Improved Segmentation for Automated Seizure Detection  
using Channel-Dependent Posteriors

A Dissertation Proposal

Submitted to

the Temple University Graduate Board

In Partial Fulfillment  
of the Requirements for the Degree  
DOCTOR OF PHILOSOPHY

by

Vinit Shah

December 2020

Examining Committee Members:

Dr. Joseph Picone, Advisory Chair, Department of Electrical and Computer Engineering,

Dr. Iyad Obeid, Department of Electrical and Computer Engineering

Dr. Yimin Zhang, Department of Electrical and Computer Engineering

Dr. Pallavi Chitturi, Department of Statistics

Dr. Mercedes Jacobson, Department of Neurology

Dr. Yaroslav Koshka, External Reader, Mississippi State University

©

Copyright

2020

By

Vinit Shah

------------------------

All Rights Reserved

# ABSTRACT

Improved Segmentation for Automated Seizure Detection  
using Channel-Dependent Posteriors

Vinit Shah

PhD Proposal

Temple University, December 2020

Advisor: Dr. Joseph Picone

The Electroencephalogram (EEG) is the primary tool used for the diagnosis of a variety of neural pathologies such as epilepsy. Identification of a critical event, such as an epileptic seizure, is difficult because the signals are corrupted by noise due to the way they are transduced. Also, because EEGs are collected by measuring extremely low voltages, signals often contain artifacts due to patient movement. Additionally, factors such as slowly evolving morphologies make accurate marking of the onset and offset of a seizure event difficult. Precise segmentation, defined as the ability to detect start and stop times within a fraction of a second, is a challenging research problem. State-of-the-art commercial systems perform poorly on this task. In addition to generating precise segmentations, a real-time seizure detection algorithm is also expected to execute with minimal latency if it is to be useful in clinical applications. In this study, we improve seizure segmentation performance by mimicking the human interpretation process.

To identify a seizure event, experts explicitly look for the evolution of sharply contoured epileptiform features such as polyspikes, spike and wave discharges/complexes. However, evolution of such features below  Hz in frequency and with a duration exceeding 10 secs creates a significant amount of ambiguity. In such cases, decision-making requires additional factors be considered such as an EEG record’s clinical correlates, history of the patient and background signals. Incorporating the history of the waveform shapes and previously detected events aids understanding of the background morphologies. Moreover, seizure events can last from three seconds to several days in duration and can vary significantly both in morphology and between patients. For this study, we only include records which are at most one hour long.

The central thesis of this is work is that separation of the seizure detection problem into a two-phase problem – epileptiform activity detection followed by seizure detection – should improve our ability to detect and localize seizure events. We use linear frequency cepstral coefficients (LFCC) and their derived features, such as temporal derivatives, in both phases of the model. In the first phase, we use a sequential neural network algorithm known as a long short-term memory (LSTM) network to identify channel-specific epileptiform discharges associated with seizures. In the second phase, the LFCC-based feature vector is augmented with posteriors that represent the onset and offset of ictal activities. This augmented vector is applied to a multichannel convolutional neural network (CNN) followed by an LSTM network.

This second phase network is designed such that the number of input channels of the CNN is equal to the dimensionality of a feature vector. During the first layer, the CNN kernels of each channel are optimized separately which means the network optimization is done on each dimension of the feature vector independently. We observe that the kernel outputs, known as feature maps, show boundaries clustered together for easily detected seizures and dispersed for the subtle seizures. Once the feature maps of the first layer are created, the feature dimensions are no longer independent, which forces the remaining layers of the model to compete for localization of the seizure boundary. We aggregate the output of both phases of the model and perform additional postprocessing to achieve high performance for seizure detection.

The multiphase model was evaluated on a blind evaluation set and was shown to detect segment boundaries within a -second margin of error. Our previous best system, which delivers state-of-the-art performance on this task, correctly detected only 9 segment boundaries. Our multiphase system was also shown to be robust by performing well on two blind evaluation sets. Seizure detection performance on the TU Seizure Detection (TUSZ) Corpus development set is sensitivity with FAs/ hours. Performance on the corresponding evaluation set is sensitivity with FAs/ hours. Performance on a previously unseen corpus, the Duke University Seizure (DUSZ) Corpus is sensitivity with FAs/ hours. Our previous best system yields sensitivity with FAs/ hours on the TUSZ development set, sensitivity with 1 FAs/ hours on the TUSZ evaluation set and sensitivity with FAs/ hours on DUSZ.

The remaining research work will focus on several incremental improvements to this system. By projecting feature maps of the CNN kernels, we have shown that kernels associated with the posterior-derived features are able to learn the history of previously detected seizures. However, the importance of such features remains unknown. We will implement a permutation feature importance algorithm which will allow us to prioritize and learn the dependencies between the features through this process.

