right) \\ & + \xi_{Y_i}[t](0,0)\log\left((1-\rho)\,\phi_i[t-1]\right) \end{aligned} \tag{5.18}$$

where here we have approximated the expectation by taking the expectation by considering $\xi_{Y_i}[t]$ and $\phi_i[t]$ independent and propagating the expectation inside the logarithm. Taking the derivative with respect to ρ we see that

$$\begin{aligned} \frac{\partial}{\partial\rho}Q(\theta \mid \theta^{(k)}) &\approx \sum_{i=1}^{M}\sum_{t=1}^{T}\left(\frac{1}{\rho}\xi_{Y_i}[t](0,1) - \frac{\phi_i[t]}{1-\rho\phi_i[t]}\xi_{Y_i}[t](0,0)\right) = 0 \\ 0 &= \rho\sum_{i=1}^{M}\sum_{t=1}^{T}\phi_i[t]\xi_{Y_i}[t](0,0) + \rho\sum_{i=1}^{M}\sum_{t=1}^{T}\phi_i[t]\xi_{Y_i}[t](0,1) \\ &\quad - \sum_{i=1}^{M}\sum_{t=1}^{T}\xi_{Y_i}[t](0,1) \\ \rho &\approx \frac{\sum_{i=1}^{M}\sum_{t=1}^{T}\xi_{Y_i}[t](0,1)}{\sum_{i=1}^{N}\sum_{t=1}^{T}\phi_i[t]\xi_{Y_i}[t](0,0) + \sum_{i=1}^{M}\sum_{t=1}^{T}\phi_i[t]\xi_{Y_i}[t](0,1)} \\ \rho &\approx \frac{\sum_{i=1}^{M}\sum_{t=1}^{T}\xi_{Y_i}[t](0,1)}{\sum_{i=1}^{N}\sum_{t=1}^{T}\phi_i[t]\gamma_{Y_i}[t](0)} \quad . \end{aligned} \tag{5.19}$$

5.5 Experimental Results

5.5.1 CHB Dataset

The first dataset used for evaluation was the CHB datset [108]. We selected 185 seizures from 24 patients for this experiment. Clinical annotations for CHB include onset and offset times. Again, the type of seizure, general or focal, and potential onset localization are not provided, making this dataset only appropriate for detection evaluation. Recordings in this dataset were made at 256 Hz in a longitudinal montage using the standard 10/20 electrode placement system [15].

5.5.2 JHH Dataset

In addition, we evaluate the R-SMMPL on patients from the JHH dataset. Expert clinical annotations from this hospital include rough onset and offset times as well as consensus of rough onset zone localizations, allowing us to evaluate both detection and localization. The dataset includes 88 seizure recordings from 15 patients. Recordings were sampled at 200 Hz using 10/20 electrode placement.

5.5.3 Preprocessing

We extracted seizure recordings with up to 10 minutes of baseline before and after the seizure annotations. Channels were normalized to mean zero and variance one. High- and low-pass filters were applied at 1.6 Hz and 50 Hz to remove DC offsets and noise. A notch filter at 60 Hz was applied to remove any possible power line contamination. One second windows with 250 ms overlap were extracted from all EEG channels. Test and train sets were separated using leave one patient out cross validation. Unlike studies which train patient-specific detectors, our evaluation focuses on generalizability to unseen patients.

5.5.4 Baseline Comparisons

We compare detection accuracies to the discriminative CNNs trained on the individual EEG channels (I-CNN) and those trained on all channels (S-CNN). These baselines let us assess the effect of the propagation model in fusing information across time and channels. We also compare to the CHMM model in [7] using the same I-CNN for likelihood scoring. The CHMM was shown to outperform standard machine learning classifiers in [7].

5.5.5 Seizure Detection

Tables 5-III and 5-IV reports the detection performance of our R-SMMPL and baseline methods for the JHH and CHB datasets, respectively. We evaluate each algorithm's performance in terms of true positive rate (TPR), true negative rate (TNR), area under the curve (AUC),

Table 5-III. Detection results for the JHH dataset

Trial	TPR	TNR	AUC	P	F1
R-SMMPL	**0.62**	0.84	0.84	**0.65**	**0.53**
CHMM	0.46	**0.96**	**0.86**	**0.65**	0.51
S-CNN	0.34	0.92	0.76	0.41	0.32
I-CNN	0.28	0.92	0.77	0.33	0.27

Table 5-IV. Detection results for the CHB dataset

Trial	TPR	TNR	AUC	P	F1
R-SMMPL	**0.67**	0.94	**0.86**	**0.58**	**0.58**
CHMM	0.59	**0.96**	0.85	0.57	0.54
S-CNN	0.48	0.95	0.84	0.48	0.44
I-CNN	0.30	0.95	0.82	0.34	0.29

precision (P), and F1 score on a frame-wise basis. Performance was evaluated based on how well the methods detected the entire seizure interval. This evaluation is more stringent than prior work, which flags a single correct detection.

The R-SMMPL outperforms all baselines in TPR, P, and F1 scores. We observe that higher TPR comes at the cost of more false positives, reflected by lower TNR. Both the R-SMMPL and CHMM outperform the CNN baselines, illustrating the positive effect of data fusion through the use of spatio-temporal models. The main difference is that the CHMM provides localization information only via heuristic analysis, whereas the R-SMMPL provides it automatically.

5.5.6 Localization Results

We evaluate the localizing ability of our model when provided with a rough seizure onset time. We stipulate that the switching variable should remain in pre-seizure baseline until the clinician annotated onset. The R-SMMPL is free to switch on any time after this point.

Figure 5-7 shows the estimated location distribution π^n for each patient, along with the clinically diagnosed onset location. In this figure, red regions represent areas our model

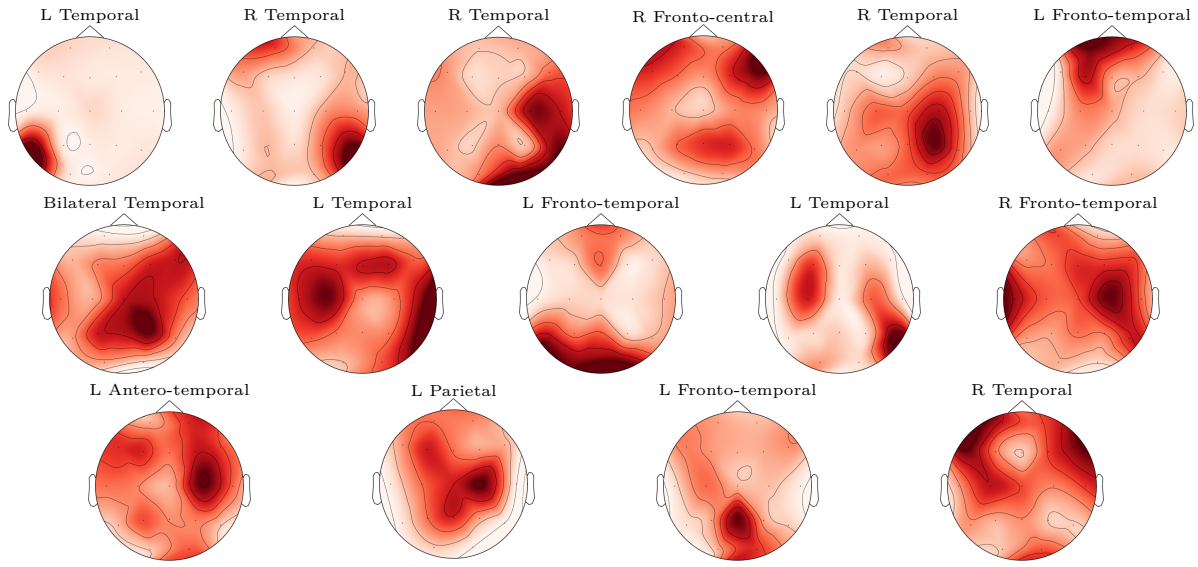


Figure 5-7. Localization results from the JHH dataset. Posterior distributions over onset locations for each patient are shown with clinician provided onset diagnoses above.

assigns high probability for the onset location. In the top row we show cases where our algorithm reported a primary mode in agreement with clinical consensus of the seizure onset zone. The second row shows cases in which the secondary modes agrees with the clinical annotations. By pooling seizure localization information across all of each patient's recordings, our model identifies likely seizure onset zones in agreement with clinical consensus in 11 of 15 patients. The R-SMMPL misidentifies the seizure onset location in just four patients. In summary, not only does the R-SMMPL *automatically* detect and track the seizure, but also leverages multiple seizure presentations to create an onset zone hypothesis for each patient. These results demonstrate the promise of R-SMMPL for clinical evaluation of epilepsy.

5.6 Conclusion

In this chapter we have presented R-SMMPL, the first unified framework that provides clinically relevant detection and localization information from scalp EEG. R-SMMPL combines a probabilistic graphical model of seizure propagation with deep learning for data driven likelihood scoring. We derive an inference and learning procedure for the model and demonstrate its detection and localization abilities on wholly unseen patients, mirroring

clinical conditions. Our methodology for automatic seizure onset zone localization by tracking seizure propagation is the first of its kind.

As in the hybrid CHMM-CNN, discriminative CNNs must be trained prior to learning the R-SMMPL graphical model. In practice, this procedure yields performance gains over previous graphical modeling approaches but requires learning to be decoupled into two separate problems. The performance of these hybrid models depends critically on the quality of the neural network likelihoods, as these likelihoods remain static during the learning of the graphical modeling component. In the second half of this thesis, end-to-end neural network approaches for seizure detection, localization, and tracking will be explored. These methods forgo the advantages of graphical modeling, namely the ability to tightly relate the behavior of the model to hypothesized phenomena in the data, in order to learn models that capture these phenomena in an end-to-end fashion. These end-to-end neural network approaches allow representations of the data and the underlying dynamics of seizure evolution to be learned concurrently.

Chapter 6

SZTrack: End-to-End Seizure Tracking and Localization Using Deep Learning

6.1 Introduction

In the first half of this thesis, we explored graphical modeling approaches for seizure detection and localization. Using hand designed models, seizure onset and propagation hypotheses were incorporated directly into the structure of the models. Leveraging the power of graphical modeling, we were able to restrict the R-SMMPL and CHMM to recognize only the desired onset and propagation of seizure activity. However, this structured approach came at the cost of end-to-end training. While seen to outperform methods from feature engineering, neural likelihood functions were required to be trained prior to graphical model learning. Thus the ultimate performance of the model was constrained by the limits of each individual training step.

For the remainder of the thesis, we explore end-to-end neural network models for seizure detection and localization. Through hybrid convolutional and recurrent structures, feature extraction and temporal analysis are unified in a single neural network architecture. As shown in the Appendix, these CNN-RNN combinations can outperform alternative architectures for seizure detection. Though these networks can be trained to predict seizure activity directly

from the EEG signal, their black box type structures makes incorporating hypothesized seizure spreading phenomena much more challenging than in the case of graphical modeling approaches. As such, novel architectures and training techniques are developed to allow for the identification clinically relevant ictal patterns.

In this chapter, we combine the problems of detection and localization to identify the onset and propagation of electrode-level seizure activity in clinical EEG data. We introduce SZTrack, the first end-to-end network for multichannel seizure activity tracking. SZTrack uses a combined convolutional and recurrent approach to perform classification of seizure activity in individual EEG electrodes at timescales of 1 second, thus generating predictive maps of seizure activity at each time-step. While the architecture operates on each EEG electrode individually, we propose two novel aggregation techniques during training to leverage multichannel phenomena in the EEG data. Our first aggregation technique is to pool the channel-wise classifications into a single patient-wise seizure detection. This strategy allows us to train the network using standard clinical annotations of the seizure onset and offset. It also accommodates the fact that seizure activity may be present in only a subset of the EEG electrodes at a given time. Our second aggregation technique is to combine the *onset information* across channels in anterior vs. posterior head regions and right vs. left hemispheres into a single SOZ prediction. Once again, this strategy allows us to train our channel-wise architecture based on coarse SOZ annotations provided during clinical review. We evaluate SZTrack on two clinical EEG datasets acquired at the Johns Hopkins Hospital and the University of Wisconsin Madison. We demonstrate that SZTrack achieves comparable seizure detection performance to state-of-the-art deep learning approaches. In addition, it can reliably localize the SOZ in a leave-on-patient-out cross validation setting. Finally, SZTrack shows promising cross-site generalization between the two datasets, which provides further evidence of its clinical utility.

6.1.1 Prior Work in Deep Multichannel EEG Analysis

In recent years, Graph Convolutional Networks (GCNs) have become popular for multichannel analysis of EEG data. Broadly, GCNs extend the traditional convolutional architectures, which operate on a regular grid, to arbitrary graphs [116, 117]. Citing the network structure of the brain, GCN approaches in epilepsy encode the underlying connectivity of the brain, for example through spatial proximity or diffusion MRI pathways, directly into the filtering operations of the network. In the work of [118], spectral features derived from the Fast Fourier Transform are analyzed using GCNs to classify 10 second windows of multi-channel EEG as either containing or not containing seizure activity. Along the same lines, the work of [119] uses temporal GCNs to detect the presence of seizure activity in long (96 second) sequences. Going one step further, the authors of [120] learn subject-specific graphs for (temporal) seizure prediction using intracranial EEG. Similarly, the work of [121] used this combination to the problem of emotion recognition from EEG, and the work of [122] apply a spatio-temporal GCN [123] to a Brain-Computer Interface (BCI) motor imagery task.

6.2 Methods

Figure 6-1 illustrates our SZTrack architecture. We first extract a hidden representation at the electrode level by applying a 1D CNN encoder to each one-second window of the time series (left). The encoding sequence for each electrode is passed through a Bidirectional Long Short-Term Memory (BLSTM) unit to determine channel-wise seizure activity. As indicated in Figure 6-1, the CNN and BLSTM parameters are shared across EEG channels, thus providing a compact deep network architecture. The following subsections describe the SZTrack components as well as the aggregation strategy used to train SZTrack for electrode level predictions from seizure onset, offset, and coarse localizations. All models were implemented in PyTorch 1.5.1. Our code is publicly available for download at `https://engineering.jhu.edu/nsa/links/`.

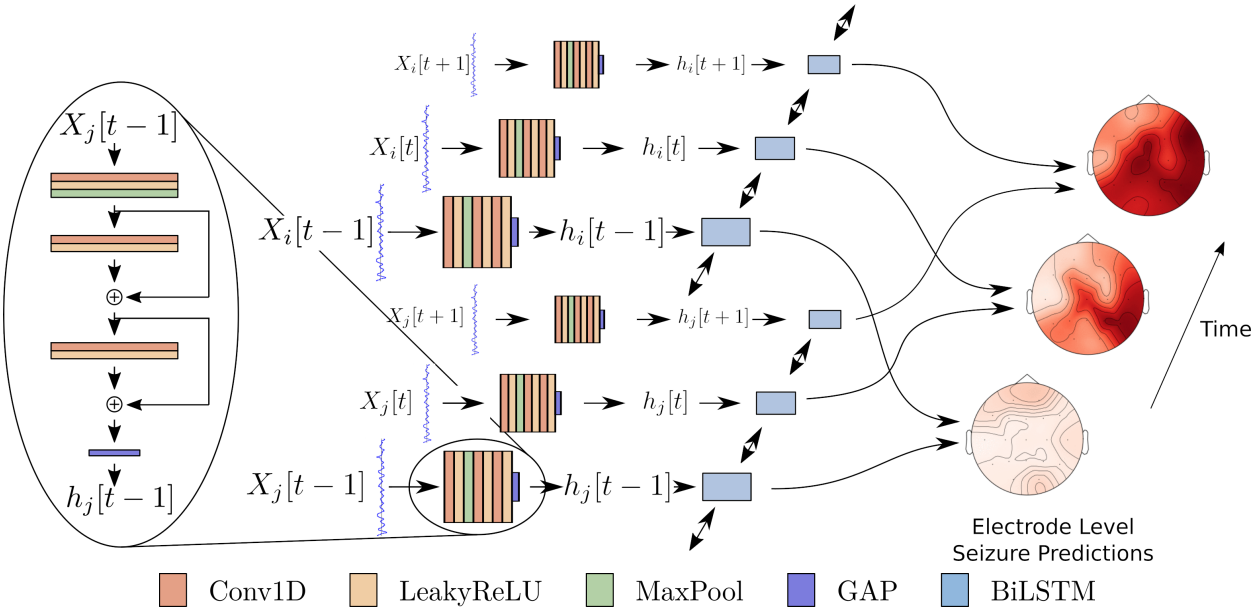


Figure 6-1. SZTrack architecture. Individual EEG electrode signals are fed through a 1D CNN (left). The sequences of representations are fed through the BLSTM layer and then classified for seizure activity in each electrode.

6.2.1 CNN Encoding to Capture Instantaneous Phenomena

The CNN encoder (Figure 6-1, left) extracts feature representations directly from one-second windows of each EEG electrode signal. Let $X_i[t]$ be the signal in EEG electrode i during window t and $\mathbf{X}$ represent the combined signals for the entire recording. The signal $X_i[t]$ is passed through a 3 layer 1D CNN encoder to generate a hidden representation $h_i[t]$. The first layer consists of 20 kernels (length = 7 samples and padding = 3 samples), followed by a LeakyReLU nonlinearity [124], Max Pooling (kernel = 2 samples) and Batch Normalization [125]. The second and third layers use 20 kernels (length = 3 samples and padding = 1 sample), followed by a LeakyReLU nonlinearity and Batch Normalization. Residual connections are added in the second and third layers to ensure a smooth flow of gradient information [75]. Finally, we apply global average pooling resulting in a length 20 hidden representation $h_i[t]$ for each electrode. CNN parameters are shared for all electrodes to ensure a consistent feature representation.

6.2.2 Seizure Tracking via Recurrent Neural Networks

The working hypothesis in focal epilepsy is that a seizure originates from a discrete SOZ and spreads over time to involve other areas of the brain [126]. This spreading pattern is unique across patients, occuring at different time scales and encompassing different spatial extents [127]. We capture this temporal evolution using a BLSTM (Figure 6-1, right), which captures both long-term and short-term dependencies [128]. Formally, the CNN encodings $h_i[t]$ are passed through a BLSTM layer of 40 hidden units. This comparatively large size provides the representational flexibility to track the seizure evolution on longer (i.e., minute-level) time scales. We have tied the BLSTM weights across the electrodes to prevent SZTrack from biasing its predictions towards certain areas of the scalp. Let $o_i[t]$ be the output of the BLSTM for electrode i at time t. The channel-wise seizure prediction $Y_i[t]$ in EEG electrode i and time window t is made via a simple softmax assignment, i.e., $P(\hat{Y}_i[t] \mid \mathbf{X}) = \text{softmax}(W^{\text{T}} o_i[t] + b)$.

6.2.3 Max Pooling for Global Seizure Prediction

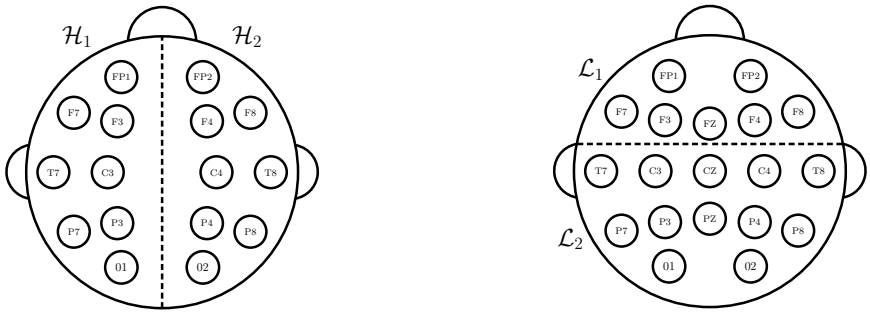
Although SZTrack is designed to *track* the temporal evolution of seizure activity, we only have access to coarse seizure onset and offset times for training. Therefore, we develop a max-pooling strategy to aggregate the electrode level predictions $\hat{Y}_i[t]$ into recording-level predictions $\hat{Y}[t]$ for each one-second window t. Formally, the global prediction $Y[t]$ at window t is predicted as the maximum predicted probability of seizure in any individual channel, i.e.,

$$P(\hat{Y}[t] = 1 \mid \mathbf{X}) = \max_i P(\hat{Y}_i[t] = 1 \mid \mathbf{X}). \tag{6.1}$$

Effectively, when one channel enters the seizure state, the network registers a seizure, accounting for the fact that the activity may concentrate in a single electrode or subset of electrodes. This flexibility allows SZTrack to learn seizure spreading patterns at the electrode resolution with only onset/offset training labels.

6.2.4 Lateralization and Anterior vs. Posterior Classification

Similar to the coarse temporal information, our EEG datasets contain only hemisphere and lobe annotations of the seizure onset, such as "left frontal" or "right temporal". Thus, in order to train SZTrack with these labels, we aggregate electrode level seizure predictions $\hat{Y}_i[t]$ according to the two partitions illustrated in Figure 6-2 (a) and (b). In one partition, Figure 6-2 (a), the EEG electrodes are divided into the left and right hemispheres, denoted $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. In the other partition, Figure 6-2 (b), the EEG electrodes are divided into anterior and posterior head regions, denoted $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively. This classification boundary is defined such that the anterior head region coarsely aligns with frontal lobe seizure foci, while posterior head region contains temporal and parietal foci.



(a) Lateralization zones (b) Lobe zones

Figure 6-2. Localization zones and electrode connectivity graph. Partition of EEG electrodes into zones to train our network based on coarse hemisphere (a) and anterior and posterior head regions (b).