Our multiphase model is able to detect segment boundaries within a five-second margin of error. There is the potential to fine tune these segments by introducing an additional postprocessing step. We will use a 10-second context window around the detected boundaries and develop a rudimentary energy detector model and use a small LSTM model to find the seizure boundary. The energy detector will serve as a baseline which simply calculates the energy of the signal each second and finds optimal boundary based on a maximum likelihood detector to maximize the energy difference between window’s left and the right context.

Our multiphase model uses a fixed duration sliding window approach. This technique prevents the model from learning the evolution of EEG events which could vastly vary in duration. For example, a phase 2 model which operates on 11-second windows cannot learn a complete evolution of a 300‑second long seizure. This can negatively impact our ability to detect longer seizures especially when they are slowly evolving. We will implement an off-line seizure detection algorithm which can be trained on the entire seizure segment regardless of its duration. This is expected to reduce the false alarm rate.

Finally, we will analyze the model’s deficiencies on specific types of patients and investigate if there are any common underlying electrographic patterns and/or pathological reasons that explain poor segmentation performance.

Improving seizure detection performance through better segmentation is an important step forward in making automated seizure detection systems clinically acceptable. The primary failure modality for these systems is segmentation. For a real-time system, accurate segmentation will allow clinicians detect a seizure as soon as it appears in the EEG signal. This will allow neurologists to take action during the early stages of the event which, in many cases, is essential to avoid permanent damage to the brain. In a similar way, accurate offset detection will help with delivery of therapies designed to mitigate postictal (after seizure) period symptoms. This will also help reveal the severity of a seizure and as a consequence, provide guidance for medicating a patient.

# Dedication

Dedication goes here…

# ACKNOWLEDGMENTS

Acknowledgements go here…

# Table of Contents

ABSTRACT

Dedication  viii

ACKNOWLEDGMENTS ix

1. INTRODUCTION  1
2. THE ELECTROENCEPHALOGRAM  13
   1. The Electroencephalogram’s Role in Neurology
   2. Normal and Abnormal EEGs
   3. The TCP Montage and 10/20 Electrode Placement
3. THE CHALLENGES OF EEG INTERPRETATION  20
   1. Contemporary Visualization Tools
   2. ACNS Standards for Seizure Identification
   3. Ictal Discharges and EEG Gray Zone Events
   4. Interrater Agreement Among Experts
4. NEURAL NETWORK MODELS  33
   1. Relevant Neural Network Architectures
      1. The Multilayer Perceptron
      2. Convolutional Neural Networks
      3. Long Short-Term Memory Networks
   2. Features
   3. A Multiphase Recognition System
      1. Data Flow Considerations
      2. Channel-based Long Short-Term Memory Networks
      3. Feature Augmentation
      4. A CNN-LSTM Network
   4. Network Optimization
   5. Postprocessing Steps
5. EVALUATION METRICS  61
   1. Fundamental Scores and Derived Measures
   2. Evaluation Metrics
      1. NIST Actual Term Weighted Value (ATWV)
      2. Dynamic Programming Alignment (DPALIGN)
      3. Epoch-Based Sampling (EPOCH)
      4. Any-Overlap Method (OVLP)
      5. Time-Aligned Event Scoring (TAES)
   3. A Comparison of Metrics
6. EXPERIMENTAL RESULTS  76
   1. Database
   2. Performance Assessment using Objective Evaluation Metrics
   3. Receiver Operating Characteristic (ROC) Analysis
   4. Segmentation Analysis
   5. Causes of Poor Segmentation
   6. Seizure Boundary Detections Within Segment Guard Bands
   7. Analysis of the P2 Kernels Associated with the History Features
7. FUTURE PLANS  100
   1. Remaining Research Tasks
   2. Outline of Future Plans
   3. Publications
      1. Work Published or Accepted for Publication
      2. Publications Under Development

REFERENCES  107

# LIST OF TABLES

1. Normal brain rhythms  14
2. Performance of machine learning architectures according to derived metrics  74
3. TUSZ v1.2.1 statistics  76
4. A comparison of performance on the TUSZ dev set  77
5. Performance on the TUSZ blind evaluation set  78
6. Performance on the DUSZ blind evaluation set  78
7. Sensitivity of the baseline model as a function of event duration  92
8. Sensitivity of the P2 model  92

# LIST OF FIGURES

1. An industry-standard 10-20 electrode placement with a temporal central parasagittal montage  15
2. A decision tree that depicts the process used to identify an abnormal EEG  17
3. 3-4 Hz spike and slow wave complexes (absence discharges)  18
4. Secondary generalized seizure  22
5. A typical seizure event: after onset (top) and near offset (bottom)  24
6. A spectrogram of a clean seizure  25
7. qEEG analysis tools used for identifying seizures  25
8. The ictal-interictal continuum (IIC) frequency range for periodic discharges  30
9. A continuum of events related to neuronal injury  31
10. Diagram of a multilayer perceptron network  35
11. CNN network topology (Jordan, 2017) 37
12. An example of feature maps generated by CNN kernels (Lee et al., 2009) 38
13. A typical recurrent neural networ 40
14. Architecture of an LSTM cell (Hochreiter & Schmidhuber, 1997) 41
15. The proposed multiphase model 45
16. A channel-based long short-term memory network architecture 49
17. Feature augmentation using the P1 posteriors 53
18. A CNN-LSTM network architecture 54
19. The learning rate schedule 58
20. Postprocessing steps 59
21. ATWV scores this segment as 1 TP and 5 FPs.65
22. ATWV scores this segment as 0 TP and 3 FN events. 65
23. DPALIGN aligns symbol sequences based on edit distance, ignoring the actual time alignments present in the reference annotation and the system output 67
24. EPOCH scoring directly measures the similarity of the time-aligned annotations. TP, FN and FP are 5, 2 and 1 respectively. 68
25. OVLP scoring is very permissive about the degree of overlap between the reference and hypothesis. The TP score for Example 1 is 1 with no false alarms. In Example 2, the system detects 2 out of 3 seizure events, so the TP and FN scores are 2 and 1 respectively. 70
26. TAES. scoring accounts for the amount of overlap between the reference and hypothesis. TAES scores Example 1 as 0.71 TP, 0.29 FN and 0.14 FP. Example 2 is scored as 1 TP, 1 FN and 1 FP. 72
27. An example that summarizes the differences between scoring metrics 72
28. ROC curves on the TUSZ dev set 80
29. ROC curves on the TUSZ eval set 80
30. ROC curves on the DUSZ Corpus 81
31. ROC curves on the TUSZ dev set using the TAES metric 82
32. ROC curves on the TUSZ eval set using the TAES metric 82
33. ROC curves on the DUSZ eval set using the TAES metric 83
34. Baseline model segmentation for the TUSZ dev set is shown. Over-detection is shown on the left while under-detection is shown on the right. Onset boundary behavior is shown in red while offset boundary behavior is shown in green. 84
35. Baseline model segmentation for the TUSZ eval set 86
36. Baseline model segmentation for DUSZ set 87
37. P2 model segmentation for the TUSZ dev set 88
38. P2 model segmentation for the TUSZ eval set 89
39. P2 model segmentation for the DUSZ eval set 90
40. Event detections within the guard bands of the reference annotation boundaries 94
41. Layer-1 CNN kernels for recurring seizures (during the event) 96
42. Layer-1 CNN kernels for recurring seizures (between the event) 97
43. Layer-1 CNN kernels for subtle seizures 99
44. Dissertation completion plan  103

# 

**INTRODUCTION**

The electroencephalogram (EEG) is the primary tool used for the diagnosis of a variety of neural pathologies such as epilepsy. Identification of epileptic seizures from an EEG is a challenging and time-consuming process. Due to increasing awareness and a growing number of epilepsy patients worldwide, neurologists are under an immense workload in hospitals (Wyllie, 2015). Further, recent advances in EEG equipment such as the introduction of low-cost consumer grade devices (Pathirana et al., 2018), comfortable scalp electrodes that facilitate long-term data collection (Grozea et al., 2011) and continuous EEG (cEEG) monitoring (Kubota et al., 2018) have tremendously increased the amount of recorded data that needs review.

Not surprisingly, within the last decade, an increasing number of researchers have proposed solutions to expedite the diagnosis process by either automating EEG event detection or providing a means by which diagnosis can be done quickly. Tools developed based on rudimentary signal processing techniques such as the quantitative EEG (qEEG) also known as trend analysis (Persyst Development Corporation, 2020; Encevis, 2020) and machine learning-based spike detection (Persyst Development Corporation, 2020; Encevis, 2020; Brain Electrical Source Analysis, 2020) are examples of successful commercialization (Jing et al., 2018; Persyst Development Corporation, 2017). Automatic interpretation of an EEG using advanced machine learning techniques (Golmohammadi et al., 2018; Roy et al., 2019; Scheuer et al., 2020) that leverage big data resources (Obeid & Picone, 2016; Seizure Prediction Project Freiburg, 2020) offers a potential solution for increasing a neurologist’s productivity so that this abundance of data can be managed.

## Recent Advances in Automated Seizure Detection

EEG signals are captured from the simultaneous firing of the brain’s pyramidal cells which are oriented perpendicular to the surface of the skull. These signals are captured at very low voltages via EEG electrodes. These signals are contaminated with environmental noise, patient movement and other physiological and biological changes that make the EEG signal extremely noisy. Such low fidelity signals make identification of EEG events a difficult problem. In many cases, identification of these events is subjective. This subjective interpretation is not simply limited to the identification of the presence of an event, but also the determination of the duration of an event and the onset/offset times of the event.