To arrive at a hemisphere or anterior and posterior prediction, we combine first order differences in $P(Y[t] = 1 \mid \mathbf{X})$, denoting seizure onset times, with electrode level predictions. Let $\Delta P_{on}[t]$ capture the transition from baseline to seizure at time t. Mathematically,

$$\Delta P_{on}[t] = \max\left(P\left(\hat{Y}[t] = 1 \mid \mathbf{X}\right) - P\left(\hat{Y}[t-1] = 1 \mid \mathbf{X}\right), 0\right) \tag{6.2}$$

As reported in Eq. (6.2), $\Delta P_{on}[t]$ will approach 1 for confident transitions into seizure, and will be 0 if predicted seizures stays the same or decreases. These differences $\Delta P_{on}[t]$ are multiplied by the seizure activity at time t, $P(Y_i[t] \mid \mathbf{X})$, summed, and normalized to create

a channel-level SOZ onset predictions $P(L_i = 1 \mid X)$.

$$P(L_i = 1 \mid X) = \frac{\sum_{t=0}^{T-2} \Delta P_{on}[t] P(\hat{Y}_i[t] = 1 \mid \mathbf{X})}{\sum_{i=1}^{M} \sum_{t=0}^{T-2} \Delta P_{on}[t] P(\hat{Y}_i[t] = 1 \mid \mathbf{X})} \tag{6.3}$$

Effectively, Eq. (6.3) computes the predicted seizure activity in channel i at time t, $P(\hat{Y}_i[t] = 1 \mid \mathbf{X})$, weighted by the onset activity $\Delta P_{on}[t]$ at that time. Notice that this aggregation relies on both accurate temporal onset detection via $\Delta P_{on}[t]$ and spatial electrode prediction via $P(Y_i = 1 \mid \mathbf{X})$ for a correct localization result. These predictions represent the posterior probability map of the SOZ electrode.

During training, the electrode onset scores $P(L_i = 1 \mid \mathbf{X})$ are aggregated according to the regions defined in Figure 6-2 (a) and (b) to create region level onset scores, $\hat{h} = P(\text{Hemi} = j \mid \mathbf{X}) = \sum_{i \in \mathcal{H}_j} (L_i = 1 \mid \mathbf{X})$ and $\hat{l} = P(\text{Region} = j \mid \mathbf{X}) = \sum_{i \in \mathcal{L}_j} (L_i = 1 \mid \mathbf{X})$. These hemisphere and anterior and posterior predictions are trained using cross-entropy loss function using true labels h and l, thus allowing SZTrack to learn electrode-level patterns from coarse clinical annotations.

6.2.5 Validation Strategy

We evaluate the seizure detection and localization performances in separate experiments using leave-one-patient-out cross validation (LOPO-CV). This cross validation strategy mimics a standard clinical review by quantifying how well each method generalizes to unseen patients. In the detection experiment, we consider the temporal overlap between clinically provided seizure labels and the seizure predictions of SZTrack. In the localization experiment, onset weight in predicted by SZTrack in each the localization based divisions in Figure 2 is considered. Our detection and localization experiments are further detailed in the following subsections.

6.2.5.1 Seizure Onset/Offset Detection

We train SZTrack using a cross-entropy loss between the recording-level seizure prediction $P(\hat{Y}[t] = 1 \mid X[t])$ and the clinician annotation of whether not a seizure is occurring at time

window t. To mitigate over-fitting, training is done for 50 epochs with a weight decay of 0.0001 and a batch size of 4. The learning rate is set at 0.01 and reduced by a factor of 0.5 every 20 epochs.

Performance is evaluated at the one-second window level and by correct classification of the seizure period within each EEG recording. We adopt the strategy of [11], in which the detection threshold is calibrated during each LOPO-CV fold to allow 2 minutes of false positive detection per hour on the training data. At the window level, we report sensitivity, specificity, Area Under the Receiver Operating Characteristc (AU-ROC), and Area Under the Precision-Recall curve (AU-PR) without assuming any temporal dependencies. At the seizure level, we first identify continuous intervals that cross the calibrated detection threshold as "predicted seizures".

The end of a seizure interval is typically corrupted by high levels of artifact (e.g., muscle and eye movements from lingering spasms). Thus, seizure offsets are clinically more difficult to identify, and post-seizure EEG is often mis-classified as a continuation of seizure activity. Since clinical evaluation of epilepsy focuses on the seizure onset and evolution behavior, we adopt a strategy that rewards true seizure detection (i.e., maximizing sensitivity) without penalizing the model for continuing post-seizure predictions. Predicted seizure intervals that overlap with annotated seizure activity are considered true positives, while intervals occurring exclusively during baseline are considered false positives. We note that this quantification strategy has been used previously in the seizure detection literature in [7] Predicted seizure intervals that overlap with annotated seizure activity are considered true positives, while intervals occurring exclusively during baseline are considered false positives. We report the seizure level metrics False Positive Rate (FPR), computed as the number of false positives per hour, and sensitivity, computed as the ratio of accurately classified seizures to missed seizures. We also report average latency for true positive detections. These metrics quantify the clinically relevant need for accurate seizure detection (sensitivity) with low latency and a small number of false positive detections.

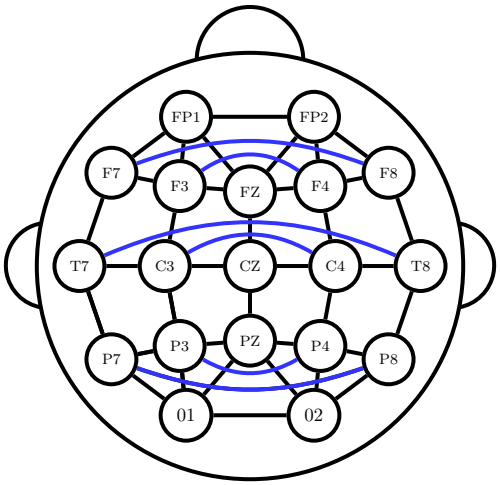


Figure 6-3. Electrode connectivity graph. Electrode connectivity graph used in GCN baselines.

6.2.5.2 Seizure Localization

We evaluate lobe and lateralization accuracy in separate experiments on recordings clipped from 15 seconds prior to 30 seconds after seizure onset. This clipping mitigates the influence of physiological confounds, such as eye and muscle movements, in the challenging localization task. In each case, the loss is a weighted combination of the cross-entropy seizure detection loss:

$$\begin{aligned} \mathcal{L} &= \lambda_{sz} CE_{sz}(\hat{y}, y) + CE_{hemi}(\hat{h}, h), \quad \text{or} \\ \mathcal{L} &= \lambda_{sz} CE_{sz}(\hat{y}, y) + CE_{region}(\hat{l}, l) \end{aligned} \tag{6.4}$$

We initialize SZTrack via the models learned in the corresponding LOPO-CV seizure detection experiment and retrain the network for 50 epochs with the combined loss function in Eq. (6.4) using the same weight decay and learning rate schedule described above. To quantify robustness, we investigate the LOPO-CV performance while sweeping the detection loss weight λ_{sz} from $(0.1 - 1.0)$.

6.2.6 Baseline Models

We compare SZTrack to an ablated model, in which we remove the BLSTM layer. This model, which we call No-BLSTM, allows us to evaluate the benefit of tracking temporal dependencies in the EEG data. We also adapt to recently published deep learning models for

seizure detection that use GCNs for spatial information fusion. In both cases, we use the graph defined in Figure 6-3, which connects neighboring and contralateral EEG electrodes.

The first model is a temporal GCN (TGCN) introduced in [119]. We rely on "Architecture II" from the paper, which achieves the best performance across most of the evaluation metrics in the original work. For comparison with SZTrack, we remove the max pooling along the temporal dimension and the average pooling along the spatial dimension, which allows the TGCN to output predictions at the electrode-level and one-second window-level resolution. The second and third GCN networks are introduced in [118] and apply the propagation rule in [129] along with graph pooling for sequence detection. Here, the Shallow-GCN model uses two GCN layers with 64 and 128 hidden units, respectively, followed by a single linear classification layer. The Deep-GCN model uses five GCN layers with increasing hidden sizes of 16, 16, 32, 64, and finally 128, followed by two linear layers with 30 and 20 hidden units before a final classification layer. Once again, we adapt the networks by removing the graph pooling layer to allow for electrode-level predictions. Unlike the previous methods, the Shallow-GCN and Deep-GCN operate on spectral input features. We construct this 10-dimensional input by extracting the spectral power in 10 equally-spaced frequency

Finally we compare SZTrack against two multichannel CNN baselines and a multichannel CNN-BLSTM baseline that have recently appeared in the seizure detection literature. These models differ from SZTrack in that they *output a single global seizure prediction at every time window*. Hence, these models are *incapable of tracking seizure activity* at the resolution of individual electrodes, which is the goal of SZTrack. The Wei-CNN baseline [101] uses 5 CNN layers followed by 2 linear layers. The CNN-2D baseline operates on the short-time Fourier transform images. Four CNN layers are applied before a final linear layer is applied for classification. The CNN-BLSTM presented in [11] extracts features from the multichannel EEG signal using a CNN before classifying the sequence of features using a BLSTM layer.

We include all models in the our detection experiments. Empirically, all baselines except the CNN-BLSTM output noisy seizure detections, since they are made independently for

each one-second window. Therefore, we smooth the predictions of these baseline models by averaging the outputs over 20 seconds. This smoothing procedure is omitted for SZTrack and the CNN-BLSTM. We assess the localization performance for SZTrack and the No-BLSTM and TCGN baselines. We omit the multichannel architectures, as they cannot output localization information at the window level. Similarly, we omit the Shallow-GCN and Deep-GCN baselines, as they do not include a temporal modeling component to capture seizure onset and evolution. In addition, we adapt the CNN-BLSTM models from our detection experiment to localization by adding a linear classification layer operating on the final hidden state of the BLSTM. We re-train these CNN-BLSTM models using the detection models as a starting point for an additional localization baseline. These models are trained explicitly for localization and λ_{sz} is set to 0 accordingly.

6.3 Results

6.3.1 Clinical EEG Datasets

JHH Dataset: The primary EEG dataset used in this experiment was the JHH dataset. Here we use all 201 seizure recordings obtained from all 34 focal epilepsy patients undergoing presurgical evaluation in the Johns Hopkins Hospital between 2016–2019. Annotations include both seizure onset and offset labels as well as patient level localization annotations, making this dataset appropriate for evaluation in both detection and localization tasks. For our analysis, the EEG recordings have been clipped to include roughly 10 minutes of pre-seizure and post-seizure activity.

UWM Dataset: Our generalization dataset consists of 53 seizure recordings from 15 pediatric patients admitted to University of Wisconsin-Madison (UWM) from February 2018 to December 2019. While smaller, this dataset also contains annotations of seizure onset and offset as well as patient level localization information. As such, we evaluate the generalization of models trained on the JHH dataset when applied to the UWM data. The data was recorded

at 256 Hz using the 10-20 common reference and was resampled to 200 Hz to be consistent with the JHH dataset.

Preprocessing: Following [6], we bandpass each recording between 0.5 to 30 Hz and remove high intensity artifacts by thresholding each recording at two standard deviations from its mean value. The EEG signals were then normalized to have mean zero and variance one. The EEG signals were then normalized to have mean zero and variance one. One second non-overlapping windows were extracted for input to the models. For efficiency, we train the JHH detection models on EEG data containing 2 minutes of pre- and post-seizure activity; however, we evaluate them on the full 20-minute recordings.

6.3.2 Detection Performance

Tables 6-I and 6-II report the window level and seizure level detection results on the JHH dataset, respectively. We observe that the SZTrack and CNN-BLSTM exhibit nearly comparable performance at the window level, with AU-ROCs of 0.895 and 0.899, respectively. This comparison with the highly-optimized multichannel CNN-BLSTM show the ability of our simpler SZTrack model to achieve state-of-the-art seizure detection performance while preserving channel-wise information. Beyond these two models, the CNN-2D achieves the next-best overall performance, as highlighted by the AU-ROC and AU-PRC measures. This performance may be due to the relatively simple architecture, which can leverage the spectral input information. The remaining baselines are clustered together, with the ablated No-BLSTM and the two GCN models (Deep and Shallow) outperforming the recently proposed TGCN and Wei-CNN methods. We note that the TGCN and Wei-CNN baselines rely on larger and more complex architectures, which may result in overfitting to the relatively modest JHH dataset.

Unlike the window-level results, where there is a clear ordering between the methods, the performance is mixed at the level of contiguous seizure detection. For example, SZTrack and the CNN-BLSTM achieve the lowest detection latency at the cost of higher false positive

Table 6-I. Window-level performance on the JHH dataset. Metrics are aggregated across one-second segments of the EEG.

Model	AU-ROC	AUC-PR	Sensitivity	Specificity
SZTrack	0.895 ± 0.112	0.644 ± 0.281	0.593 ± 0.305	0.936 ± 0.065
CNN-BLSTM	0.899 ± 0.085	0.635 ± 0.241	0.590 ± 0.048	0.945 ± 0.048
No-BLSTM	0.797 ± 0.101	0.438 ± 0.224	0.518 ± 0.211	0.883 ± 0.092
TGCN	0.760 ± 0.124	0.485 ± 0.177	0.591 ± 0.183	0.821 ± 0.157
Deep-GCN	0.786 ± 0.0978	0.394 ± 0.218	0.485 ± 0.208	0.887 ± 0.078
Shallow-GCN	0.792 ± 0.097	0.412 ± 0.177	0.488 ± 0.183	0.892 ± 0.157
Wei-CNN	0.764 ± 0.156	0.488 ± 0.280	0.405 ± 0.279	0.921 ± 0.109
CNN-2D	0.824 ± 0.147	0.527 ± 0.247	0.470 ± 0.253	0.921 ± 0.109

Table 6-II. Seizure level performance on the JHH dataset. Results are calculated over the duration of the seizure interval.

Model	FPs/hr	Sensitivity	Latency (s)
SZTrack	13.05	0.865	12.35
CNN-BLSTM	16.46	0.919	10.84
No-BLSTM	8.17	0.894	21.07
TGCN	7.99	0.859	25.41
Deep-GCN	8.05	0.823	23.56
Shallow-GCN	8.77	0.835	27.49
Wei-CNN	7.5	0.77	21.27
CNN-2D	10.2	0.84	25.39

predictions per hour. In contrast, the No-BLSTM, TGCN and static GCNs (Deep and Shallow) make fewer false positive detections but have notably higher latency. In terms of sensitivity, the CNN-BLSTM performs the best at 0.919 with the No-BLSTM model a close second at 0.894. The two GCN models performed comparably with sensitivities of 0.823 (Deep) and 0.835 (Shallow), which is on par with SZTrack. Finally, the Wei-CNN method achieves considerably lower sensitivity than the others, perhaps due to the larger architecture and lack of temporal modeling. Taken together, our seizure detection experiment demonstrates the clinical utility of our simple channel-wise architecture and information fusion strategy.

To assess cross-site generalization, Tables 6-III and 6-IV report the respective window and

Table 6-III. Window level generalization detection results on the UWM dataset. Seizure detection performance when applying the JHH models to data from UWM. We ran a LOPO-CV on UWM to calibrate the seizure versus baseline detection threshold. However, we did not retrain the neural network weights.

Model	AU-ROC	AUC-PR	Sensitivity	Specificity
SZTrack	0.813 ± 0.164	0.380 ± 0.301	0.427 ± 0.288	0.950 ± 0.060
CNN-BLSTM	0.857 ± 0.116	0.393 ± 0.298	0.329 ± 0.291	0.954 ± 0.083
No-BLSTM	0.724 ± 0.213	0.350 ± 0.312	0.287 ± 0.273	0.961 ± 0.073
TGCN	0.691 ± 0.205	0.257 ± 0.240	0.270 ± 0.228	0.894 ± 0.107
Deep-GCN	0.679 ± 0.219	0.285 ± 0.293	0.211 ± 0.206	0.958 ± 0.063
Shallow-GCN	0.699 ± 0.214	0.302 ± 0.299	0.245 ± 0.227	0.962 ± 0.063
Wei-CNN	0.849 ± 0.126	0.406 ± 0.291	0.536 ± 0.332	0.900 ± 0.145
CNN-2D	0.782 ± 0.157	0.382 ± 0.272	0.442 ± 0.221	0.956 ± 0.056

Table 6-IV. Seizure level generalization detection results on the UWM dataset. Seizure detection performance when applying the JHH models to data from UWM. We ran a LOPO-CV on UWM to calibrate the seizure versus baseline detection threshold. However, we did not retrain the neural network weights.

Model	FPs/hr	Sensitivity	Latency (s)
SZTrack	14.05	0.639	5.48
CNN-BLSTM	2.83	0.523	12.28
No-BLSTM	7.48	0.517	12.65
TGCN	15.14	0.613	14.85
Deep-GCN	10.57	0.528	15.66
Shallow-GCN	8.77	0.557	16.39
Wei-CNN	11.11	0.701	5.87
CNN-2D	10.21	0.728	13.89

seizure level detection performance of the JHH models when evaluted on the UWM dataset. In this case, we recalibrate the detection threshold for each of the JHH models (obtained via LOPO-CV) on the UWM dataset, but we do not retrain the model parameters on the new data. Consequently, there is a performance decline across all models when translated from the JHH adult cohort to the UWM pediatric population. Nonetheless, we observe the same general trends. Namely, SZTrack shows comparable performance to the strictly detection based CNN-BLSTM model at the window level with AU-ROCs of 0.813 and 0.857, respectively. At the seizure level, SZTrack exhibits higher sensitivity (0.639) at the cost

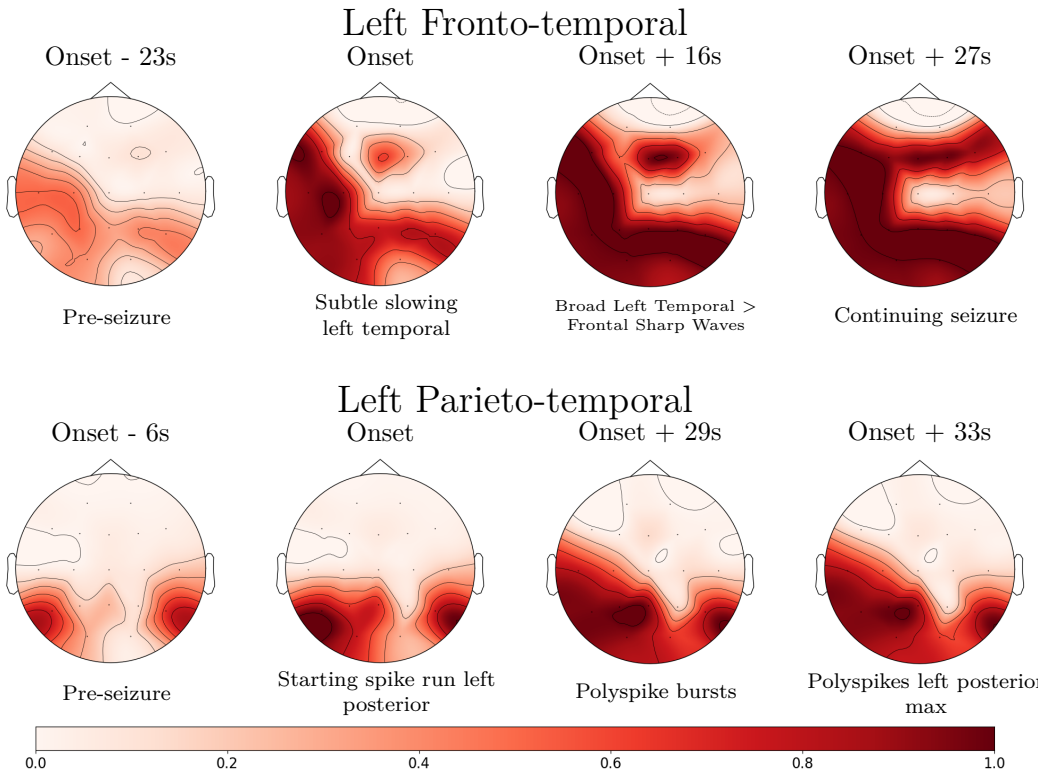


Figure 6-4. Seizure activity tracking. Seizure activity tracking in two JHH patients. Clinical SOZ annotations are given for each patient. Where clinical annotations are provided, images show seizure activity tracking corresponding to annotation times.

of more FPs/hr (14.05) when compared with the CNN-BLSTM, which had sensitivity and false positive rate of 0.523 and 2.83. Surprisingly, the Wei-CNN shows high generalization performance, with an AU-ROC of 0.849, exceeding its AU-ROC of 0.764 in the JHH dataset. This might be linked to the fact that the Wei-CNN was optimized for detection on the publicly available Children's Hospital of Boston (CHB) dataset [45], which also contains pediatric patients. Similarly, the CNN-2D shows a high level of generalization stability across datasets, perhaps indicating the robustness of spectral information.