These differences in the annotation process can be attributed to the type of training that a neurologist has received, different clinical methodologies employed by healthcare providers and a natural ambiguity in the signal that is not completely addressed by standard interpretation guidelines (American Clinical Neurophysiology Society, 2012). The diagnosis process is conservative to avoid overdosing patients or administering unnecessary medications since in many cases these medications can cause more harm than good to patients (Benbadis, 2010). Interrater agreement (IRA) among experts for identifying key events such as spike and wave events or periodic discharges is very low for critically ill ICU patients (Halford et al., 2015; Shah et al., 2020). Similarly, IRA for the seizure events is low as well (Haider et al., 2016; Halford et al., 2015; Swisher et al., 2015).

With the recent explosive growth in the field of machine learning, and related fields such as data science, there has been renewed interest in applying machine learning techniques based on big data to the automated seizure detection problem. Yuan et al. (2018) proposed a channel-aware attention network to weight the EEG channels based on their contribution to the seizure events. The fused attention networks implemented using local (channel) and global information performed at an accuracy of with an F1-score of . The evaluation metrics used in this study are a single unit of merit (e.g., accuracy) instead of specifying sensitivity and specificity as two separate scores. It is difficult to estimate error rates such as false positive rates and miss rates of the seizure class from these measures. In fact, this study was performed on the Children’s Hospital of Boston, Massachusetts Institute of Technology (CHB-MIT) database (Shoeb, 1981) which contains only subjects and includes an evaluation subset that consists of only patients and frames of data.

Hosseini et al. (2016) performed real-time seizure detection with the aid of a high-performance computing cluster by separating the real-time seizure detection problem into three stages: (1) feature extraction, (2) independent component analysis (ICA), and (3) Support Vector Machine (SVM) ensemble learning. In the first stage, the model extracts spectral and statistical features such as signal/frequency energies, entropy, skewness, etc. These features are transformed via an infinite-ICA algorithm (Knowles & Ghahramani, 2007) and applied as inputs into multiple SVM classifiers. A final decision is made by taking a majority vote among all the SVM classifiers. This model yielded sensitivity with specificity on the CHB-MIT database. The false positive (FP) rate for these ensemble models was , which is quite high given that clinically acceptable performance requires an FP rate less than (Golmohammadi et al., 2018).

Vidyaratne & Iftekharuddin (2017) used a harmonic wave packet transform (HWPT) and fractal dimension estimates of the EEG signal as features to find repetitive morphologies of a seizure event. A Relevance Vector Machine (RVM) was trained using these feature vectors which yielded a sensitivity with 0.243 false positives per hour. Golmohammadi et al. (2018) used a hybrid neural network model which combines convolutional neural networks (CNNs) and long-short term memory (LSTM) networks to learn spatio-temporal features of the seizures. This state-of-the-art model yielded a sensitivity of with a false positive rate of false positives per hours on the development set of the Temple University Hospital Seizure Detection Corpus (TUSZ) (Shah et al., 2018).

Scheuer et al. (2020) performed an interrater agreement study between three expert neurologists and their state-of-the-art seizure detection model to compare the performance of a popular commercial software package, P14 Persyst – The Seizure Detector (Persyst Development Corporation, 2020), to human experts. The authors address the real-time seizure detection problem using multiple approaches including: (1) artifact reduction, (2) per-channel processing to identify focal seizures, (3) empirical null technology to better capture the non-normal statistics of the EEGs and (4) spike detection. A total of different features were selected and integrated using convolutional neural networks (CNNs). The algorithm was able to detect seizures at a sensitivity of with false positive per day. The median latency for their software was seconds. Latency is a very important figure of merit for this type of technology because it impacts how the technology can be applied in clinical settings such as critical care.

## Recent Advances in Automatic Segmentation Algorithms

A major challenge in the seizure detection problem is the problem of accurately determining the onset and offset of a seizure. We refer to this as the segmentation problem. Limited research has been conducted on segmentation and it is often limited to databases which do not represent real-world scenarios or contain many of the challenging artifacts observed in clinical data. Shoeb et al. (2011) focused on seizure offset detection, which is referred to as post-ictal detection. In this study, the authors created different frequency bands using all available channels of the EEG signal and created a feature vector from -second long segments of the signal, which we refer to as windows. An offset point was estimated using these feature vectors and used as input to an SVM algorithm. An accuracy of with a mean absolute error of  secs for the offset detection was measured using the CHB-MIT database.

Klatchko et al. (1998) focused on only the onset of the seizure event where they used directed graphs which use conjunctive subsets of EEG wave shapes. The wave shape’s contribution was calculated across multiple channels via affine transforms. The resulting feature vectors are calculated on a window size of  seconds. The overlap between the waveform mappings and predefined seizure clusters are used to provide a probability of a seizure: a higher overlap means higher probability. These probabilities are further fused with an EEG event classifier such as a spike detector. The fusion process reduced the false positive rate by and the false negative rate by . Similar to the previous study by Shoeb et al., this study was conducted on the CHB-MIT database and only examined seizures with a duration between six and seconds.

Orosco et al. (2016) proposed a seizure detection model using linear discriminant analysis (LDA) and neural network (NN) classifiers. The discrete wavelet transform and power spectral density were used to generate features of an EEG signal below  Hz in frequency. The authors reported a mean deviation of  secs from onset boundaries and  secs from the offset boundaries. The maximum deviation ranged from to  secs for detected seizures. This study was also performed on the CHB-MIT database.