Figure 6-4 illustrates the seizure tracking output $P(\hat{Y}_i[t] \mid \mathbf{X})$ by SZTrack for two patients from the JHH dataset, as superimposed in red on a topographic scalp plot. These recordings contain annotations of seizure spreading created by epileptologists during clinical workup. Seizure activity maps are provided at the time of annotation to show concordance between annotated seizure activity and SZTrack predictive outputs. As seen, the seizure activity

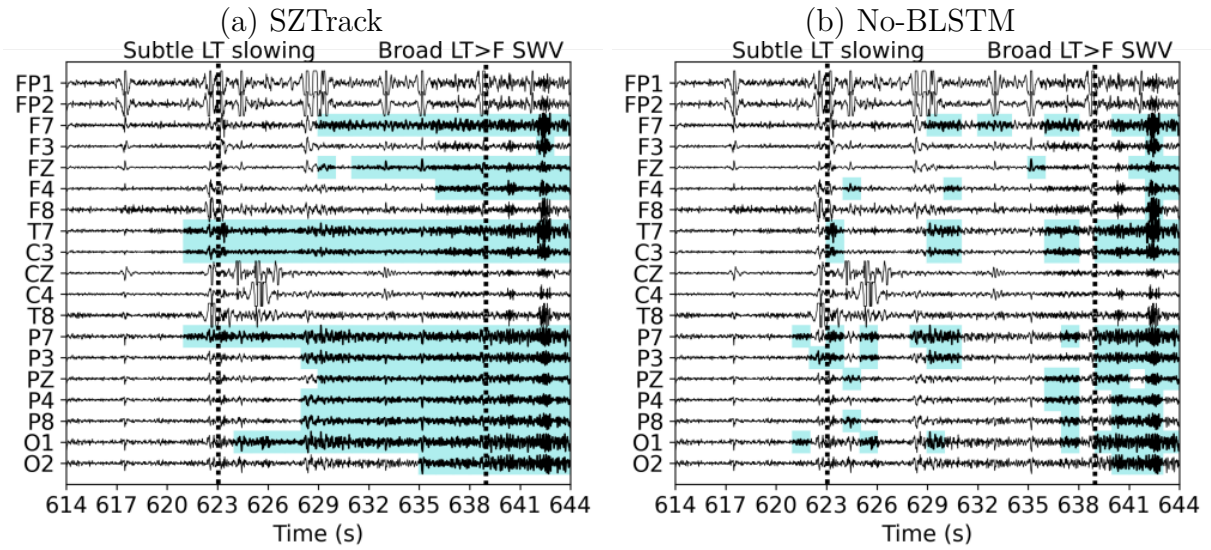
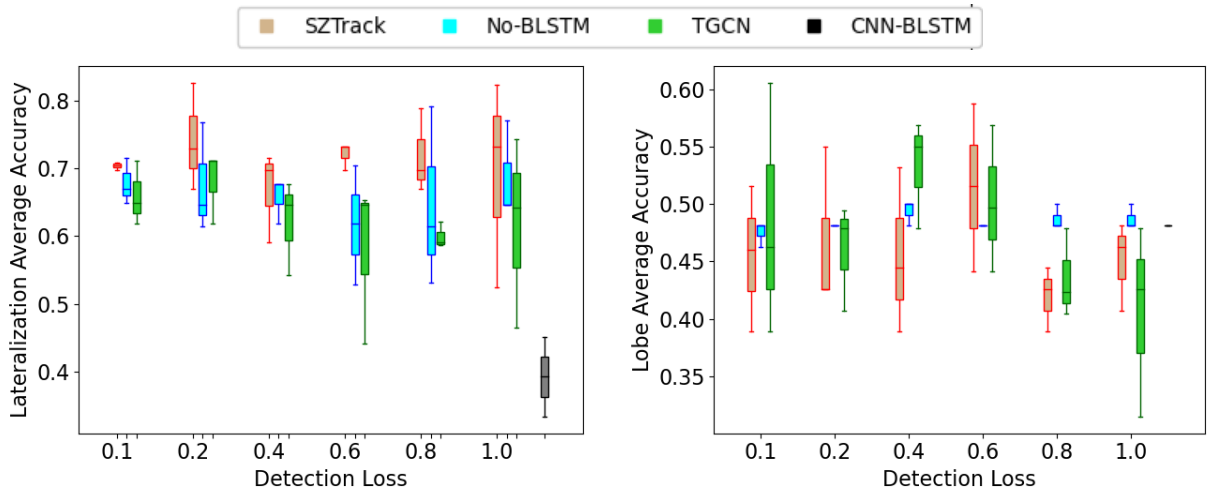


Figure 6-5. SZTrack and No-BLSTM output comparison. Channel-wise predictions for the fronto-temporal seizure shown on the top row of Figure 4 are superimposed on the EEG signal. In (a) SZTrack makes a confident prediction of seizure onset in the temporal channels which spreads to the parietal and frontal areas. In (b) No-BLSTM responds to isolated seizure activity at the onset but does not provide a temporally stable prediction.

automatically learned by SZTrack from the EEG data shows strong agreement with clinically observed spreading patterns during the seizure. We emphasize that due to our LOPO-CV training strategy, SZTrack had no *a priori* knowledge of these patients prior to generating the predictions in Figure 6-4. To the best of our knowledge, this is the first tacking result of its kind reported in the literature.

Figure 6-4 illustrates the model predictions made by SZTrack and the No-BLSTM baseline for the fronto-temporal seizure shown on the top row of Figure 4. We have used the open-source EEG visualization software EPViz to overlay the channel-wise predictions in blue on top of the EEG signals. As seen, SZTrack (a) predicts seizure activity originating in the temporal lobe channels T7 and P7. This prediction agrees with the clinically annotated onset information "subtle slowing left temporal" at 623 seconds in the EEG. Seizure activity quickly spreads to further involve left temporal, parietal, and left frontal electrode channels. Once again this prediction concurs with the clinical note of "Broad Left Temporal > Frontal Sharp Waves" at 639 seconds. In contrast, the No-BLSTM baseline correctly detects the left



(a) Lateralization accuracy (b) Anterior vs. Posterior accuracy

Figure 6-6. Localization sweep results. Average localization accuracy in JHH when varying the weight on the detection loss. Boxplots are shown for the SZTrack, No-BLSTM, and TGCN models. A horizontal dashed line shows performance for the CNN-BLSTM model.

temporal onset but does not provide contiguous predictions. Consequently, it fails to capture the clinically observed spreading pattern.

6.3.3 Localization Performance

Figure 6-6 illustrates the average lateralization and lobe classification performances in JHH dataset as the detection loss weight λ_{sz} is swept from zero to one. Accuracy results from each location class are averaged and boxplots from three separate runs are displayed for SZTrack, No-BLSTM, and the TGCN. The CNN-BLSTM baseline score is shown by the single gray boxplot, as this model was evaluated independently of seizure detection. The lateralization results are striking, as SZTrack uniformly outperforms the No-BLSTM and TGCN baselines, achieving its highest average accuracy of 0.826 at $\lambda_{sz} = 0.6$. The lobe identification task appears more difficult, as there is a universal decline in performance across all methods. In this case, the TGCN slightly outperforms SZTrack, achieving a maximum average lobe detection accuracy of 0.605 at $\lambda_{sz} = 0.1$ versus SZTrack's 0.587 for $\lambda_{sz} = 0.6$. Nonetheless, SZTrack achieves robust detection and lateralization performance, thus illustrating its potential clinical utility. We also note that this result is the first demonstration of end-to-end seizure localization

from scalp EEG reported in the literature. With that said, further prospective analyses are required to evaluate the impact of SZTrack on the current clinical workflow.

For a qualitative evaluation, Figure 6-7 illustrates the LOPO-CV localization results on the JHH dataset for a single hyperparameter setting ($\lambda_{sz} = 0.2$ for lateralization and $\lambda_{sz} = 0.6$ for lobe classification). The seizure onset maps, as denoted by $P(\mathbf{L} \mid \mathbf{X})$, are shown superimposed on head plots in red. Lateralization and lobe images for each patient are displayed on the left and right, respectively, with the expert-determined SOZ provided below. For ease of comparison, we have added small circles to the corner associated with the clinical SOZ. A green circle indicates a concordance between SZTrack and clinical annotations while a red circle indicates disagreement. In 20 of 34 patients, SZTrack identifies both the correct hemisphere and lobe. SZTrack identifies the correct lobe or hemisphere in all of the 14 remaining patients.

As a preliminary study of generalization, we selected a random LOPO-CV fold and applied the trained SZTrack model to the UWM data *with no fine tuning*. Figure 6-8 illustrates localization maps $P(\mathbf{L} \mid \mathbf{X})$ for the hemisphere and lobe identification, as averaged across the seizure recordings for each patient. As seen, SZTrack correctly localizes both partitions in 8 of the 15 patients. In 5 of the patients, SZTrack correctly localizes either hemisphere or lobe. It misses completely in only 2 of the 15 patients. This result suggests that our SZTrack architecture is capturing salient information regarding seizure onset location that generalizes across different epilepsy cohorts.

6.4 Discussion

SZTrack introduces a channel-wise architecture that consists of a CNN encoder, operating on one second windows of the EEG signal, followed by a BLSTM to capture both short- and long-term temporal dependencies. While the SZTrack architecture analyzes each EEG channel individually, our novel training strategy allows the network to learn and predict

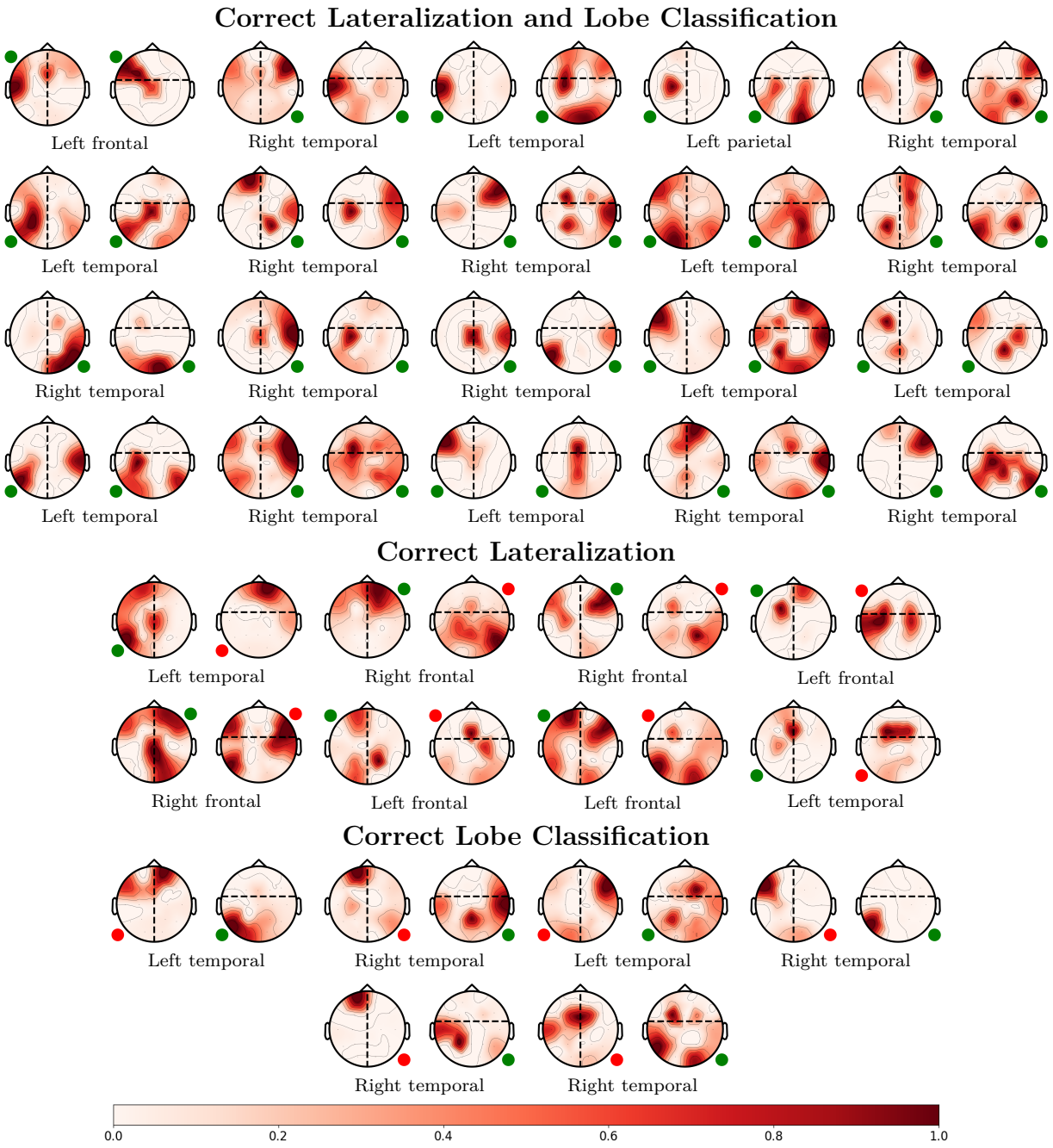


Figure 6-7. Localization results from the JHH dataset. Patient-wise lateralization and lobe classification for SZTrack in JHH. Predicted SOZ locations are superimposed on the head figure in red. The small circle indicates the coarse clinical SOZ annotation, where green indicates concordance with clinical annotations and red circle indicates disagreement. SZTrack correctly localizes both the hemisphere and lobe in 21 of 34 patients. In 12 of 34 patients, SZTrack correctly localizes either hemisphere or lobe; it misses completely in just one patient.

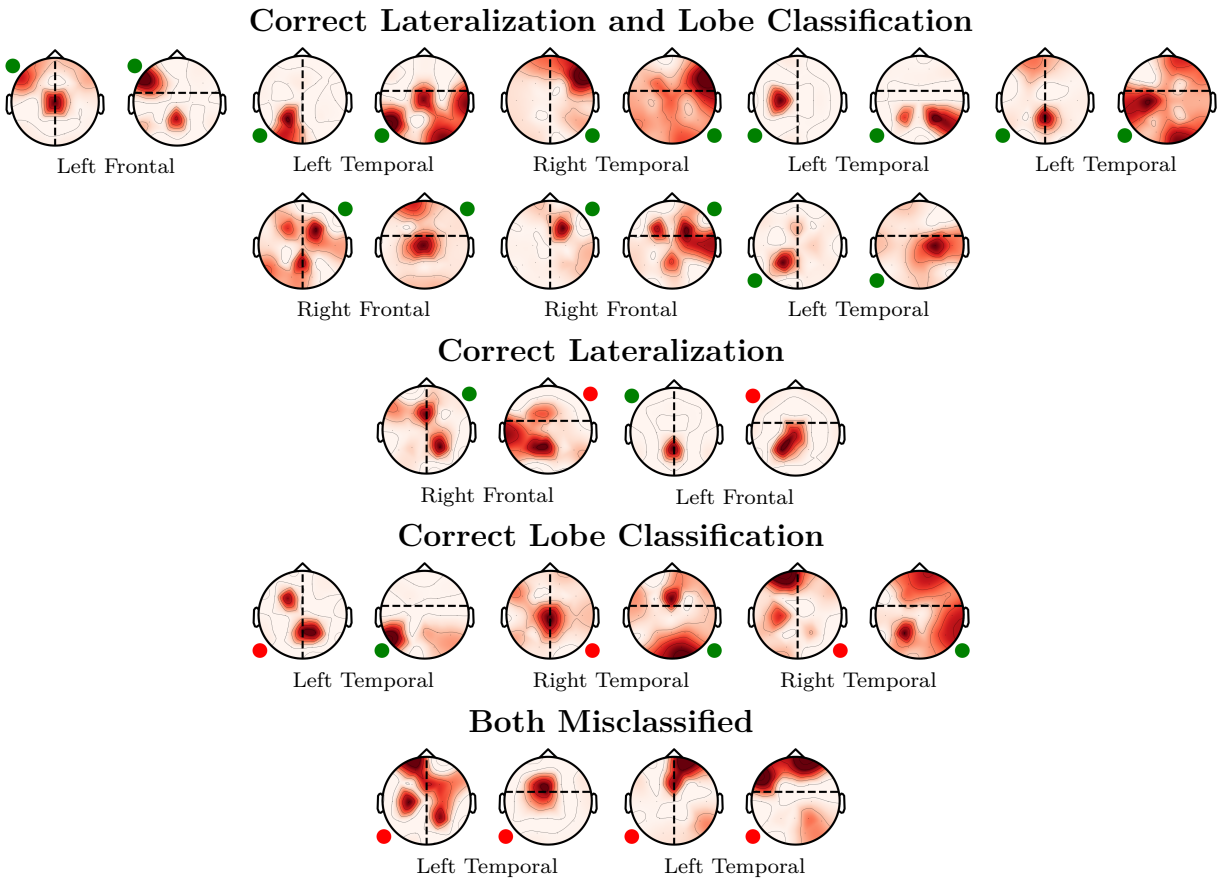


Figure 6-8. UWM dataset generalization results. Lateralization and lobe classification results when applying a SZTrack model trained on JHH to data from UWM. Predicted SOZ locations are shown superimposed on the head figure in red. The small circle indicates the coarse clinical SOZ annotation.

seizure activity based on the multichannel data. These predictions show high concordance with clinician determination of the seizure onset and offset. In addition, we have demonstrated the first end-to-end *seizure localization* results based on multichannel scalp EEG. Excitingly, SZTrack is able to *track* spatiotemporal seizure activity at higher resolution than the clinician annotations used for training.

In terms of seizure detection, SZTrack performs comparably to the benchmark CNN-BLSTM architecture. This result demonstrates that our max-pooling aggregation strategy can account for multichannel phenomena to a similar extent as a larger cross-channel architecture.

Going one step further, Figs. 6-4 and 6-5 demonstrate that SZTrack can learn channel-level seizure onset and propagation, a task that is impossible for the CNN-BLSTM. Specifically, it is not possible to disentangle which EEG channel(s) are driving the CNN-BLSTM prediction at any given time. In Figure 5 the added benefit of the LSTM for temporal seizure activity in the SZTrack architecture can be seen when comparing to model outputs provided by the No-BLSTM baseline. While the No-BLSTM baseline identifies some seizure activity near the annotated onset, it does not produce temporally contiguous seizure predictions after this onset. With the added LSTM layer, SZTrack correctly infers the start and continuing evolution of seizure activity. We stress that the tracking patterns are learned solely from annotations of the seizure onset and offset interval; our dataset does not contain fine-grained information at the level of individual channels.

When compared with GCN baselines, SZTrack exhibits higher seizure detection performance. These GCN models encode hypothesized network structures of the brain directly in their architectures in an attempt to capture “biologically informed" relationships in the data. Empirically, we observe that this approach does not directly lead to increases in detection performance, as SZTrack outperforms these graph based approaches. In fact, the results in Tables 6-I and 6-II suggest that it may be more valuable to incorporate multichannel information during training (e.g., our max-pooling strategy) rather than directly into the network architecture. This is particularly true if there is a mismatch between the assumed graph structure and the actual EEG data.

In addition to detection efficacy, we make a first attempt to perform and validate end-to-end seizure onset localization from scalp EEG. SZTrack outperforms the ablated No-BLSTM, demonstrating the necessity of modeling temporal dependencies for this challenging task. SZTrack also outperforms the TGCN, which relies on both a GCN layer for cross-electrode information sharing and 1D temporal convolutions for time-series modeling. Similar to the detection task, perhaps the lower TGCN performance reflects a mismatch between the assumed graph and the actual data dependencies. The CNN-BLSTM, trained only for seizure

localization, shows the worst performance of all the models in lateralization, and near-chance anterior vs. posterior detection accuracy. This model analyzes all EEG channels concurrently in its architecture, which likely blurs subtle differences indicative of the seizure onset location.

We note a general decrease in anterior vs. posterior classification for all models. We hypothesize that this decrease is partially attributed to class imbalance. Most patients in our datasets have temporal lobe onsets, which greatly reduces the number of examples to learn patterns associated with other onset locations. Interestingly, in Figure 7 we note that in roughly 6 of 8 cases where SZTrack fails to correctly localize the SOZ in the anterior vs. posterior classification task, during the lateralization taskSZTrack places the mode of its SOZ probability in the correct anterior or posterior head region. This indicates that without the confounding effects of the dataset imbalance, SZTrack may learn to correctly classify anterior or posterior head regions even without being explicitly trained to do so.

The coarse division of onset zones into anterior versus posterior may also contribute to the performance decrease, as it does not accommodate seizures originating at the border of this division or complex multi-focal cases. Particularly, we note that electrodes F7 and F8 may be involved in both frontal and temporal lobe onsets [14], while our division scheme necessitates that these electrodes be placed in only one head region. Future work will consider a more comprehensive channel grouping strategy that allows for soft assignments and overlapping classes. In the anterior vs. posterior classification task, we noted more variation in performance across models as the hyperparameter λ_{sz} is swept across its range. While the CNN-BLSTM drastically under-performed in the lateralization task, we note that its performance in anterior vs. posterior classification is on par with SZTrack and the other two baselines, likely due to the general difficulty of this task.

When applied to a generalization dataset, SZTrack shows robustness without the need for retraining. In the detection task, SZTrack maintains a stable AU-ROC across the JHH and UWM datasets. In the localization feasibility study, SZTrack models trained in the JHH dataset correctly identify the SOZ to the hemisphere and anterior vs. posterior level

consistent with models trained and tested in the original dataset. These results demonstrate the robustness of SZTrack to real-world changes in clinical condition.

While we have demonstrated the retrospective clinical utility of SZTrack in detecting, tracking, and localizing seizure activity in two separate scalp EEG datasets, we note that our method has several limitations. While SZTrack achieves high seizure detection, the CNN-BLSTM surpasses SZTrack in performance in the original JHH dataset and UWM generalization dataset. Future work will incorporate a multi-channel component to SZTrack, similar to the CNN-BLSTM, to leverage cross-channel dependencies when making a prediction. Another issue is the relatively poor anterior vs. posterior classification performance. As described above, some issues in anterior vs. posterior classification may stem from our choice of dividing boundary between anterior and posterior SOZs. Specifically, onsets that occur near the boundary (e.g., fronto-temporal SOZ) are likely difficult for SZTrack to disambiguate. This scenario motivates a finer evaluation and training strategy for anterior vs. posterior prediction. In the future, we will explore data augmentation and aggregation techniques to increase the amount of extra-temporal seizure data for model training.

6.5 Conclusion

In this chapter, we have introduced SZTrack, a novel deep neural network architecture for seizure activity tracking in multichannel EEG. Forgoing the structured prediction afforded by graphical models, a neural network architecture is trained to identify seizure activity. Through cross electrode parameter sharing and novel predictive output aggregation, SZTrack achieves comparable seizure detection performance as deep models that use GCNs for direct information sharing between EEG electrodes. In addition, our aggregation techniques allow SZTrack to predict electrode level seizure activity from coarser clinical annotations. We also evaluate SZTrack on the difficult task of seizure localization, where it achieves high hemisphere and above-chance anterior vs. posterior region classification accuracy. The localization performance also generalizes across sites with no fine tuning. SZTrack represents the first

end-to-end neural network for seizure tracking, detection, and localization, establishing an important benchmark for the field.

As SZTrack operates on each channel of the EEG signal individually, the network is incapable of truly incorporating cross-channel information at test time. We circumvent this limitation through our cross channel seizure detection aggregation during training. In analogy to the performance gains due to higher order information fusion seen in the R-SMMPL, we hypothesize that similar multi-scale information sharing between global and electrode level signals could improve localization. Furthermore, while SZTrack is able to effectively lateralize seizures, the coarse region based aggregation used to train the network suffers in the anterior vs. posterior task. In the next chapter, we explore a relaxation of this aggregation to allow for more accurate localization.