Correa et al. (2015) also performed real-time seizure detection via setting threshold for a wavelet decomposition model. The input to the model were features collected from the power spectrum of various EEG rhythms (e.g., alpha, beta, theta) and their temporal derivatives. The detected events were far by to  secs from the reference onsets and  to  secs from the reference offsets.

Common flaws among many of these studies include databases that contain a limited number of subjects, a lack of evaluation data that is a good representation of the types of common artifacts seen in clinical data, and a lack of a standard scoring methodology. For example, studies performed by Shoeb et al. (2011) and Yuan et al. (2018) use subjects ranging in age from  to , which is very limited compared to the age range seen in a typical urban public hospital. Further, in the study from Klatchko et al. (1998), an open-set testing methodology was not observed. Patients are included in both the training and evaluation sets. Shoeb et al. (2011) only evaluates the performance of their model in terms of offset where as Klatchko et al. (1998) evaluates in terms of onset. Focusing on only one of these measures only partially reflects the quality of the segmentation. Performance was evaluated based on either smaller fixed-sized intervals or the amount of overlap between the reference and detected events.

There is a lack of FDA approved automated software in the commercial market which can detect seizures within a range of performance that is clinically acceptable (Haider et al, 2016). This is due, in part, to the difficulties related to the interpretation process, inconsistencies in annotation standards and low IRA among experts. As a consequence, many publications use lenient scoring metrics such as the Any-Overlap (OVLP) method (Wilson et al., 2003) for evaluating performance of their systems. This method, as the name suggests, incorporates any amount of overlap between a detected event and ground truth to call the hypothesis event a true positive (TP). Recognition systems which use such a lenient metrics tend to yield very good performances since the metric pays no attention to the accuracy of the segment onset and offset times. The system can hypothesize a target event around a true event and, regardless of detected event’s duration, get a positive score. Models optimized using such metrics tend to start detecting seizure events at the middle or the end of the event when they have already evolved substantially. This type of scoring does not accurately reflect the needs of clinicians. Systems evaluated using such generous metrics do not accurately assess the quality of the segmentations of the detected events. The segmentation problem has not been adequately addressed in the literature.

On the other hand, there has been increasing interest in the development of computerized seizure detection algorithms in the machine learning community. A variety of supervised and unsupervised machine learning algorithms such as K-nearest neighbors (KNN), SVMs and NNs have been proposed to identify the seizure events (Roy et al., 2019; Golmohammadi et al., 2018; Shoeb et al., 2011). Since EEGs contain information in temporal and spatial domains, sequential algorithms such as recurrent neural networks (RNN) or hidden Markov models (HMMs) and structural algorithms such as convolutional neural networks (CNNs) are perfect candidates for seizure/no-seizure classification. Seizure events can last from 3 secs (i.e. absence discharges) to days (i.e. refractory status epilepticus). Although these algorithms are able to learn short term dependencies of an EEG signal, capturing long term information becomes expensive. Systems often integrate features at multiple time scales or use networks with extended memory to model such behavior. However, the impact of these approaches on the segmentation problem has not been adequately studied, and hence, is the focus of this dissertation.

## Research Plans and Contributions

In this dissertation, we mimic the human interpretation process and address the problem of seizure segmentation by dividing the problem into a two-phase problem – epileptiform activity detection followed by seizure detection. We use linear frequency cepstral coefficients (LFCC) and their derived features, such as temporal derivatives, in both phases of the model. In the first phase, we use a sequential neural network algorithm known as a long short-term memory (LSTM) network to identify channel-specific epileptiform discharges associated with seizures. In the second phase, the LFCC-based feature vectors are augmented with the first phase’s posteriors which represent the onset and offset of ictal activities. We use a multichannel convolutional neural network (CNN) followed by an LSTM network in the second phase of the model.

This second phase network is designed such that the number of input channels of the CNN is equal to the dimensionality of a feature vector. During the first layer, the CNN kernels of each channel are optimized separately which means the network optimization is done on each dimension of the feature vector independently. We observe that the kernel outputs, known as feature maps, show boundaries clustered together for easily detected seizures and dispersed for the subtle seizures. The feature maps created in the first layer are correlated, which forces the remaining layers of the model to compete for localization of a seizure boundary. We aggregate the output of both phases of the model and perform additional postprocessing to achieve high performance for seizure detection.

The remaining part of the dissertation will be focused on understanding the benefits of the proposed model for the segmentation process and the impact this model has on seizure detection performance. There are four major tasks to be completed:

**Task 1:** Analysis of the convolutional neural network kernels assigned for the augmented feature vectors.

**Task 2:** Fine tune the segmentation boundaries using traditional energy-based detectors and novel neural network algorithms. Compare and contrast algorithms and their pros and cons.

**Task 3:** Implement an offline seizure detection algorithm to perform classification on detected seizure segments to reduce the false alarm rate.

**Task 4:** Investigate the underlying reasons about why the algorithm works well on specific seizures and fails on others. This will include correlating performance with the duration of the event and the type of seizure. Error analysis on specific sets of patients should reveal the underlying pathologies that result in degraded performance.

## Proposal Outline

The remainder of this proposal is organized into seven chapters. In Chapter 2, we discuss the basics of the EEG signal and introduce key concepts such as a montage. We review relevant standard terminology used by neurologists to analyze an EEG signal. This chapter will give us a brief understanding of the difference between a normal and abnormal EEG signal. We will apply this knowledge to characterize the segmentation problem discussed in the following chapters.

In , we will discuss a few standard EEG interpretation guidelines, our annotation process and discuss what makes the EEG interpretation process difficult. We will discuss the primary epileptiform features that are observed in epileptic patients. We review the types of artifacts that make annotation of a seizure event challenging. We discuss the gray zone areas during ictal-interictal continuum (IIC) where the identification of segment boundaries of the epileptiform and seizure events become very difficult. This chapter provides a detailed understanding of the challenges in identifying seizure events and explains why IRA among experts on such epileptiform events is typically low.

In Chapter 4, we provide a mathematical background of the algorithms used for seizure detection. We discuss the motivations behind the proposed multiphase system and the feature augmentation process. In the first phase of the model, we detect the epileptiform activities associated with the seizures which is followed by a seizure detection model in the second phase. The second phase of the model incorporates first phase’s detections as additional feature vectors. The first phase model’s hypotheses and LFCC feature vectors used in the multiphase model fine tunes the seizure segment boundaries. The motivation behind the design of first and second phase models are discussed in detail.

In , we will discuss the deficiencies of traditional scoring metrics. The rationale behind our preferred choices for metrics and the drawbacks of these metrics are discussed. We demonstrate the analysis and insight that can be obtained when the proper metric is chosen. We introduce a new metric, Time-aligned Event Scoring (TAES), which generates fractional scores that are proportional to the amount of overlap between the reference and hypothesis events. Such a metric is more appropriate for evaluating the segmentation performance of a system since it accurately measures partially detected events. We provide examples for each metric and end this chapter with an example showing how combining multiple scoring metrics gives insight into the performance of a system. Using these metrics, the behavior of a sequential pattern recognition system can be understood without the need for a manual error analysis step. This allows the system to be automatically tuned to optimize this objective metric.

In , we compare the results of the proposed approach with a previously developed state-of-the-art seizure detection system on three separate databases. We compare the segmentation performance of the two best performing systems and show that our proposed approach gives promising results in terms of accurately detecting segmentation boundaries.

Finally, in , we present our conclusions and describe the remaining research to be completed in this dissertation. A timeline and publication plan are presented.

The impact of this dissertation will be a better understanding of the importance of incorporating EEG history features in an augmented feature vector. We will analyze the importance of processing each type of feature independently in the earlier stages of the neural networks. We will visualize the kernels of the networks to observe what is being learned by the model on per-channel basis.

Improved seizure detection performance and accurate segmentation will allow clinicians to detect seizures in real time. This will allow neurologists to take action during the early stages of the event. This is essential to avoid permanent damage to the brain. In a similar way, accurate offset detection will help with the delivery of therapies designed to mitigate postictal (after seizure) period symptoms. This will also help reveal the severity of a seizure and as a consequence, provide guidance for medicating a patient.

# 

**THE ELECTROENCEPHALOGRAM**

This chapter provides an introduction to the science of electroencephalograms (EEGs). We briefly discuss scalp EEG electrode placement methods and the type of signal preprocessing that is commonly performed. We discuss ways to better visualize and interpret EEG signals via differencing these electrode signals. This is followed by a discussion on how identification and annotation of specific EEG events is performed.

## The Electroencephalogram’s Role in Neurology

The EEG is a primary tool for capturing brain’s electrical activity. Over the years, new technologies such as fMRI (Szaflarski et al., 2017) and MEG (Liu et al., 2000) have emerged for capturing neural functioning but the EEG still remains the primary technique for diagnosis due to its cost-effectiveness and convenience. Non-invasive scalp EEG recording methods are the preferred way for collecting data in clinical settings. The most common use of an EEG is in the diagnosis of epilepsy. This disease affects approximately 40 million individuals worldwide (Wyllie, 2015). Epilepsy patients are under the constant threat of experiencing an epileptic seizure. This is a life-changing diagnosis because it affects the quality of life of a patient (e.g., loss of driving privileges).

In recent years, long-term and continuous EEG (cEEG) monitoring are increasingly being used in hospitals (Kubota et al., 2018). Due to advancements in the EEG technology, these monitoring technologies allow the EEG signal to be recorded for durations ranging from several hours to days. Increasing the number of EEG records that require manual review has significantly increased the workload for the neurophysiologists. The delay between an EEG test and interpretation by a trained healthcare provider has further impacted the ability to provide high quality patient care. The need to reduce the amount of data that needs manual review and the lag between testing and interpretation has created a major market opportunity for automatic seizure detection software in clinical settings (Golmohammadi et al., 2018; Persyst Development Corporation, 2020).