Chapter 7

SZLoc: An End-to-End Framework for Seizure Onset Zone Localization

7.1 Introduction

In the final chapter of this thesis, SZLoc, a neural network architecture for end-to-end seizure localization, is presented. In conjunction with this novel neural network architecture, a multi-task learning approach employing multiple weakly supervised loss functions is developed to train the architecture to recognize relevant ictal patterns. Using two interconnected signal paths for global and single-channel predictions, SZLoc can be trained to learn multi-scale information for accurate seizure localization. Finally, by leveraging the information rich outputs of the SZLoc network, seizure onset can be localized in the original signal space while providing SOZ maps at the recording and patient level.

Where seizure detection requires only the identification of ictal activity in one or more EEG channels, seizure localization is a challenging problem requiring information across both space and time to be considered. For localization, not only must seizure signals be identified in each EEG channel, their temporal and spatial relationships must be considered. Similar to the R-SMMPL model presented in Chapter 5, SZLoc uses separate signal paths for analyzing electrode level and global seizure activity to capture multi-scale information relevant to localization. However, where the R-SMMPL separated feature extraction and analysis of seizure propagation into neural network and graphical modeling components,

respectively, SZLoc connects these signal paths directly within the neural network structure. This approach eliminates the need for individual components of the model to be trained separately, allowing feature extraction and propagation to be learned concurrently using backpropagation.

By incorporating convolutional, recurrent, and transformer layers, the architecture is capable of identifying seizure activity directly from the EEG signal and relating this activity through the course of seizure onset and propagation. Adopting an approach similar to SZTrack presented in Chapter 6, SZLoc uses a combined convolutional and recurrent network structure to capture evolving seizure activity. Extending beyond SZTrack, SZLoc contains a transformer layer applied spatially across the extracted features for each EEG signal at every time point. This addition allows SZLoc to use self-attention to better identify the cross channel highly correlated activity indicative of seizure onset. Additionally, features from the global CNN are input into the transformer as well, allowing global information to inform the representations extracted at the electrode level.

As in previous modeling approaches, the lack of specific spatial and temporal onset labels complicates training, as only rough onset times and lobar localization annotations are available. To compensate, we adopt a multi-task learning approach, developing novel methods for training the network in this weakly supervised setting. Departing from the zone based SOZ classification task approach used in SZTrack, onset maps are defined for each patient individually. Eliminating the need for catch-all zones with ambiguous boundaries, groups of potential onset channels are defined for each patient. Localization maps are trained with separate loss functions which reward correct localization, penalize onset predictions outside annotated onset regions, and maximize the margin between the positive and negative areas. Furthermore, SZLoc is trained to detect seizure activity prior to and after onset, allowing the network to learn the complicated dynamics of seizure onset without high resolution labels. Thus, only the transition from baseline to seizure is enforced during training, reducing potential problems due to temporal onset label noise.

Localization from SZLoc is evaluated on 45 second seizure onset recordings taken from the JHH dataset. 15 seconds of pre-seizure baseline as well as 30 seconds of EEG after potential onset are included. Onset maps are generated based on coarse localizations to lobe and hemisphere for each patient. While the network is trained on each recording individually, evaluation for localization is performed both on recording and patient aggregated onset maps. This aggregation alleviates issues due to unclear onsets in some recordings and mirrors the clinical practice of evaluating multiple seizure presentations for congruent onset information.

7.1.1 Prior Work

By comparing and combining feature representations in a feed forward manner, transformers allow related information to be shared between separate feature instances. In NLP, this has allowed the transformer architecture to train powerful language models efficiently while eliminating the complications in training recurrent architectures [85]. Recognized for their ability to analyze sequences efficiently, transformers have also found use in speech applications [130]. Noting their ability to analyze related, but not necessarily sequential, representations, transformers have been used spatially in computer vision applications such as [87, 88].

Similar spatial applications of transformer architectures have been made in previous EEG literature. [131] investigates several self attention based architectures for EEG classification and notes potential advantages to attention applied in the temporal, spatial, and feature dimensions. The work presented in [132] combines convolutional layers with a transformer applied across time to perform sleep staging over long EEG recordings. Inspired by successes in language modeling, [133] pretrains a tranformer sequence model coupled with a CNN feature extractor and evaluates the model on several downstream tasks. [134] evaluates a variety of pure transformer and hybrid CNN/transformer models applied spatially and temporally towards a motor imagery classification task. After preprocessing, a transformer is applied across channels in [135] to produce importance weightings between channels. A

compressed version of the multi-channel EEG signal is then windowed and analyzed with a temporal transformer in a motor imagery BCI task. In [136], spatial and temporal attention are combined into one transformer layer and evaluated on an emotion recognition task. Similarly, a convolutional layer is embedded within a transformer in [137] to classify visual stimuli from EEG activity.

7.2 Methods

In the following section, the SZLoc architecture, derivation of onset maps, and training is described. In the first three subsections, the architectural components of SZLoc, are detailed. The fourth and fifth subsections address methods to leverage the high dimensional outputs of SZLoc to produce interpretable, clinically relevant information. In the next subsection, an ensemble of weakly supervised loss functions are introduced to allow SZLoc to be trained to localize SOZ using multi-task learning. Finally, we give implementation details, including training hyperparameters and data augmentation techniques.

7.2.1 CNN-Transformer Feature Extraction

In the feature extraction phase, representations at the global and electrode level are extracted from 1 second windows of EEG signal. A schematic of the feature extraction portion of SZLoc is shown in Figure 7-1. In the global CNN, a 1D CNN applied across all electrode channels extracts a 160 dimensional feature representation of the entire EEG signal. This signal path is shown on the left side of Figure 7-1, with extracted features shown in yellow. Applied on each of the 19 channels of the 10-20 system individually, the electrode CNN extracts length 160 features, shown in red in Figure 7-1, for each separate electrode. A transformer layer, consisting of an encoder-decoder structure, fuses information between both the global and electrode level representations of the EEG signal. All electrode representations as well as the global representation are provided as input to the encoder. Encoded inputs are then shared with the original electrode level feature representations in the decoder. Informed by both

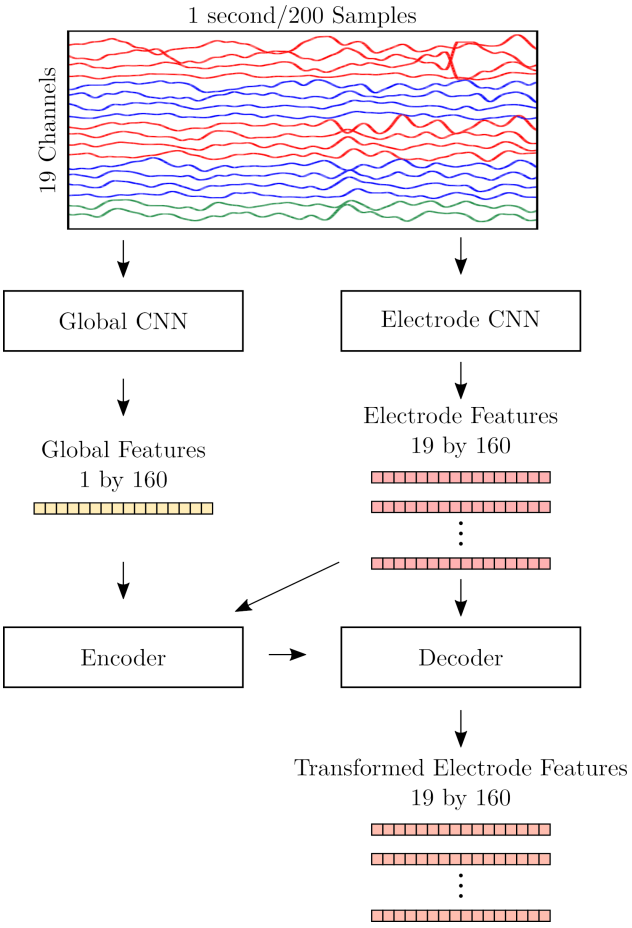


Figure 7-1. Network feature extraction

global and electrode level information, representations of the EEG signal in each electrode are output from the decoder, shown in orange in Figure 7-1.

Two separate CNN architectures are used for the global and electrode signal paths. Each CNN consists of an embedding layer, followed by a cascade of residual blocks and convolutional projection layers. Each residual block first applies a convolution layer followed by batchnorm and a PReLU activation. Another application of convolution and batchnorm follow, with a residual connection between the input of the block prior to the final PReLU activation. Convolutions within the block use the same number of channels as the block's input and use kernels with a length 7 and a stride of 1. Projection layers are used in between convolutional blocks to double the number of channels while halving the length of the representation and employ a kernel size of 3 and a stride of 2 unless otherwise noted. Finally global average

pooling at the end of the CNN is applied to reduce the time-series channels to a single length 160 feature vector.

In the global CNN, the original 19 channel EEG signal is passed through an embedding convolution with kernel size 7 and an output of 80 channels. Next a residual convolution block with 80 channels is applied. A projection layer increases the number of channels to 160, followed by another convolution block. The representation is then passed through a convolution layer with 160 channels of input and output, kernel size 1, and a stride of 2 to reduce the length of the representation. A final residual block is applied to 160 channel signal before global average pooling. Dropout is applied to the final feature representation during training.

The electrode CNN uses a similar architecture but increases the depth of the CNN. The same CNN is trained across each channel, ensuring consistency in the final feature representation and using the multi-channel signal most efficiently. The individual electrode signal is passed through an embedding convolution with kernel size 7 and 20 output channels. Next 3 residual blocks and projection layers are applied, increasing the representation from 20 to 40, 80, and 160, while halving the length of the representation after each projection. A final residual block is applied with 160 features before global average pooling and dropout are applied.

The decoder and encoder layers of the transformer follow [85], however only one layer is used in both the encoder and decoder. In preliminary experimentation, increased transformer depth led to poorer generalization, likely due to overfitting. The feedforward dimension is set to 256 and dropout is applied. The 19 length 160 representations from from the electrode CNN and 160 dimensional representation from the global CNN are fed into the encoder of the transformer. The 19 electrode representations are input into the decoder, resulting in a length 160 feature representation of each electrode channel. Thus each electrode channel representation incorporates information from the other channels as well as the global EEG signal for effective multi-scale information fusion.

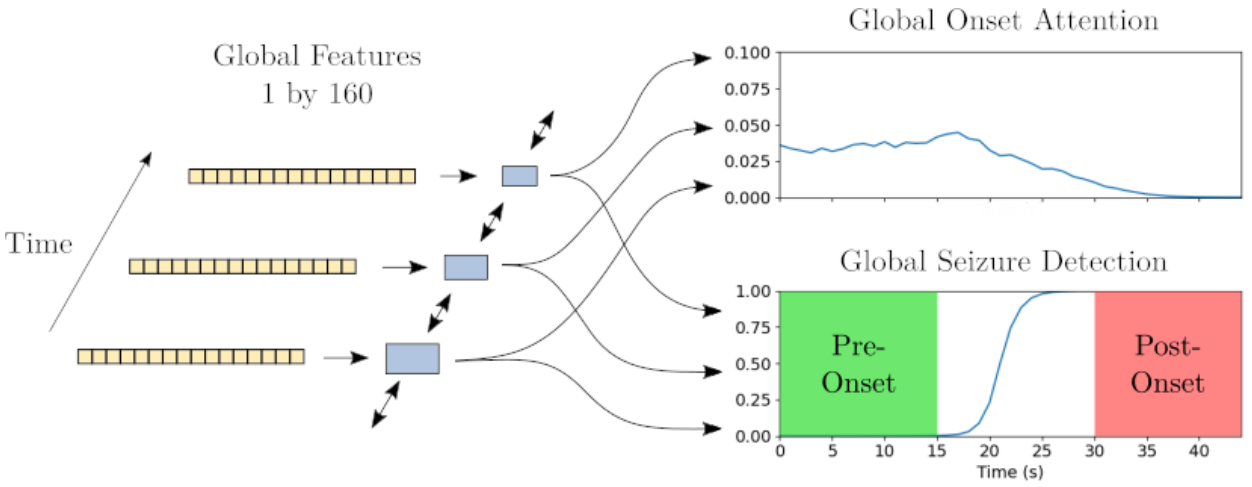


Figure 7-2. Global signal temporal analysis. Global features shown in yellow are analyzed by a bidirectional GRU network, indicated by the blue rectangles. At each window, output from the GRU is used to produce a global seizure prediction $S^g[t]$ and onset attention score $a^g[t]$.

7.2.2 Global Seizure Activity Analysis

Shown in Figure 7-2, features from the global signal path are used to predict seizure activity and derive global onset attention. Global feature representations are input into a bidirection GRU network. Outputs from the GRU network are used to derive global onset attention $a^g[t]$ and seizure detection $S^g[t]$ for every window t of the seizure onset recording. Detailed in Section 7.2.6, seizure detection S^g will be trained to identify pre-seizure and post-onset seizure activity. Global onset attention a^g will be used to generate SOZ localization maps for each recording and is further described in Section 7.2.5.

The bidirectional GRU uses 2 layers with a hidden size of 80. Thus the combined output from the forward and backward directions is 160 dimensional. To generate predictions $\hat{S}^g[t]$, the length 160 representations at each timestep are fed into a linear layer with an output of 2 and a softmax is applied. Global onset attention scores $a^g[t]$ are generated using a linear layer which reduces the 160 length representation to a scalar at each time window. The softmax operation is applied across time windows so that $\sum_{t=1}^{T} a^g[t] = 1$.

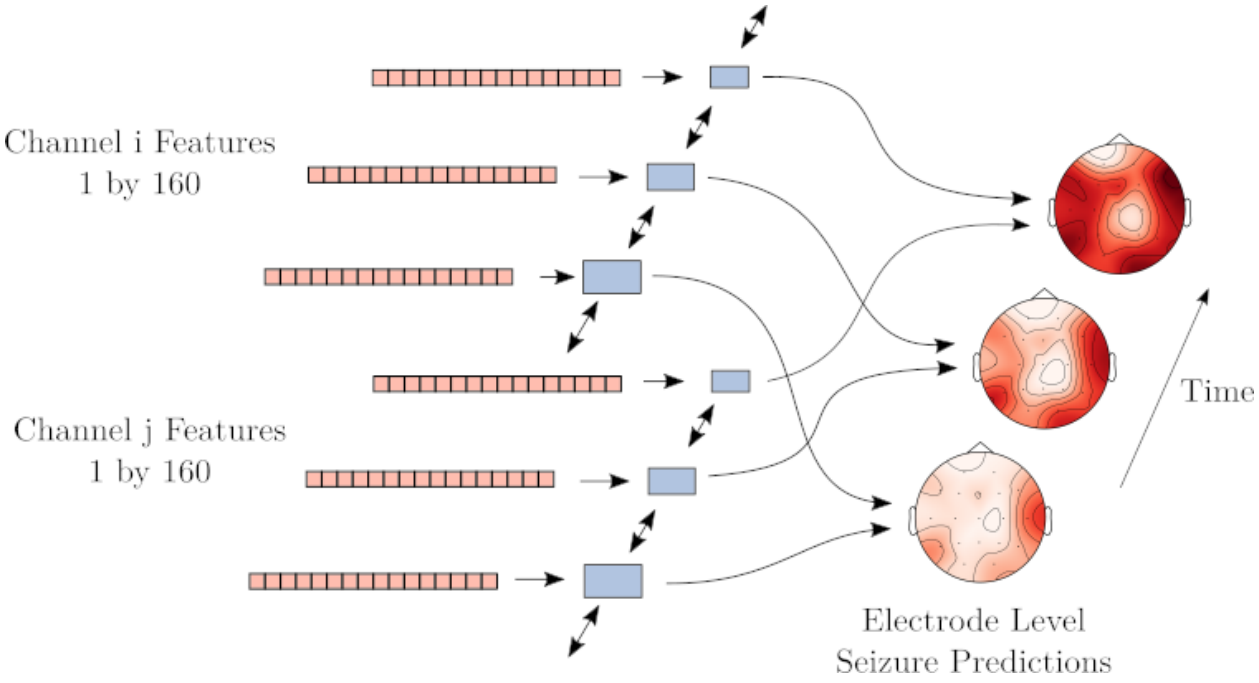


Figure 7-3. Single-Channel Detection

7.2.3 Electrode Level Seizure Prediction

After feature extraction, each channel's feature vector is analyzed for evolving seizure activity using a bidirectional GRU as shown in Figure 7-3. The same GRU network is used for each channel. For each channel i of the 19 channel 10-20 system and time window t, the GRU network is used to generate an electrode level prediction of seizure activity $\hat{Y}_i[t]$. These predictions are shown on the right of Figure 7-3 as topographic plots of seizure activity superimposed on the layout of the 10-20 system.

Following the global signal path, the electrode signal path is fed into a 2 layer bidirectional GRU with a hidden size of 80. The 160 dimensional output is passed through a linear classification layer and a softmax is applied to generate electrode level seizure activity predictions $\hat{Y}_i[t]$. As in the electrode CNN for feature extraction, the same GRU and linear layers are trained across channels to efficiently leverage information from the multi-channel signal during training.

7.2.4 Seizure Detection and Onset Attention from Electrode Predictions

While the electrode level signal path provides only seizure activity predictions $\hat{Y}_i[t]$, we derive overall seizure detection predictions $\hat{S}^e[t]$ and attention onset scores $a^e[t]$ from these more granular network outputs. Following methods used in the SZTrack network presented in Chapter 6, seizure activity is pooled across channels to create an overall prediction of seizure.

$$\hat{S}^e[t] = \max_i \hat{Y}_i[t] \tag{7.1}$$

The overall predicted seizure is taken to be the maximum predicted seizure activity across all channels at time window t. This approach allows us to train the channel predictions on seizure level labels, avoiding the need for labels of seizure activity in each individual channel. Again following SZTrack, the first order difference of the overall seizure prediction derived from the individual electrodes is used to derive onset attention scores.

$$a^e[t] = \max\left(\hat{S}^e[t] - \hat{S}^e[t-1], 0\right) \tag{7.2}$$

These predictions and onset scores will be used to generate localization maps as well as to train the network to identify seizure activity.

7.2.5 Generating Seizure Level Onset Maps

As in Chapter 6, the onset attention scores and electrode level seizure activity predictions are combined to generate SOZ localization map predictions $\hat{O}$ for each seizure recording. In this subsection, we use the generic onset score variable a in place of the specific signal path variables a^g and a^e. However, in the following sections we will adopt the same superscript notation, e.g. $\hat{O}^g$ and $\hat{O}^e$, to indicate the source of onset scores used to generate an SOZ map. Differing from the approach taken in the previous chapter, seizure localization maps are derived in two steps. First, elementwise multiplication is used to generate the matrix P from attention scores a and electrode predictions $\hat{Y}_i$, $P_{it} = a[t]\hat{Y}_i[t]$. Next, a summation

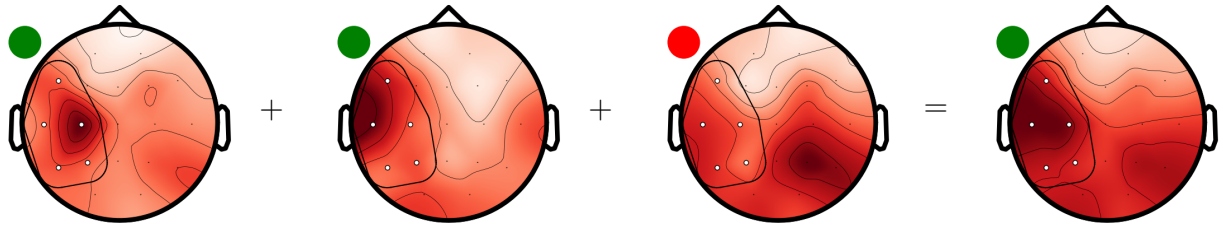


Figure 7-4. Localization aggregation for an example patient. SOZ is correctly localized to the left temporal region in 2 recordings. The remaining recording predicts SOZ in the right parietal region. After aggregation over all 3 recordings, the total SOZ prediction is in the left temporal region as indicated by clinical annotations.

across time windows and normalization are performed to generate SOZ maps.

$$\hat{O}_i = \frac{\sum_{t=2}^{45} P_{it}}{\sum_{i=1}^{19} \sum_{t=2}^{45} P_{it}} \tag{7.3}$$

Effectively, the matrix P represents the electrode channels i and time windows t which contribute to the localization map O. By analyzing P, the specific EEG windows considered responsible for seizure onset can be directly identified in the native signal space.

As in SZTrack we aggregate localizations across a patient's recordings to arrive at a patient level localization map. Figure 7-4 depicts aggregation graphically, where the combination of two correct and one incorrect localization results in a correct patient level SOZ prediction. This process mirrors clinical practice, as congruent SOZs across multiple recordings are typically needed to generate a hypothesis of a patient's SOZ. Due to propagation effects, seizure activity originating in one part of the brain may first manifest as recognizable seizure activity in distant electrodes. Furthermore, seizure onset may be obscured by artifact or otherwise unrecognizable, rendering an individual seizure recording devoid of SOZ localization information. By aggregating over multiple recordings, more reliable SOZ maps can be generated for a given patient.

7.2.6 Weakly Supervised Loss Functions

To train the architecture from inexact temporal and spatial onset annotations, we develop several methods of weakly supervised training employed in a multi-task learning paradigm.