**Table 1**. Normal brain rhythms

|  |  |  |  |
| --- | --- | --- | --- |
| **Rhythm** | **Frequency (Hz)** | **Amp. Range (μv)** | **Activity** |
| Delta | 0-4 | 20-200 | Deep sleep |
| Theta | 4-7 | 20-100 | Creativity, intuition |
| Alpha | 8-13 | 20-60 | Relaxation |
| Beta | 13-30 | 2-20 | Memory |
| Gamma | 30-100 | 20-70 | Cognition, learning |

Scalp EEG signals contain information related to cerebral activity within the frequency range of to Hz frequency where majority of the information resides below Hz in the spectrum (Shoeb et al, 2011). The frequency content of normal brain rhythms is summarized in . Electrical signals are measured from an array of electrodes placed around the scalp, as shown in . These signals are typically sampled at  Hz, a common sample frequency used in clinical settings because it captures all of the significant information about brain activity but is low enough to minimize file storage requirements. These signals have amplitudes in the microvolt range making them extremely susceptible to noise. As a result, these signals are filtered and amplified prior to digitization.

The information captured within Hz is further divided into frequency bands called “EEG rhythms.” EEG rhythms are frequency bands which are separated based on the cognitive and behavioral activities observed from a human brain. To identify specific types of activities, EEG signals are divided into different frequency bands (Tatum et al., 2014. The frequency ranges of these rhythms with their corresponding nomenclature are: Delta:  Hz, Theta:  Hz, Alpha:  Hz, Beta:  Hz and Gamma:  Hz. These rhythms play an important role in the diagnostic process because each rhythm conveys specific information related to a subject’s mental status and the environmental conditions around it. One of the first steps during the diagnosis process is to identify if an EEG signal depicts a normally functioning brain.

A screenshot of a cell phone

Description automatically generated

**Figure** **1.** An industry-standard 10-20 electrode placement with a temporal central parasagittal montage (TCP) overlaid

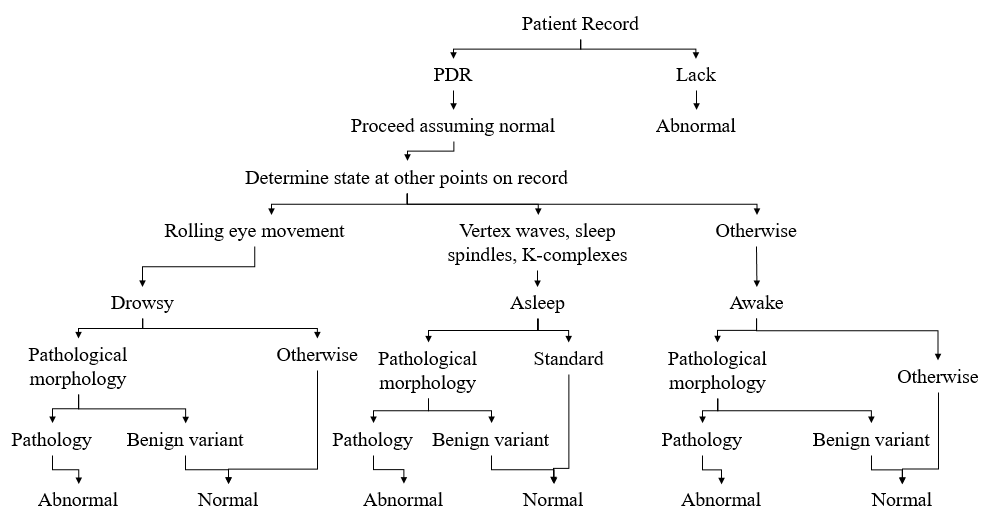
## Normal and Abnormal EEGs

For normal EEGs, each of these rhythms are expected to have specific amplitude ranges and focal regions. For example, an alpha rhythm can range from microvolts and can be observed mainly in the occipital region of the brain (the back portion of the brain) or any other region other than the front. Activities within this rhythm are normally observed during eye blinks, when eyes are closed (also known as a posterior dominant rhythm) or during resting conditions (e.g., sleep spindles). A beta rhythm can have an amplitude around  microvolts and can be mainly observed in the motor cortex region of the brain during mental or physical activities. Delta and theta rhythms exhibit amplitudes greater than  microvolts and can be observed during deep sleep and drowsiness, respectively. The normal amplitude range for these rhythms is within  microvolts (Krauss & Fisher, 2011). Activities observed below this range can be attributed to various factors such as medication effects or poor data acquisition system. Activity observed above this range can be identified as normal (e.g., drowsiness), abnormal (e.g., rhythmic delta activity) or artifactual (e.g., muscle movements).