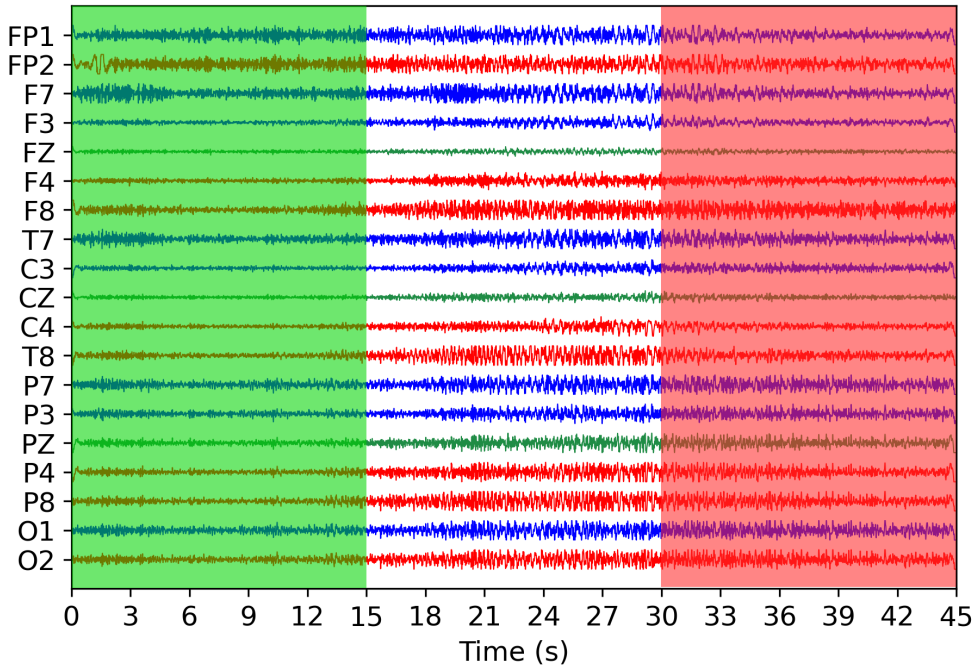


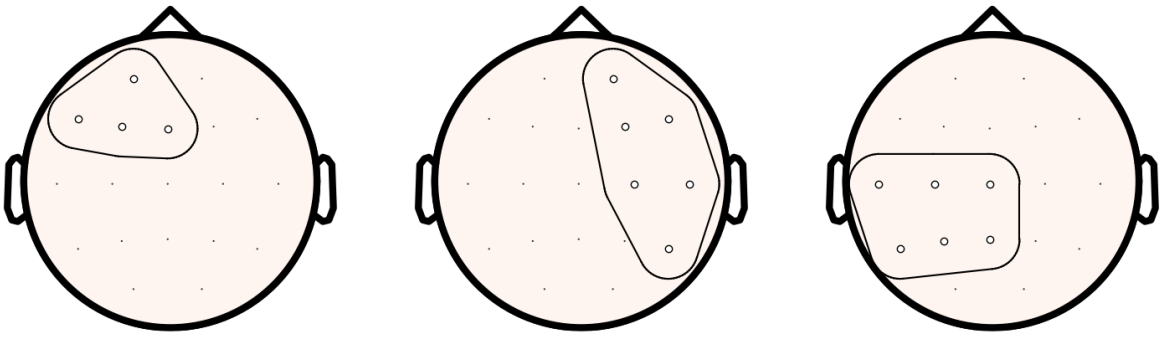
Figure 7-5. Detection regions for seizure detection training. Cross-entropy loss is applied for pre-seizure and post-onset regions. No loss is applied from 15–30 seconds.

By identifying an ensemble of desirable properties, four novel loss functions are defined. The first loss function allows SZLoc to be trained for seizure detection while letting the network learn to identify onset without high resolution onset annotations. The remaining three loss functions reward and penalize desirable properties for seizure onset maps.

The detection loss function allows SZLoc to be trained from inexact seizure onset time annotations. Depicted in Figure 7-5, cross entropy loss is applied to the pre-seizure and post-onset seizure predictions only. Mathematically,

$$\mathcal{L}_{detection}(\hat{S}) = -\frac{1}{30}\sum_{t=1}^{15}\log(1-\hat{S}[t]) - \frac{1}{30}\sum_{t=31}^{45}\log(\hat{S}[t]) \tag{7.4}$$

where $\hat{S}$ may be generated by the global or electrode signal paths. It is assumed that seizure onset occurs between 15 and 30 seconds during the recording, with the exact time being unknown. As such, these timepoints are omitted from the detection loss function. This approach enforces that the architecture learns to transition from pre-seizure into seizure while allowing the network the freedom to identify seizure onset activity occuring between 15 and



(a) Left Frontal (b) Right Temporal (c) Left Parietal

Figure 7-6. Example seizure onset map labels. Electrode channels are designated as potential seizure onset locations. (a) shows potential channels for a left frontal seizure onset zone. (b) shows a right temporal SOZ while (C) depicts a left parietal onset zone.

30 seconds.

For SOZ localization, we develop three loss functions designed to capture complimentary behavior relevant to task. These functions reward correct localization, penalize incorrect localization, and maximize the margin between predictions in the correct and incorrect regions. Example SOZ map labels, O are given in Figure 7-6. Potential electrode channels corresponding to clinical annotations of SOZ are identified for each patient. In Figure 7-6 (a), electrode channels which correspond to a left frontal SOZ are enclosed. Simlarly, in Figure 7-6 (b) and (c) maps for right temporal and left parietal onset zones are shown. To simplify notation, we denote the set of channels with the onset region for a given map as $\mathcal{P}(O)$ and let $\bar{O} \triangleq \hat{O} / \max_i \hat{O}_i$ be SOZ predictions normalized such that the maximum value over all channels is 1.

To reward correct localizations we define the loss function

$$\mathcal{L}_{loc+}\left(\bar{O}, O\right) = \frac{\sum_{i \in \mathcal{P}(O)} \left(1 - \bar{O}_i\right)^2}{|\mathcal{P}(O)|} \tag{7.5}$$

This loss function applies the ℓ_2 norm to enforce correct predictions in the ground truth SOZ and scales the loss such that it ranges between zero and one. Analogously, we define the negative loss function

$$\mathcal{L}_{loc-}\left(\bar{O}, O\right) = \frac{\sum_{i \notin \mathcal{P}(O)} \bar{O}_i^2}{19 - |\mathcal{P}(O)|} \tag{7.6}$$

which penalizes incorrect predictions of seizure onset with a similar ℓ_2 penalty and scaling. The final localization based loss maximizes the margin between predictions inside and outside the ground truth SOZ.

$$\mathcal{L}_{margin}\left(\bar{O}, O\right) = \frac{1}{2}\left(1 - \max_{i \in \mathcal{P}(O)} \bar{O}_i + \max_{i \notin \mathcal{P}(O)} \bar{O}_i\right) \tag{7.7}$$

Max pooling is performed over channels within the positive and negative regions and the difference is taken. By maximizing the difference, the network is encouraged to localize onset activity within the annotated SOZ while minimizing the weight placed outside of it. 1 is added and the result is divided by 2 to ensure the loss function ranges between 0 and 1.

7.2.7 Implemtation Details

7.2.7.1 Data Augmentation

To combat the limited size of the dataset, three data augmentation techniques are implemented. The first technique injects random Gaussian noise into the raw EEG signals. For each EEG channel, let $M_i[t] \sim \text{Bernoulli}(0.5)$ be a random variable indicating whether or not Gaussian noise will be added. Similarly, the variance of the additive noise for each window is sampled from a Gaussian distribution $V_i[t] \sim \mathcal{N}(0, 0.1)$ with variance 0.1. Mathematically, noise augmentation $G(\cdot)$ applied in channel i at time t can be written as

$$G(X_i[t]) = X_i[t] + M_i[t]V_i[t]\epsilon \tag{7.8}$$

where $\epsilon \in \mathbb{R}^{200}$ is generated from a multivariate normal distribution with mean 0 and standard deviation 1. This data augmentation technique applies noise to each channel and window individually, reinforcing invariance to changes in noise conditions across channels and time.

In addition to random Gaussian noise, signal time reversal and cross-hemispheric signal flipping is applied. For signal time reversal, the EEG signal in each window for all channels is reversed. Note that while each window is reversed, the order of the windows remains consistent, ensuring that seizure onset and propagation information is preserved. Cross-hemispheric flipping interchanges left and right channels, creating a mirror image of the

original EEG signal and onset labels. This flipping ensures that the global signal path is invariant to mirror image transformations. The electrode signal path remains unaffected as no sense of spatial position is preserved inherently in this part of the network. In contrast to the window based additive noise augmentation, to preserve important phase and spatial relationships between channels, both of these techniques are applied for all electrodes and time windows for a given sequence. Each augmentation technique is applied with a probability 0.5 to each sequence.

7.2.7.2 Training Details

During training, the four loss functions are applied to outputs of the network in a multi-task learning paradigm. Complimentary information from each loss function is used to arrive at a final SOZ prediction. Detection losses are applied to outputs from both the global and aggregated electrode detection predictions. Positive, negative, and margin based losses are applied to SOZ localizations generated using both global onset attention scores and electroe onset attention scores. Multiplicative factors for $\mathcal{L}_{loc+}$ are set to 2 while the remaining loss factors are left at unity. Mathematically,

$$\begin{aligned}\mathcal{L}_{total} =& \mathcal{L}_{detection}(\hat{S}^g) + \mathcal{L}_{detection}(\hat{S}^e) \\ &+ 2\mathcal{L}_{loc+}(\hat{O}^e, O) + \mathcal{L}_{loc-}(\hat{O}^e, O) + \mathcal{L}_{margin}(\hat{O}^e, O) \\ &+ 2\mathcal{L}_{loc+}(\hat{O}^g, O) + \mathcal{L}_{loc-}(\hat{O}^g, O) + \mathcal{L}_{margin}(\hat{O}^g, O)\ .\end{aligned} \tag{7.9}$$

SZLoc is implemented in Pytorch 1.9. The network is trained for 100 epochs using the Adam [91] optimization algorithm with a batch size of 5 and a learning rate of 0.001. Weight decay of 0.001 is applied and dropout in the CNN, transformer, and GRU are set to 0.1.

7.2.8 Validation Strategy

We train the SZLoc architecture using leave-one-patient-out cross validation (LOPO-CV). Results are aggregated across all left out patients. Using this strategy, we are able to mimic SZLoc's application to unseen patients to evaluate generalization. SZLoc and competing

baselines are evaluated for 5 random initializations and results are then averaged across these test runs. To evaluate correct or incorrect localization, we consider the maximum value in the predicted onset map $\hat{O}$. If the maximum predicted weight $\arg\max_i \hat{O}$ is within the labeled SOZ for a recording or patient $\mathcal{P}(O)$, the localization is considered correct.

7.2.8.1 Evaluating Multi-Task Onset Attention

To evaluate the contribution of the global and electrode sources of onset attention, localization losses for the global and electrode are included and excluded in different test configurations. In the first experiment, localization losses for both global and electrode onset attention generated SOZ maps are used to train the model. Next, by setting multiplicative factors for the relevant loss functions to 0, we evaluate the performance of the models using electrode and global onset attention individually. By comparing performance between paradigms, the gains in localization accuracy due to each source of localization attention can be quantified.

7.2.8.2 Baseline Models

As SZLoc represents a novel step towards end-to-end seizure localization not before seen in the literature, we opt to evaluate SZLoc in comparison to ablated and reordered versions of the original architecture. The CGT baseline represents the SZLoc components reordered such that the GRU follows convolutional feature extraction and precedes the transformer layer. In the CG baseline, the transformer layer is completely omitted, rendering the network similar to SZTrack. In fact, without global onset attention losses applied, the CG baseline operates analogously to SZTrack for SOZ localization. Note that because the CNN output, and both input and output for the GRU and transformer are of length 160, the GRU and transformer can be reordered and omitted without need to adjust network dimensions. Finally, the SZLoc-No Connect and CGT No-Connect models represent versions of the SZLoc and CGT networks where global features are not included in the transformer encoder input.

Table 7-I. Localization results with electrode onset attention a^e and global onset attention a^g applied in conjunction. Patient aggregated and indvidual recording results are presented for each model

Model	Patient Electrode	Seizure Electrode	Patient Global	Seizure Global
SZLoc	24.2 ± 1.0	109.6 ± 8.2	23.6 ± 0.5	112.0 ± 5.4
CGT	15.6 ± 2.4	83.4 ± 10.9	14.6 ± 1.9	84.6 ± 8.6
SZLoc-No Connect	19.6 ± 2.6	102.0 ± 4.3	18.6 ± 0.5	102.0 ± 4.3
CGT-No Connect	16.4 ± 1.8	81.4 ± 7.9	17.8 ± 3.0	80.6 ± 8.9
CG	19.4 ± 2.6	87.2 ± 3.7	20.2 ± 2.2	91.8 ± 3.0

Table 7-II. Localization results with electrode onset attention a^e and global onset attention a^g applied separately. Multiplicative factors for loss functions corresponding to localization with the omitted source of onset attention are set to zero. Patient aggregated and indvidual recording results are presented for each model.

Model	Patient Electrode	Seizure Electrode	Patient Global	Seizure Global
SZLoc	23.0 ± 1.5	105.2 ± 7.0	19.2 ± 2.0	101.6 ± 3.3
CGT	19.0 ± 2.2	93.6 ± 3.3	19.2 ± 1.0	92.4 ± 12.8
SZLoc-No Connect	20.8 ± 2.9	101.6 ± 4.6	17.6 ± 1.5	94.2 ± 7.7
CGT-No Connect	17.2 ± 1.6	94.8 ± 6.6	18.6 ± 3.1	89.4 ± 1.8
CG	20.0 ± 3.0	102.0 ± 7.0	17.2 ± 2.5	90.2 ± 4.8

7.3 Results

7.3.1 Clinical Dataset

45 second seizure onset segments are extracted from each seizure recording in the JHH dataset. 15 seconds of pre-onset seizure baseline are included, as well as the following 30 seconds seizure EEG. All 34 patients are included for a total of 201 seizures. EEG recordings are high-pass filtered at 0.5 Hz and low-pass filtered at 30 Hz. EEG signals from each channel are normalized to mean 0 with standard deviation 1. After normalization, amplitudes greater than 2 standard deviations are thresholded. One second windows with no overlap are extraced for input into the network.

7.3.2 Localization Results

Localization results for each network using all loss functions are presented in Table 7-I. Onset localization results based on electrode onset attention and global onset attention are given

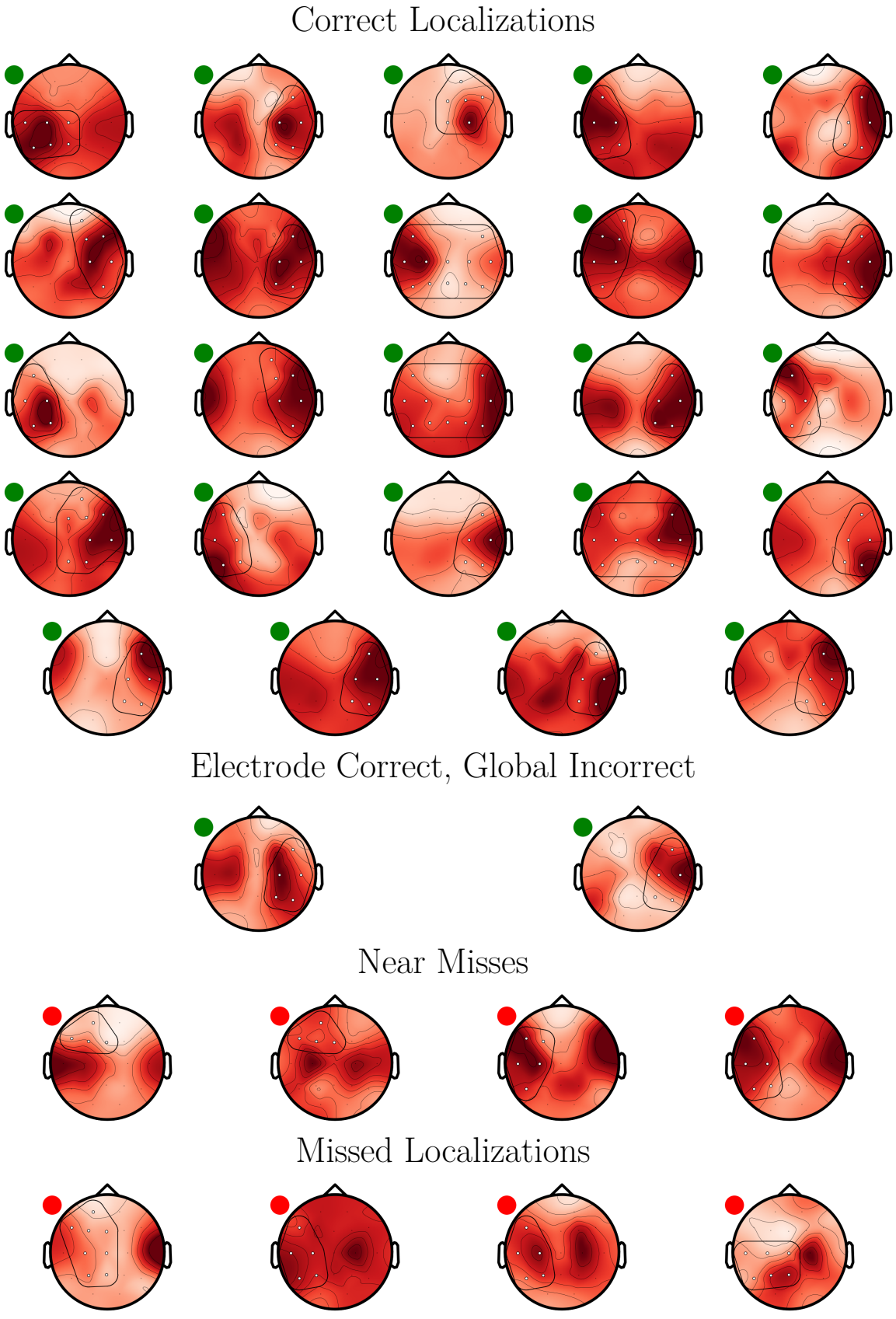


Figure 7-7. Channel attention localization results

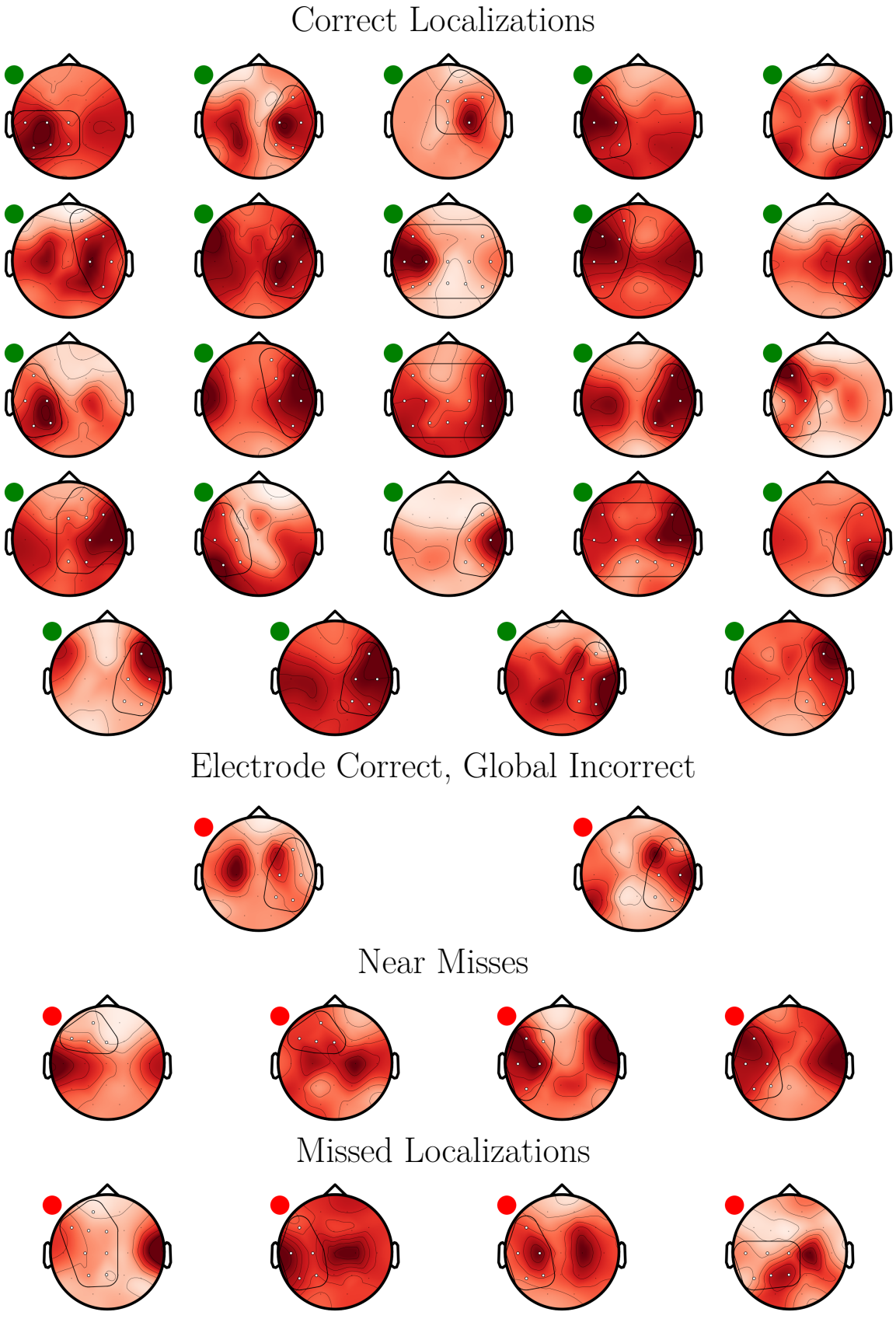
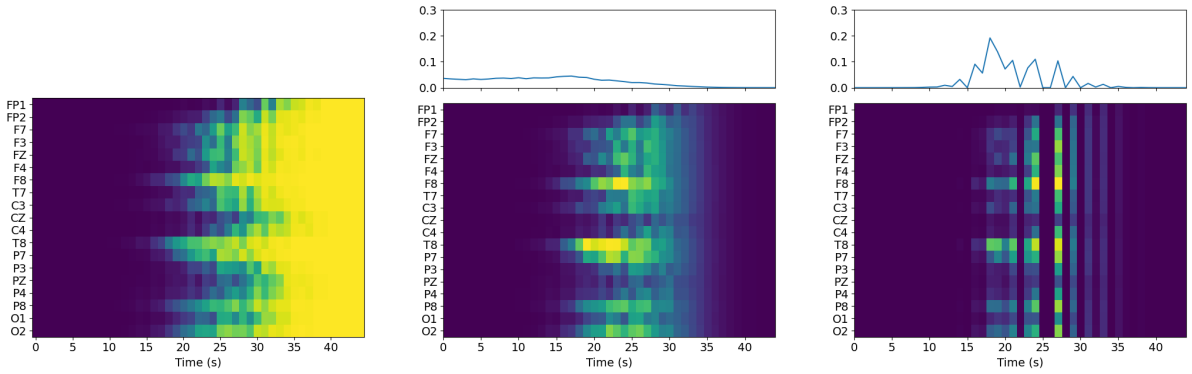


Figure 7-8. Channel attention localization results

aggregated for each patient and at the seizure level. Results are averaged over 5 random initializations of each network. In all localization metrics, SZLoc outperforms all competing baselines. At best, out of 34 patients, SZLoc correctly localizes onset zones in an average of 24.2 ± 1.0 patients with onset attention derived from electrode predictions. SZLoc identifies onset zones correctly in an average of 112.0 ± 5.4 for best recording level performance using global attention based localization. Of the competing baselines, the CG network identifies the largest number of correct seizure onset zones at the patient level, with an average of 20.2 ± 2.2, using multi-channel attention. For individual seizures, the SZLoc-No Connect network correctly predicts 102.0 ± 4.3 seizures for both multi-channel and single-channel onset attention. Across all metrics, the CGT and CGT-No Connect models perform the worst, with best patient level localization of 17.8 ± 3.0 by the CGT-No Connect model and best seizure level localization of 84.6 ± 8.6 by the CGT.

Table 7-II gives results for each model with attentions a^e and a^g applied trained separately. Again the SZLoc architecture outperforms all competing baselines, achieving patient and seizure correct localizations of 23.0 ± 1.5 and 105.2 ± 7.0 with electrode onset attention. Next best performing at the patient level is the SZLoc-No Connect (20.8 ± 2.9) and CG at the recording level (102.0 ± 7.0), both with electrode onset. While the SZLoc architecture performs worse, interestingly the competing baselines trained with individual onset sources on average outperform versions of the models with both electrode and global losses applied.

Figure 7-7 shows patient level aggregated SOZ localization maps for the best performing random initialization. In total, SZLoc placed the mode of the SOZ onset map within the annotated SOZ for 26 out of the 34 patients. On inspection of the remaining 8 patients, we note that in 4 cases SZLoc placed significant weight within the annotated SOZ or identified channels just outside of the SOZ as likely onset zones and heuristically designate these cases as near misses. In only 4 cases did SZLoc fail to accurately place substantial SOZ prediction weight near or within the annotated SOZ. In Figure 7-8 localization maps are shown for each patient when localizations are derived using global onset attention a^g. The location of each



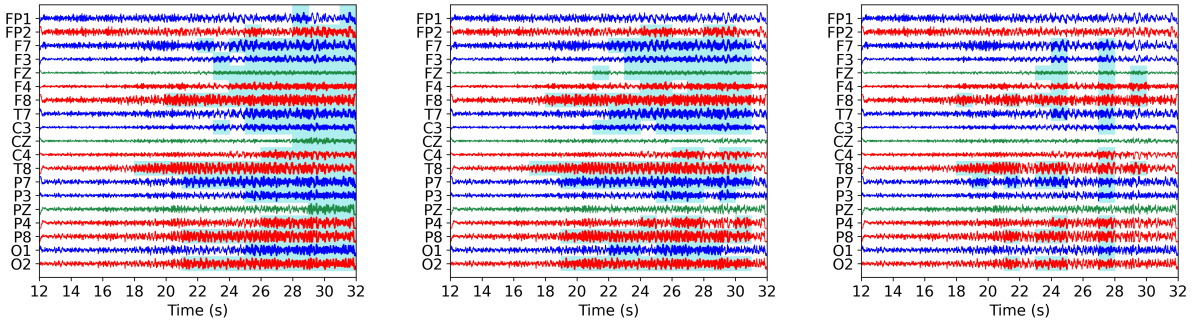
(a) Channel Predictions (b) Global attention onset (c) Electrode attention onset

Figure 7-9. Seizure and onset predictions for patient 5 overlayed on an EEG recording of a right temporal seizure. (a) Seizure predictions $\hat{Y}$ from each individual channel are shown. Seizure activity begins in channel T8 and spreads to neighboring channels. (b) Derived localization attention P^g and a^g. (c) Derived localization attention P^e and a^e.

patients onset map is unchanged between Figures 7-7 and 7-8. Performance is similar to the electrode case, however 2 additional patients are incorrectly now incorrectly classified.

Figures 7-9 and 7-10 demonstrate SZLoc's ability to generate predictions of seizure activity in each electrode channel $\hat{Y}_i$ and employ onset attention to identify likely sources of seizure onset via P. Figures 7-9 (a) and 7-10 (a) show individual electrode predictions as a heatmap and superimposed on the EEG channel, respectively. Beginning at or shortly after the annotated seizure onset at 15 seconds, SZLoc begins to predict seizure activity in electrode T8. As this ictal activity spreads to neighboring electrodes, SZLoc accurately identifies this propagation.

In Figure 7-9 (b) and (c), onset attention a and matrices P are shown derived from global and electrode onset attention, respectively. Figure 7-10 (b) and (c) shows matrices P^g and P^e superimposed on the original EEG signal. In this visualization, the matrices P can be used to identify the EEG morphologies contributing to SZLoc's generated SOZ maps in the original signal space. At the top of Figure 7-9 (b), a^g remains nearly constant from the beginning of the seizure before declining to 0 near 30 seconds. Figure 7-9 (c) shows the electrode derived onset attention a^e, which peaks closer to 15 seconds after the onset of the seizure. As such, the derived matrix P^g highlights many more channels and windows as potential



(a) Channel Predictions (b) Global attention onset (c) Electrode attention onset

Figure 7-10. Seizure and onset predictions for patient 5 overlayed on an EEG recording of a right temporal seizure. (a) Seizure predictions $\hat{Y}$ from each individual channel are shown. Seizure activity begins in channel T8 and spreads to neighboring channels. (b) Derived localization attention P^g using a^g. (c) Derived localization attention P^e using a^e.

onset locations while P^e focuses more on specific channels and times. This contrast can be observed in Figure 7-10 (b) and (c), where P^g identifies many more times and channels as contributing to localization map than P^e.

7.4 Discussion

Using a combined convolutional, transformer, and recurrent network, SZLoc blends information across electrode and global signal paths to generate SOZ localization maps at the patient and recording level. With an ensemble of weakly supervised loss functions, SZLoc is trained using a multi-task learning framework to incorporate separate aspects of the seizure localization task. When trained using all localization loss functions, SZLoc achieves the best localization scores across both sources of onset attention and on the patient and recording levels. Similarly, when training is performed with only one source of attention, SZLoc performs best in all metrics.

When trained using each source of attention separately, SZLoc's performance remains best across models, however its overall performance is decreased. Interestingly, many of the competing baselines exhibit better performance when trained using only one source of onset attention. This result indicates that while the full ensemble of weakly supervised loss

functions may confound learning in the competing baselines, SZLoc is able to effectively leverage our multi-task framework and achieve its best performance with all losses applied.

By aggregating over recordings, the proportion of correct SOZ maps is improved. For example, when considering electrode derived onset maps, an average 24.2 of 34 (71.1%) patients are correctly localized while 109.6 of 201 (54.5%) recordings are correctly localization by SZLoc. Thus by aggregating over a patient's recordings, more accurate localizations can be generated. Noting near misses in Figure 7-7 which place significant weight in the correct onset zone, the quality of patient level SOZ maps could be further refined by identifying recordings containing better localization information. In future work, this could be accomplished by including a scoring mechanism to weight the contribution of each recording when aggregating over a patient's seizure presentations.

Not only does SZLoc provide seizure onset maps as shown in Figures 7-7 and 7-8, Figure 7-10 demonstrates how SZLoc's outputs can be used to identify seizure activity in the original EEG signal space. In Figure 7-10 (a), electrode level predictions are superimposed on the EEG signal, allowing all predicted seizure activity to be identified. By leveraging the combined onset attention a and electrode level predictions $\mathbf{Y}$, the matrices P can be used to identify specific EEG windows and channels contributing to SZLoc's final recording level localization as shown in Figure7-10 (b) and (c).

In Figures 7-9 and 7-10 electrode level predictions as well as onset attention and matrices P are shown. In Figure 7-9 (b), global onset attention a^g can be observed to remain constant from the beginning of the recording and declining slowly to zero near 30 seconds. However, as the matrix P^g is derived by multiplying electrode predictions $\hat{Y}$ with global onset attention a^g, time windows before 15 seconds where $\hat{Y}$ is near 0 do not contribute to the final localization prediction. Still, when comparing Figure 7-9 (b) and (c) it is apparent that more channels and time windows contribute to the onset map when using a^g than when a^e is used. This difference can be noted further when comparing Figure 7-10 (b) and (c), where the electrode case more correctly focuses on the true onset of seizure in the right temporal region.

This difference in onset behavior indicates a possible source for the difference in performance between electrode and global onset attention. Using a^g, more channels and time windows contribute to the final recording level localization. The sharper a^e may generate more accurately localized onset attention, resulting in increased localization performance. However, due to the non-monotonicity of the S^e, the electrode onset matrix P^e highlights non temporally contiguous regions of the EEG signal. Further refinements are thus needed to focus onset attentions a^e and a^g closer to the seizure onset.

7.5 Conclusion

In this chapter we present SZLoc, a hybrid convolutional, transformer, and recurrent neural network for seizure localization. By fusing information across multiple signal paths, multi-scale global and electrode level spatio-temporal seizure activity contributes to SOZ localization. Using an ensemble of weakly supervised loss functions, desirable properties of accurate SOZ maps are balanced to generate localization maps at the recording and patient level in a multi-task learning paradigm. By leveraging the information rich output of the SZLoc model, seizure activity and EEG activity contributing to the final localization can be visualized in the original signal space. Taken together, SZLoc is capable of providing clinically useful information at multiple scales to aid in the localization of focal epileptic seizures.

The SZTrack architecture presented in the previous chapter is improved upon in multiple ways. While an identical aggregation for generating recording level SOZ maps is employed, by separating the procedure into two operations the intermediate resuls are used to identify EEG onset windows in the original signal space. Through the incorporation of a transformer layer, multi-scale information is successfully leveraged in SZLoc, while SZTrack analyzed each EEG channel individually. By reframing seizure localization from a classification task, as in SZTrack, and instead training the network to enforce desirable properties of correct localizations, SZLoc is able to move beyond the coarse region based localizations of the previous work.

Discussion and Conclusions

In this thesis we have presented methods for detection, tracking, and localization of focal epileptic seizures using clinically recorded scalp EEG signals. While previous approaches in the scalp EEG literature have focused primarily on detection alone, methods presented here seek to capture clinically relevant phenomena to provide clinicians with automated tools for SOZ localization. In clinical practice, SOZ localization from scalp EEG is challenging, requiring time consuming analysis of hours of EEG signals performed by highly trained neurologists. Automated methods capable of providing clinically relevant diagnostic information have the potential to aid in the clinical workflow.

The methods presented in this thesis are directly informed by clinical practice. In analyzing an EEG recording of a seizure, a clinician will identify the onset of seizure activity and trace its propagation as it spreads through the EEG signal space. Mirroring this analysis, the methods here seek to capture the structured spatio-temporal seizure spreading phenomena through graphical modeling and deep neural network approaches.

The first half of this thesis explores methods based on graphical modeling. In Chapter 3, the CHMM model is introduced. By representing the EEG signal as a collection of interrelated HMM chains, seizure activity was tracked through the space of EEG channels. Despite being trained for seizure detection alone, we observe that the CHMM provides information useful for tracking the onset of seizure activity in individual EEG channels. However, localization in the CHMM was performed heuristically, as the model was only capable of identifying onset in a post hoc fashion.

Circumventing the need for hand designed feature engineering, Chapter 4 explored the incorporation of CNN likelihoods into the original CHMM model. By training CNNs for seizure prediction directly on the EEG signal, seizure detection performance was improved over the purely graphical modeling approach in Chapter 3. However, this gain in performance came with the limitation that models could no longer be trained in an end-to-end fashion. Neural likelihoods must be pretrained before the CHMM can be learned using their outputs. Still, the powerful feature extraction from neural likelihoods led to an overall gain in performance.

In Chapter 5, we presented the R-SMMPL model. Improving upon many of the limitations of the CHMM, the R-SMMPL used a hierarchy of interrelated random variables to fuse information at multiple scales. Effectively functioning as a global seizure detector, a switching chain was incorporated to control the seizure propagation dynamics of a CHMM similar to the one presented in Chapter 3. This switching chain was informed by a CNN used to identify seizure activity in the global EEG signal. By enforcing a global seizure state, onset was limited to a single channel for each recording while all chains were restricted to exit the seizure state simultaneously. Furthermore, by learning a distribution of localizations for each patient, clinically relevant SOZ maps of potential onset were directly output from the model. These improvements extended the utility of the original CHMM model while still allowing the model to predict seizure activity at the resolution of individual channels.

In Chapter 6, we make the transition from hybrid graphical modeling and neural network approaches to end-to-end neural network approaches for seizure detection, tracking, and localization. Noting the seizure detection efficacy of the CNN-BLSTM presented in the appendix, the SZTrack network described in Chapter 6 applies a similar convolutional and recurrent hybrid network to EEG signals extracted from each individual electrode. By aggregating seizure activity across EEG channels, a global seizure annotations can be used to train the network to identify seizure activity at the electrode level. Effectively, this allows SZTrack to learn to identify seizure activity at resolutions higher than the seizure annotations provided. Coarse seizure localization classifications are derived from the model outputs,

allowing SZTrack to identify the source of seizure activity to hemispheres and anterior vs. posterior regions.

Chapter 7 develops SZLoc, a combination convolutional, transformer, and recurrent neural network for seizure localization. SZLoc extends the SZTrack model by incorporating information fusion across electrodes. Through the addition of a global signal path for multi-channel EEG analysis, the electrode level signal representations are informed by the overall seizure activity. By combining global and electrode level signal paths in one nueral network, SZLoc is capable of incorporating multi-scale information for accurate SOZ localization. SZLoc is trained in a mutli-task learning paradigm using an ensemble of weakly supervised loss functions. These losses encourage the network to produce outputs conducive to seizure localization. Using the information rich outputs of SZLoc, seizure activity can be tracked, EEG segments contributing to seizure onset can be identified, and localization maps can be generated all from a single network.

The graphical modeling and neural network based strategies provided here leverage the advantages of each modeling strategy. Graphical models allow structured predictions to be rigidly defined. Thus, graphical models can naturally incorporate clinical hypotheses of seizure spreading easily. In contrast, neural networks are black box architectures which typically produce only classification outputs. Interpreting these outputs is difficult, as the complicated decision functions learned by neural network are opaque to current methods of analysis. In the methods presented here, we circumvent this problem by restricting the outputs of our networks to be interpretable intermediate results. By aggregating these results, the networks can be trained for complicated detection tasks while maintaining interpretability.

Future work may leverage the strengths and weaknesses of these approaches for better hybridizations of graphical models and neural networks. While the R-SMMPL model effectively combined hierchical information to directly produce patient level maps of SOZ, its performance was fundamentally limited by the neural network likelihoods used in the model. As these likelihoods were trained without recurrent components and ommitting the

aggregation techniques developed in Chapters 6 and 7 for SZTrack and SZLoc, they exhibited far greater noise in their predictions than the later developed methods. By combining the R-SMMPL with networks trained using ideas from SZLoc, more accurate structured predictions may be possible than can be generated from either architecture alone.

Recent advancements in neural network training have been shown to improve the quality of neural network predictions. Specifically, using semi-supervised learning, networks can be trained to produce representations of unlabled data. Networks pretrained using semi-supervised techniques have been shown to outperform identical networks trained exclusively for downstream tasks. As such, semi-supervised training may be appropriate for pre-training future networks to extract powerful representations of the EEG signal.

Similarly, transformer models have been used effectively on sequential and time-series data. While the auto-regressive nature of RNN models closely mirrors the generative processes of the EEG signal, transformer architectures allow deeper networks to be trained more efficiently using more data. In conjunction with semi-supervised approaches, transformers for EEG sequence modeling may allow for further improvements in seizure analysis applications.

In this thesis, an array of approaches to the challenging task of seizure detection, tracking, and localization have been described. While these approaches explore a wide variety of modeling techniques, common to all is the ability to identify seizure activity at the level of individual electrodes and track this activity as seizures propagate. Each model seeks to capture clinically informed seizure spreading phenomena using the advantages of the modeling framework employed. The methods presented extend beyond the traditional seizure detection paradigm to provide clinicians with clinically relevant information useful in diagnosis and theraputic planning. Taken together, this work represents an exploration of potential avenues towards the development of automated tools for assisting in the annotation of scalp EEG in patients with focal epilepsy.

References

1. Miller, J. W. & Goodkin, H. P. *Epilepsy* (Wiley, 2014).
2. French, J. A. Refractory epilepsy: clinical overview. *Epilepsia* **48,** 3–7 (2007).
3. Lüders, H. O., Najm, I., Nair, D., Widdess-Walsh, P. & Bingman, W. The epileptogenic zone: general principles. *Epileptic disorders* **8,** 1–9 (2006).
4. Theodore, W. H. *Presurgical focus localization in epilepsy: PET and SPECT* in *Seminars in nuclear medicine* **47** (2017), 44–53.
5. Duncan, J. S., Winston, G. P., Koepp, M. J. & Ourselin, S. Brain imaging in the assessment for epilepsy surgery. *The Lancet Neurology* **15,** 420–433 (2016).
6. Craley, J., Johnson, E. & Venkataraman, A. *A novel method for epileptic seizure detection using coupled hidden markov models* in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2018), 482–489.
7. Craley, J., Johnson, E. & Venkataraman, A. A Spatio-Temporal Model of Seizure Propagation in Focal Epilepsy. *IEEE Transactions on Medical Imaging,* 1–1 (2019).
8. Craley, J., Johnson, E. & Venkataraman, A. *Integrating convolutional neural networks and probabilistic graphical modeling for epileptic seizure detection in multichannel EEG* in *International Conference on Information Processing in Medical Imaging* (2019), 291–303.
9. Craley, J., Johnson, E., Jouny, C. & Venkataraman, A. *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* (2019), 253–261.
10. Craley, J. *et al.* Automated seizure activity tracking and onset zone localization from scalp EEG using deep neural networks. *PLOS ONE* **In Review** (2021).
11. Craley, J., Johnson, E., Jouny, C. & Venkataraman, A. Automated inter-patient seizure detection using multichannel Convolutional and Recurrent Neural Networks. *Biomedical Signal Processing and Control* **64,** 102360 (2021).
12. Haas, L. F. Hans berger (1873–1941), richard caton (1842–1926), and electroencephalography. *Journal of Neurology, Neurosurgery & Psychiatry* **74,** 9–9 (2003).
13. Krauss, G. L. & Fisher, R. S. *The Johns Hopkins atlas of digital EEG: an interactive training guide* (Johns Hopkins University Press, 2006).
14. Marcuse, L. V., Fields, M. C. & Yoo, J. J. *Rowan's Primer of EEG E-Book* (Elsevier Health Sciences, 2015).
15. Jurcak, V. *et al.* 10/20, 10/10, and 10/5 systems revisited: their validity as relative head-surface-based positioning systems. *Neuroimage* **34,** 1600–1611 (2007).

16. Urigüen, J. A. & Garcia-Zapirain, B. EEG artifact removal—state-of-the-art and guidelines. *Journal of neural engineering* **12,** 031001 (2015).

17. Fisher, R. S. *et al.* ILAE official report: a practical clinical definition of epilepsy. *Epilepsia* **55,** 475–482 (2014).

18. Zach, M. & et al. National and State Estimates of the Numbers of Adults and Children with Active Epilepsy – United States, 2015. *CDC MMWR* **66,** 821–825 (31 2017).

19. Organization, W. H. *Epilepsy Fact Sheet* https://www.who.int/news-room/fact-sheets/detail/epilepsy. Accessed: 2010-03-10.

20. Rosenow, F. & Lüders, H. Presurgical evaluation of epilepsy. *Brain* **124,** 1683–1700 (2001).

21. Wilson, S. B. & Emerson, R. Spike detection: a review and comparison of algorithms. *Clinical Neurophysiology* **113,** 1873–1881 (2002).

22. Van Donselaar, C. A., Schimsheimer, R.-J., Geerts, A. T. & Declerck, A. C. Value of the electroencephalogram in adult patients with untreated idiopathic first seizures. *Archives of neurology* **49,** 231–237 (1992).

23. Cascino, G. D. Video-EEG monitoring in adults. *Epilepsia* **43,** 80–93 (2002).

24. Tufenkjian, K. & Lüders, H. O. Seizure semiology: its value and limitations in localizing the epileptogenic zone. *Journal of Clinical Neurology* **8,** 243–250 (2012).

25. Ghougassian, D. F., d'Souza, W., Cook, M. J. & O'Brien, T. J. Evaluating the utility of inpatient video-EEG monitoring. *Epilepsia* **45,** 928–932 (2004).

26. Ryvlin, P., Cross, J. H. & Rheims, S. Epilepsy surgery in children and adults. *The Lancet Neurology* **13,** 1114–1126 (2014).

27. Jayakar, P. *et al.* Diagnostic test utilization in evaluation for resective epilepsy surgery in children. *Epilepsia* **55,** 507–518 (2014).

28. Téllez-Zenteno, J. F., Ronquillo, L. H., Moien-Afshari, F. & Wiebe, S. Surgical outcomes in lesional and non-lesional epilepsy: a systematic review and meta-analysis. *Epilepsy research* **89,** 310–318 (2010).

29. Struck, A. F., Hall, L. T., Floberg, J. M., Perlman, S. B. & Dulli, D. A. Surgical decision making in temporal lobe epilepsy: A comparison of [18F] FDG-PET, MRI, and EEG. *Epilepsy & Behavior* **22,** 293–297 (2011).

30. Knake, S. *et al.* The value of multichannel MEG and EEG in the presurgical evaluation of 70 epilepsy patients. *Epilepsy research* **69,** 80–86 (2006).

31. Barkley, G. L. & Baumgartner, C. MEG and EEG in epilepsy. *Journal of clinical neurophysiology* **20,** 163–178 (2003).

32. Fuchs, M., Wagner, M., Kohler, T. & Wischmann, H. Linear and Nonlinear Current Density Reconstructions. *J Clinical Neurophysiology* **16,** 267–295 (1999).

33. Fuchs, M., Ford, M., Sands, S. & Lew, H. Overview of Dipole Source Localization. *Phys Med Rehabil Clin N Am* **15,** 251–262 (2004).

34. Wang, J., Williamson, S. & Kaufman, L. Magnetic Source Imaging Based on the Minimum-Norm Least-Squares Inverse. *Brain Topography* **5,** 365–371 (1993).

35. Grech, R. *et al.* Review on Solving the Inverse Problem in EEG Source Analysis. *Journal of NeuroEngineering and Rehabilitation* **5,** 25 (2008).

36. Gorodnitsky, I. & Rao, B. Sparse Signal Reconstruction from Limited Data Using FOCUSS: A Re-weighted Minimum Norm Algorithm. *IEEE Transactions on Signal Processing* **45** (1997).

37. Cuffin, B. EEG Localization Accuracy Improvements using Realistically Shaped Head Models. *IEEE Trans Biomedical Engineering* **43,** 299–393 (1996).

38. Crouzeix, A., Yvert, B., Bertrand, O. & Pernier, J. An Evaluation of Dipole Reconstruction Accuracy with Spherical and Realistic Head Models in MEG. *Clinical Neurophysiology* **110,** 2176–2188 (1999).

39. Foged, M. T. *et al.* Diagnostic added value of electrical source imaging in presurgical evaluation of patients with epilepsy: a prospective study. *Clinical Neurophysiology* **131,** 324–329 (2020).

40. Gotman, J. Automatic recognition of epileptic seizures in the EEG. *Electroencephalography and clinical Neurophysiology* **54,** 530–540 (1982).

41. Andrzejak, R. G. *et al.* Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E* **64,** 061907 (2001).

42. Osorio, I., Zaveri, H. P., Frei, M. G. & Arthurs, S. *Epilepsy: the intersection of neurosciences, biology, mathematics, engineering, and physics* (CRC press, 2016).

43. Hopfengärtner, R. *et al.* Automatic seizure detection in long-term scalp EEG using an adaptive thresholding technique: a validation study for clinical routine. *Clinical Neurophysiology* **125,** 1346–1352 (2014).

44. Hopfengärtner, R., Kerling, F., Bauer, V. & Stefan, H. An efficient, robust and fast method for the offline detection of epileptic seizures in long-term scalp EEG recordings. *Clinical Neurophysiology* **118,** 2332–2343 (2007).

45. Shoeb, A. H. & Guttag, J. V. *Application of machine learning to epileptic seizure detection* in *International Conference on Machine Learning* (2010), 975–982.

46. Zandi, A. S. *et al.* Automated real-time epileptic seizure detection in scalp EEG recordings using an algorithm based on wavelet packet transform. *IEEE Transactions on Biomedical Engineering* **57,** 1639–1651 (2010).

47. Kaleem, M., Guergachi, A. & Krishnan, S. Patient-specific seizure detection in long-term EEG using wavelet decomposition. *Biomedical Signal Processing and Control* **46,** 157–165 (2018).

48. Bhattacharyya, A. & Pachori, R. B. A multivariate approach for patient-specific EEG seizure detection using empirical wavelet transform. *IEEE Transactions on Biomedical Engineering* **64,** 2003–2015 (2017).

49. Faust, O., Acharya, U. R., Adeli, H. & Adeli, A. Wavelet-based EEG processing for computer-aided seizure detection and epilepsy diagnosis. *Seizure* **26,** 56–64 (2015).

50. Acharya, U. R. *et al.* Automated diagnosis of epileptic EEG using entropies. *Biomedical Signal Processing and Control* **7,** 401–408 (2012).

51. Wu, D. *et al.* Automatic Epileptic Seizures Joint Detection Algorithm Based on Improved Multi-Domain Feature of cEEG and Spike Feature of aEEG. *IEEE Access* **7,** 41551–41564 (2019).

52. Ghosh-Dastidar, S., Adeli, H. & Dadmehr, N. Mixed-band wavelet-chaos-neural network methodology for epilepsy and epileptic seizure detection. *IEEE transactions on biomedical engineering* **54,** 1545–1551 (2007).

53. Adeli, H., Ghosh-Dastidar, S. & Dadmehr, N. A wavelet-chaos methodology for analysis of EEGs and EEG subbands to detect seizure and epilepsy. *IEEE Transactions on Biomedical Engineering* **54,** 205–211 (2007).

54. Ocak, H. Automatic detection of epileptic seizures in EEG using discrete wavelet transform and approximate entropy. *Expert Systems with Applications* **36,** 2027–2036 (2009).

55. Bandarabadi, M., Teixeira, C. A., Rasekhi, J. & Dourado, A. Epileptic seizure prediction using relative spectral power features. *Clinical Neurophysiology* **126,** 237–248 (2015).

56. Sridevi, V. *et al.* Improved Patient-Independent System for Detection of Electrical Onset of Seizures. *Journal of Clinical Neurophysiology* **36,** 14 (2019).

57. Murphy, K. P. Machine learning: a probabilistic perspective (2012).

58. Alickovic, E., Kevric, J. & Subasi, A. Performance evaluation of empirical mode decomposition, discrete wavelet transform, and wavelet packed decomposition for automated epileptic seizure detection and prediction. *Biomedical signal processing and control* **39,** 94–102 (2018).

59. Breiman, L. Random forests. *Machine learning* **45,** 5–32 (2001).

60. Jelinek, F. *Statistical methods for speech recognition* (MIT press, 1997).

61. Esmaeili, S., Araabi, B., Soltanian-Zadeh, H. & Schwabe, L. *Variational Bayesian learning for Gaussian mixture HMM in seizure prediction based on long term EEG of epileptic rats* in *2014 21th Iranian conference on biomedical engineering (ICBME)* (2014), 138–143.

62. Dash, D. P., Kolekar, M. H. & Jha, K. Multi-channel EEG based automatic epileptic seizure detection using iterative filtering decomposition and Hidden Markov Model. *Computers in biology and medicine* **116,** 103571 (2020).

63. Direito, B. *et al.* Modeling epileptic brain states using EEG spectral analysis and topographic mapping. *Journal of neuroscience methods* **210,** 220–229 (2012).

64. Wulsin, D. F., Fox, E. B. & Litt, B. Modeling the complex dynamics and changing correlations of epileptic events. *Artificial intelligence* **216,** 55–75 (2014).

65. Nefian, A. V. *et al. A coupled HMM for audio-visual speech recognition* in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on* **2** (2002), II–2013.

66. Dong, W., Pentland, A. & Heller, K. A. Graph-coupled hmms for modeling the spread of infection. *arXiv preprint arXiv:1210.4864* (2012).

67. Rezek, I. & Roberts, S. J. *Estimation of coupled hidden Markov models with application to biosignal interaction modelling* in *Neural Networks for Signal Processing X, 2000. Proceedings of the 2000 IEEE Signal Processing Society Workshop* **2** (2000), 804–813.

68. Zhong, S. & Ghosh, J. *HMMs and coupled HMMs for multi-channel EEG classification* in *Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on* **2** (2002), 1154–1159.

69. Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review* **65,** 386 (1958).

70. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25,** 1097–1105 (2012).

71. Hinton, G. & et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine* **29,** 82–97 (2012).

72. LeCun, Y. *et al.* Backpropagation applied to handwritten zip code recognition. *Neural computation* **1,** 541–551 (1989).

73. Russakovsky, O. *et al.* Imagenet large scale visual recognition challenge. *International journal of computer vision* **115,** 211–252 (2015).

74. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

75. He, K., Zhang, X., Ren, S. & Sun, J. *Deep residual learning for image recognition* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), 770–778.

76. Van den Oord, A. *et al.* WaveNet: A Generative Model for Raw Audio. *CoRR* **abs/1609.03499.** arXiv: 1609.03499 (2016).

77. Hammer, B. On the approximation capability of recurrent neural networks. *Neurocomputing* **31,** 107–123 (2000).

78. Schuster, M. & Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* **45,** 2673–2681 (1997).

79. Hochreiter, S. & Schmidhuber, J. Long Short-Term Memory. *Neural Computation* **9,** 1735–1780. eprint: https://doi.org/10.1162/neco.1997.9.8.1735 (1997).

80. Graves, A. & Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks* **18,** 602–610 (2005).

81. Huang, Z., Xu, W. & Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).

82. Hu, X. *et al.* Scalp EEG classification using deep Bi-LSTM network for seizure detection. *Computers in Biology and Medicine* **124,** 103919 (2020).

83. Cho, K. *et al.* Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

84. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

85. Vaswani, A. *et al. Attention is all you need* in *Advances in neural information processing systems* (2017), 5998–6008.

86. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

87. Parmar, N. *et al. Image transformer* in *International Conference on Machine Learning* (2018), 4055–4064.

88. Qi, D. *et al.* Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966* (2020).

89. Ba, J. L., Kiros, J. R. & Hinton, G. E. *Layer Normalization* 2016. arXiv: 1607.06450 [stat.ML].

90. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *nature* **323,** 533–536 (1986).

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

92. Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012).

93. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *Journal of Big Data* **6,** 1–48 (2019).

94. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521,** 436–444 (2015).

95. Craik, A., He, Y. & Contreras-Vidal, J. L. Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of neural engineering* **16,** 031001 (2019).

96. Tăuţan, A.-M., Dogariu, M. & Ionescu, B. *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)* (2019), 2377–2381.

97. Yuan, Y., Xun, G., Jia, K. & Zhang, A. A Multi-View Deep Learning Framework for EEG Seizure Detection. *IEEE journal of biomedical and health informatics* **23,** 83–94 (2018).

98. Gao, Y., Gao, B., Chen, Q., Liu, J. & Zhang, Y. Deep Convolutional Neural Network-Based Epileptic Electroencephalogram (EEG) Signal Classification. *Frontiers in Neurology* **11** (2020).

99. Khan, H., Marcuse, L., Fields, M., Swann, K. & Yener, B. Focal onset seizure prediction using convolutional networks. *IEEE Transactions on Biomedical Engineering* **65,** 2109–2118 (2017).

100. Taherisadr, M., Joneidi, M. & Rahnavard, N. *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)* (2019), 1–6.

101. Wei, Z., Zou, J., Zhang, J. & Xu, J. Automatic epileptic EEG detection using convolutional neural network with improvements in time-domain. *Biomedical Signal Processing and Control* **53,** 101551 (2019).

102. Zou, L., Liu, X., Jiang, A. & Zhousp, X. *Epileptic Seizure Detection Using Deep Convolutional Network* in *2018 IEEE 23rd International Conference on Digital Signal Processing (DSP)* (2018), 1–4.

103. 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).

104. Vidyaratne, L., Glandon, A., Alam, M. & Iftekharuddin, K. M. *Deep recurrent neural network for seizure detection* in *2016 International Joint Conference on Neural Networks (IJCNN)* (2016), 1202–1207.

105. Affes, A., Mdhaffar, A., Triki, C., Jmaiel, M. & Freisleben, B. *A Convolutional Gated Recurrent Neural Network for Epileptic Seizure Prediction* in *International Conference on Smart Homes and Health Telematics* (2019), 85–96.

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

107. Ayodele, K., Ikezogwo, W., Komolafe, M. & Ogunbona, P. Supervised domain generalization for integration of disparate scalp EEG datasets for automatic epileptic seizure detection. *Computers in Biology and Medicine,* 103757 (2020).

108. Goldberger, A. L. *et al.* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101,** e215–e220 (2000).

109. Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. An introduction to variational methods for graphical models. *Machine learning* **37,** 183–233 (1999).

110. Nocedal, J. & Wright, S. J. *Numerical Optimization* (Springer, 1999).

111. Esteller, R., Echauz, J., Tcheng, T., Litt, B. & Pless, B. *Line length: an efficient feature for seizure onset detection* in *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE* **2** (2001), 1707–1710.

112. Logesparan, L., Casson, A. J. & Rodriguez-Villegas, E. Optimal features for online seizure detection. *Medical & Biological Engineering & Computing* **50,** 659–669 (July 2012).

113. Shoaran, M., Farivar, M. & Emami, A. *Hardware-friendly seizure detection with a boosted ensemble of shallow decision trees* in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2016), 1826–1829.

114. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).

115. Kschischang, F. R. & et al. Factor graphs and the sum-product algorithm. *IEEE Transactions on information theory* **47,** 498–519 (2001).

116. Zhou, J. *et al.* Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434* (2018).

117. Such, F. P. *et al.* Robust spatial filtering with graph convolutional neural networks. *IEEE Journal of Selected Topics in Signal Processing* **11,** 884–896 (2017).

118. Wagh, N. & Varatharajah, Y. *EEG-GCNN: Augmenting Electroencephalogram-based Neurological Disease Diagnosis using a Domain-guided Graph Convolutional Neural Network* in *Machine Learning for Health* (2020), 367–378.

119. Covert, I. *et al.* Temporal graph convolutional networks for automatic seizure detection. *arXiv preprint arXiv:1905.01375* (2019).

120. Lian, Q., Qi, Y., Pan, G. & Wang, Y. Learning graph in graph convolutional neural networks for robust seizure prediction. *Journal of neural engineering* **17,** 035004 (2020).

121. Yin, Y., Zheng, X., Hu, B., Zhang, Y. & Cui, X. EEG emotion recognition using fusion model of graph convolutional neural networks and LSTM. *Applied Soft Computing* **100,** 106954 (2021).

122. Li, X. *et al. Classify EEG and reveal latent graph structure with spatio-temporal graph convolutional neural network* in *2019 IEEE International Conference on Data Mining (ICDM)* (2019), 389–398.

123. Yan, S., Xiong, Y. & Lin, D. *Spatial temporal graph convolutional networks for skeleton-based action recognition* in *Proceedings of the AAAI conference on artificial intelligence* **32** (2018).

124. Maas, A. L., Hannun, A. Y., Ng, A. Y., *et al. Rectifier nonlinearities improve neural network acoustic models* in *Proc. icml* **30** (2013), 3.

125. Ioffe, S. & Szegedy, C. *Batch normalization: Accelerating deep network training by reducing internal covariate shift* in *International conference on machine learning* (2015), 448–456.

126. Brazier, M. A. Spread of seizure discharges in epilepsy: anatomical and electrophysiological considerations. *Experimental neurology* **36,** 263–272 (1972).

127. Grinenko, O. *et al.* A fingerprint of the epileptogenic zone in human epilepsies. *Brain* **141,** 117–131 (2018).

128. Graves, A., Fernández, S. & Schmidhuber, J. *Bidirectional LSTM networks for improved phoneme classification and recognition* in *International conference on artificial neural networks* (2005), 799–804.

129. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

130. Karita, S. *et al. A comparative study on transformer vs rnn in speech applications* in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (2019), 449–456.

131. Cisotto, G. *et al.* Comparison of Attention-based Deep Learning Models for EEG Classification. *arXiv preprint arXiv:2012.01074* (2020).

132. Qu, W. *et al.* A residual based attention model for eeg based sleep staging. *IEEE journal of biomedical and health informatics* **24,** 2833–2843 (2020).

133. Kostas, D., Aroca-Ouellette, S. & Rudzicz, F. BENDR: using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *arXiv preprint arXiv:2101.12037* (2021).

134. Sun, J., Xie, J. & Zhou, H. *EEG Classification with Transformer-Based Models* in *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)* (2021), 92–93.

135. Song, Y., Jia, X., Yang, L. & Xie, L. Transformer-based Spatial-Temporal Feature Learning for EEG Decoding. *arXiv preprint arXiv:2106.11170* (2021).

136. Liu, J., Zhang, L., Wu, H. & Zhao, H. Transformers for EEG Emotion Recognition. *arXiv preprint arXiv:2110.06553* (2021).

137. Bagchi, S. & Bathula, D. R. EEG-ConvTransformer for Single-Trial EEG based Visual Stimuli Classification. *arXiv preprint arXiv:2107.03983* (2021).

138. Schölzel, C. *Nonlinear measures for dynamical systems* version 0.5.2. June 2019.

139. Shoeb, A. H. *Application of machine learning to epileptic seizure onset detection and treatment* PhD thesis (Massachusetts Institute of Technology, 2009).

140. Qu, H. & Gotman, J. A seizure warning system for long-term epilepsy monitoring. *Neurology* **45,** 2250–2254 (1995).

141. Nagaraj, S. B., Stevenson, N. J., Marnane, W. P., Boylan, G. B. & Lightbody, G. Neonatal seizure detection using atomic decomposition with a novel dictionary. *IEEE Transactions on Biomedical Engineering* **61,** 2724–2732 (2014).

Appendix I

CNN-BLSTM Hybrid for Deep Seizure Detection

I.1 Introduction

In this appendix 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, including work presented in Chapters 3 and 4 of this thesis, have focused on one-dimensional CNNs applied individually to each EEG channel [8, 9, 101–103]. This approach ignores clinically relevant cross-channel phase information, as seizures are often characterized by atypical synchronization between channels [13, 14]. 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 seconds [105–107]. The RNN then provides a 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 the JHH 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 [101], and it can easily be integrated into the existing clinical infrastructure.

I.2 Materials and Methods

I.2.1 EEG Data and Preprocessing

We validate our model on the JHH dataset of 34 patients. For each seizure, we include a maximum of 10 minutes of pre-seizure and post-seizure baseline. While the variety in patient representations in the JHH dataset 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.

I.2.2 An End-to-End Detection Framework

I.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 Figure I-1. In the

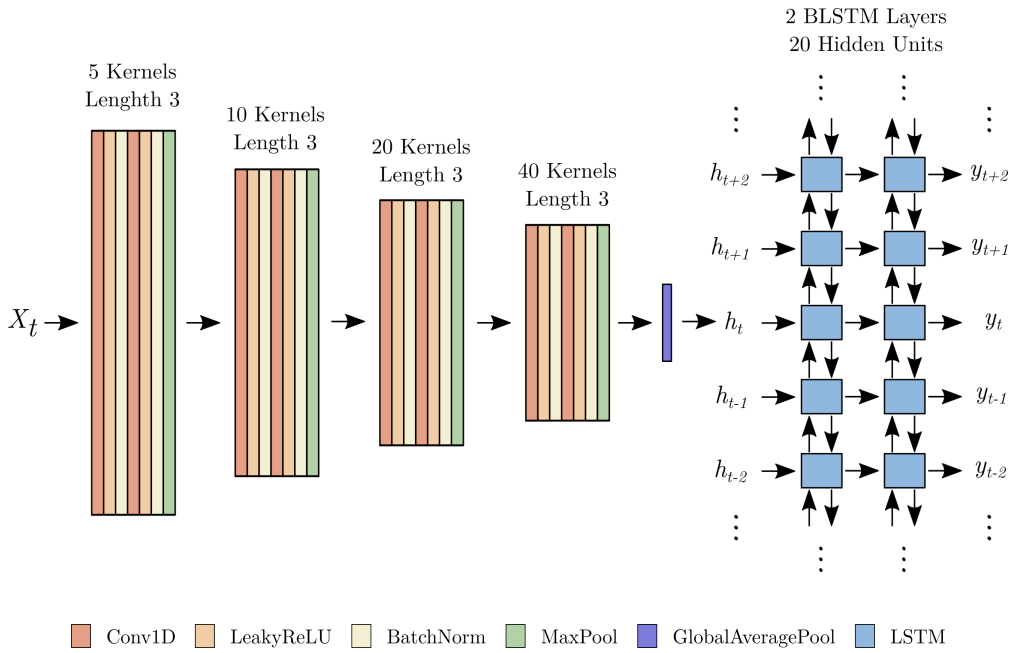


Figure I-1. Our CNN-BLSTM architecture for inter-patient seizure detection. A convolutional encoder converts EEG signal X_t to hidden representations h_t. These representations are classified by a two layer BLSTM to predict seizure labels y_t.

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 Figure I-1. 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 [74], 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

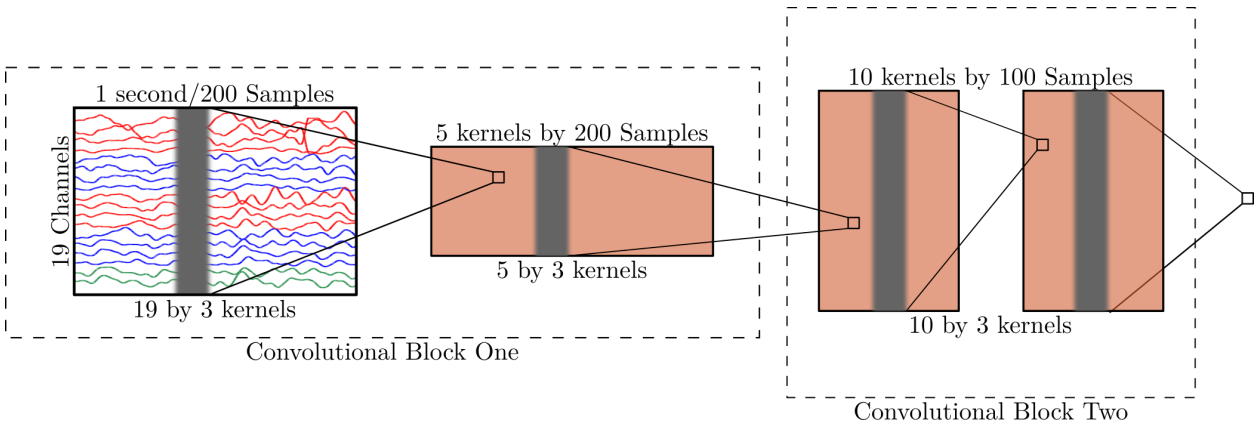


Figure I-2. The first two convolutional blocks of the CNN encoder. One second of preprocessed EEG signal is fed directly into the first layer of the CNN. An example input for each convolution is shown in gray while the corresponding output of the convolution is shown in the next layer as a square. Each 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.

roughly the same amount of computation. The convolution and max pooling procedures are illustrated in Figure I-2. 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, 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.

I.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 minutes 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.

I.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 vanishing and exploding gradient problem [94]. 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 31 windows, a learning rate of 0.01, and the ADAM optimizer [91]. 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 [91]. 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.

I.2.3 Baseline Comparison Methods

I.2.3.1 Feature Based Classification

Our first set of baseline methods employs the two-stage feature selection and classification pipeline discussed in Section 2.2. 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 [50] and [52, 53], 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 [138]. 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 single seconds of EEG signal, we apply the same temporal smoothing and calibration described in I.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 second 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.

I.2.3.2 Convolutional Models

We implement the CNN network from Wei et al. [101] (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 [108, 139]. 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 Figure I-3, 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

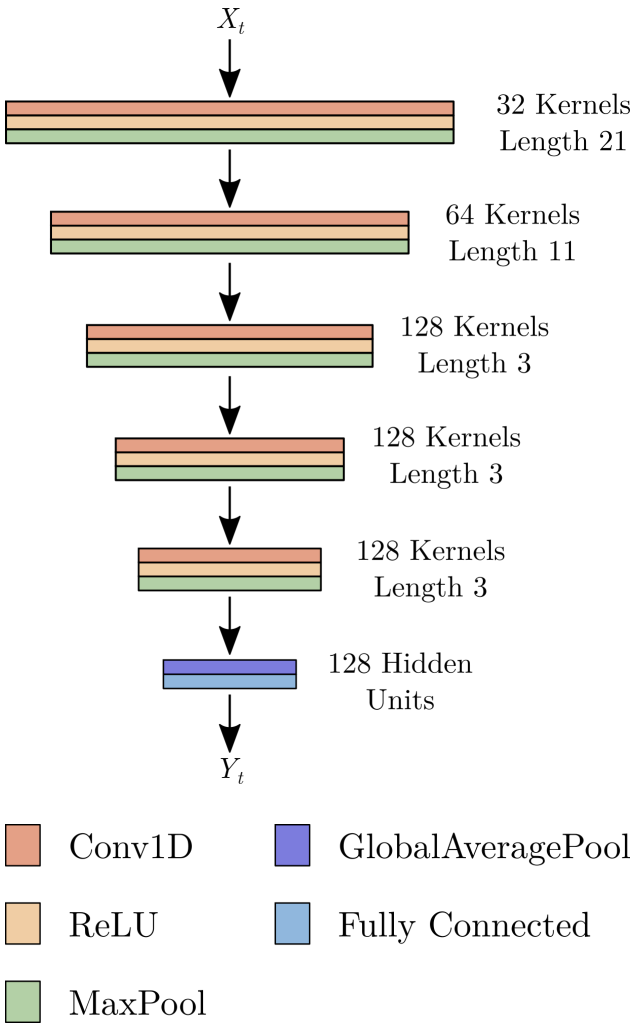


Figure I-3. Wei-CNN baseline Architecture

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 I.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 Figure I-4. The EEG signal is windowed into one second non-overlapping segments and an FFT is calculated.

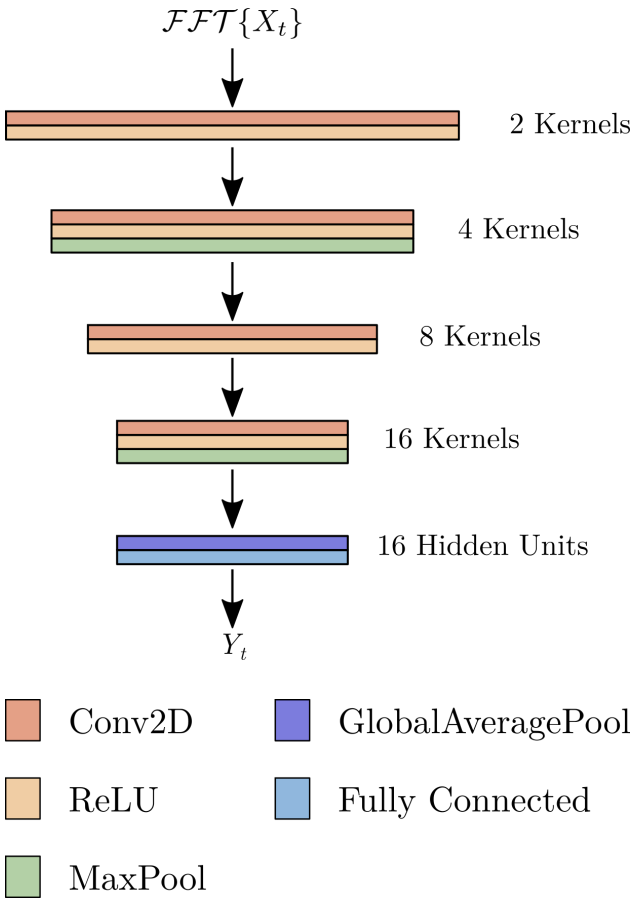


Figure I-4. CNN-2D FFT image baseline architecture.

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 [97, 99, 100] where time-frequency decompositions are used in conjunction with 2D convolutions.

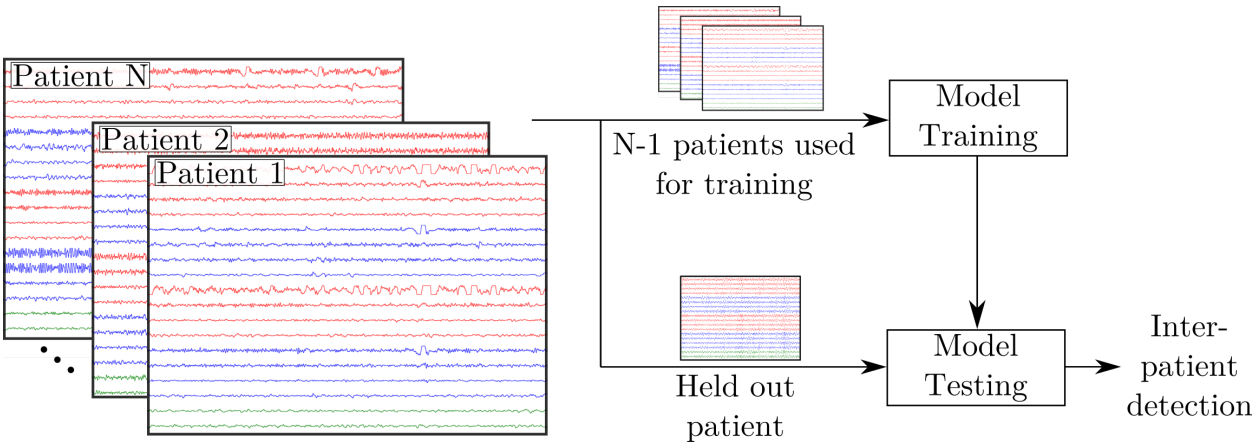


Figure I-5. Cross validation procedure, in which one patient is left-out for testing while the rest of the dataset is used for training. This procedure is repeated for each patient and the performance is averaged across all N folds.

I.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* [140]. 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 Figure I-5. 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 [141] 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.

I.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.

At the seizure level, we consider contiguous seizure classifications produced by each model. Namely, if the model prediction exceeds the threshold determined in Section I.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.

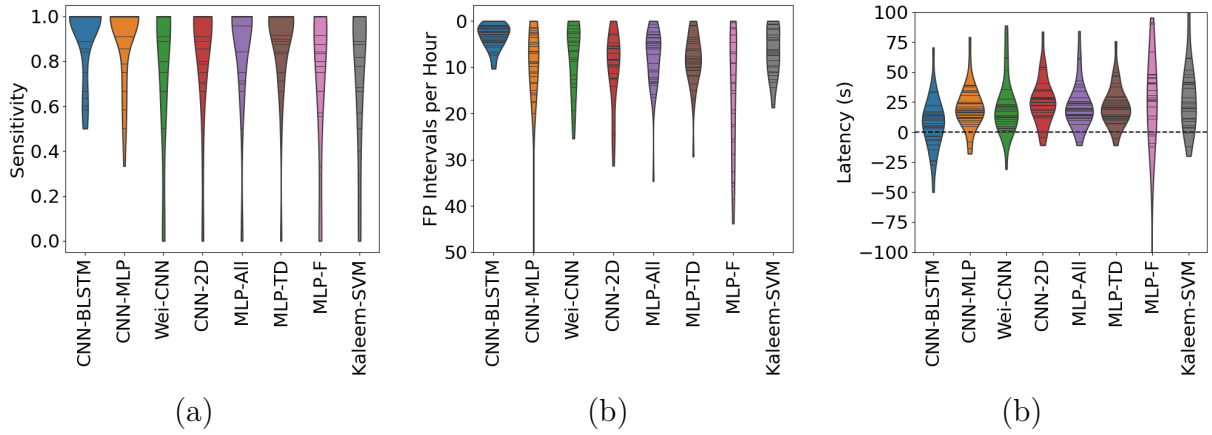
Table I-I. IID Window Level Results

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

I.3 Experimental Results

I.3.1 Window Level Accuracy

Table I-I 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



(a) (b) (b)

Figure I-6. 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.

0.7054 and 0.4304, retrospectively.

In addition, Table I-I 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 0.9531. All models exhibit specificities above 0.9, with the CNN-MLP achieving the highest specificity of 0.9555.

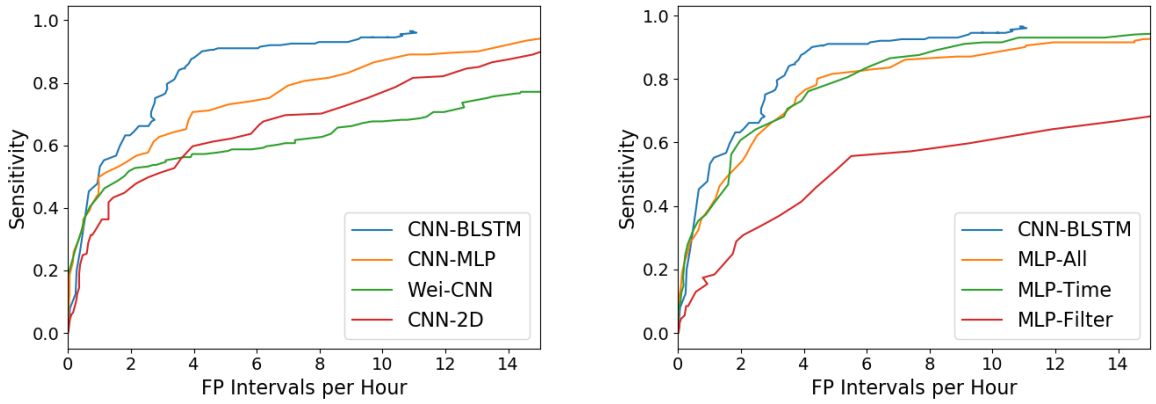
I.3.2 Seizure Level Results

Figure I-6 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 Figure I-6 (a) we see that the CNN-BLSTM maintains high sensitivity across the dataset, while baseline models fail to generalize to some patients. In addition, Figure I-6 (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

I-II and I-III in the appendix report the patient-specific performance metrics. Table I-II shows performance for the CNN-based models (CNN-BLSTM, CNN-MLP, Wei-CNN, CNN-2D) while Table I-III 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/hr. The CNN-MLP, Wei-CNN, and CNN-2D all exhibit lower sensitivities at 0.90, 0.77, and 0.84, respectively, and FPs/hr of 9.6, 7.5, and 10.2, respectively. Thus the CNN-BLSTM achieves the highest sensitivity with the lowest false positive rate. Finally, Figure I-6 (c) shows the spread of onset latencies for each model. In the CNN-BLSTM onset latency is distributed around 10 seconds while other models report a higher average latency.

Sensitivity versus FPR plots are shown in Figure I-7, 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 I-II and I-III. 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.

Figure I-8 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



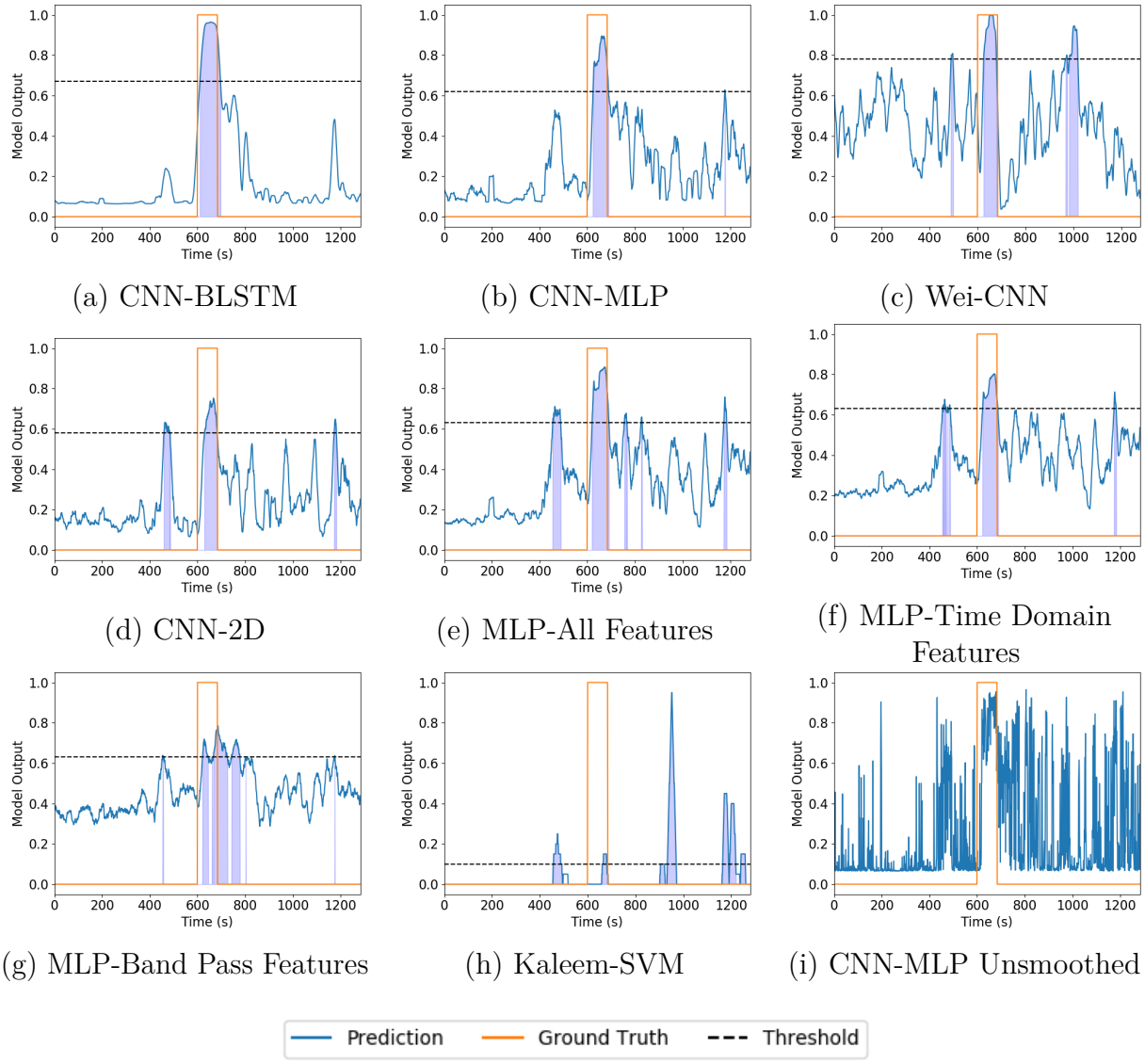
(a) End-to-End Baselines (b) Feature-Based MLP Baselines

Figure I-7. Sensitivity versus false positive rate curves for each model. The metrics are calculated as the seizure detection threshold is swept is swept from 0 to 1 for each patient. The threshold sweep is performed globally and not calibrated separately for each patient.

recording occurring past the offset annotation.

The prediction output of baseline methods are shown in Figure I-8 (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 Figure I-8 (c), where the Wei-CNN makes three spurious false positive detections throughout the recording. Figure I-8 (i) shows the unsmoothed prediction output for the CNN-MLP model. When comparing this image to the smoothed CNN-MLP output in Figure I-8 (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 Figure I-9. 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 seconds after the start of the recording, in the figure indicated by the vertical dashed black line. The CNN-BLSTM detects the seizure at



(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

Figure I-8. 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.

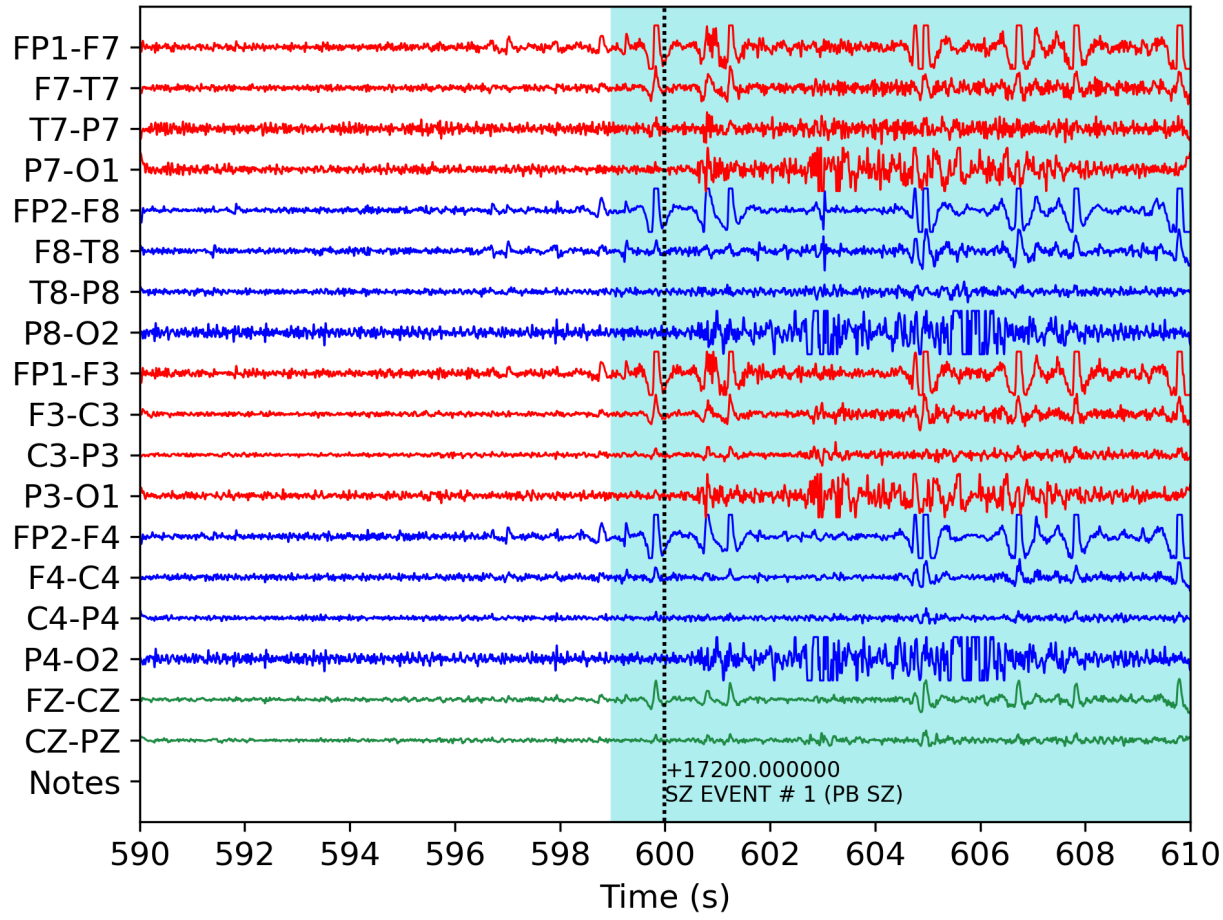


Figure I-9. EEG recording and CNN-BLSTM classification corresponding to Figure I-8 (a). Seizure onset annotation is depicted by the vertical dashed line at 600 seconds. CNN-BLSTM seizure classification is shown shaded in light blue. The CNN-BLSTM declares the onset of a seizure at 599 seconds, in accordance with the clinical annotation.

599 seconds, 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.

I.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 I-II 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/hr during testing was 3.3.

When examining the output of an individual model in Figure I-8 (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 Figure I-8 (b). However, we note that our BLSTM network 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 Figure I-8 (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 Figure I-6 (c), the average onset latency for the CNN-BLSTM is evenly

distributed around 10 seconds. 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 Figure I-9, 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 I-I 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 minute sample of EEG. Bandpass, FFT, line-length, and power features combined could be computed in less than 5 seconds. However, using a freely available Python package, the non-linear features sample entropy and LLE took roughly 60 and 310 seconds, respectively, far too long for use in a clinical environment. By comparison, the CNN-BLSTM took roughly 0.15 seconds 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.

I.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 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.

I.6 Results by Patient

Table I-II. JHH CNN Seizure Results by Patient

	CNN-BLSTM			CNN-MLP			Wei-CNN			CNN-2D		
Patient	FPs/hr	Sensitivity	Latency (s)	FPs/hr	Sensitivity	Latency (s)	FPs/hr	Sensitivity	Latency (s)	FPs/hr	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

Table I-III. JHH MLP Seizure Results by Patient

	MLP-All			MLP-Time Domain Features			MLP-Filterbank Features			KaleemSVM		
Patient	FPs/hr	Sensitivity	Latency (s)	FPs/hr	Sensitivity	Latency (s)	FPs/hr	Sensitivity	Latency (s)	FPs/hr	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