Of course, these amplitude and frequency ranges are merely examples of activities associated with the rhythms observed in normal patients. Since scalp EEG electrodes capture multiple frequencies originating from multiple sources, a mix of all these rhythms can be observed everywhere on the scalp, especially if the patient possesses abnormal brain behavior. These activities can be separated into mild abnormalities (e.g., EEG asymmetry between hemispheres, a lack of reactivity) and severe abnormalities (e.g., signals exhibiting epileptiform/ictal features, burst suppression).

There are also exceptions to the definition of an abnormal event. For example, an alpha rhythm can have as much as a asymmetry between hemispheres. Even the complete absence of an alpha rhythm is considered normal behavior among adults (Krauss & Fisher, 2011). A basic decision tree summarizing the diagnostic process used to identify an abnormal EEG is shown in . The normal/abnormal EEG event classification problem is discussed in more detail in Lopez de Diego (2017).

It is important to note that the interpretation of EEG signals based on their frequency ranges (rhythms) is done quite differently than other disciplines in signal processing. The focus for EEG interpretation is on the dominant frequencies in the lower part of the spectrum. The frequency content in an EEG rhythm is measured based on the number of cycles per second observed in the waveform ­– a process often called peak-picking in the signal processing literature. For example, morphologies such as a  Hz spike and wave (absence) complex usually shows significant energy in its high frequency components.



**Figure 2.** A decision tree that depicts the process used to identify an abnormal EEG

The most prominent frequencies in this morphology are  Hz slow wave delta rhythms which complete three cycles per second. Hence, the name “ Hz” as a prefix regardless of the abundance of much higher frequency components. In , examples of Hz spike and wave complexes are shown. This can be considered a type of seizure referred to as an “absence seizure” (petit-mal) (Albuja, 2020). From the plot, it can be seen that there are many other sharper activities evolving from the first channel (FP1-F3) and fifth channel (FP2-F4). But each of these small events repeat at a rate of approximately  Hz in frequency and hence the term “ Hz spike and slow wave complexes/discharges” is used to describe this type of waveform.

A close up of text on a white background

Description automatically generated

**Figure 3.** 3-4 Hz spike and slow wave complexes (absence discharges) (Albuja, 2020)

## The TCP Montage and 10/20 Electrode Placement

A scalp EEG recording is most frequently collected using a electrode placement in clinical settings (Tatum et al., 2014), as shown in . A placement is also popular (American Clinical Neurophysiology Society, 2012). Most expert clinicians are trained to interpret signals collected using these two standards. Interpretation is usually preferred to be done on a noise-reduced version of the signal known as a montage (Tatum et al., 2014). In a typical montage, signals are differenced to reduce noise and localize the EEG activities. This aids interpretation process by removing common noise from adjacent or reference electrodes.

There are two types of montages commonly used: unipolar and bipolar. Unipolar montages are created by simply subtracting signals collected from each electrode with only one reference electrode. Bipolar montages, on the other hand, subtracts signals from adjacent electrodes. This aids in localization of an EEG events – determining the location in the brain where the event originated. Identification of such focal regions is done by checking phase reversals between EEG channels. The montages are created in a way that signal subtraction is done in two directions: (1) a longitudinal direction (nasion to inion) and (2) a transverse direction (left to right ear). Both these directions help in identifying specific EEG events. For example, a triphasic wave can be observed from the longitudinal direction since it tends to propagate from the frontal lobe to occipital lobe of the brain. On the other hand, a vertex wave is best seen from the transverse direction because it appears at the midline of the skull from where the transverse montage is created.

Popular montages are discussed in great detail in Ferrell et al. (2020), Lopez et al. (2016), and Gross et al. (2016). For our research, we have always preferred the popular temporal central parasagittal (TCP) bipolar montage. The  electrode placement system with a TCP montage is shown in . There are four segments focused in the longitudinal direction (front to back / vertical line) and one in the transverse direction (red horizontal line).

# 

**THE CHALLENGES OF EEG INTERPRETATION**

With a growing interest in machine learning approaches to EEG interpretation, the literature has evolved significantly in recent years to the point where standardized literature on the interpretation process and diagnosis is openly available (American Clinical Neurophysiology Society, 2012; American Academy of Neurology, 2020). One of the most difficult aspects of building a machine learning system is understanding the process experts use to interpret the data. With recent interest in computer automated interpretation, the literature describing the manual interpretation process has become much richer.

There are multiple factors that play an important role in making a decision about a diagnosis. In addition to evidence in the EEG signal, clinicians incorporate information such as a patient’s medication history, physiological condition and clinical correlates during a recording session (e.g., a video recording, an EMG signal) (Krauss & Fisher, 2011). See Ferrell et al. (2020) for a more complete discussion of the types of physiological signals collected during a clinical EEG. More recently, fMRI technology has played an important role in pinpointing asymmetry and abnormalities in the brain (Szaflarski et al., 2017; Kesavadas & Thomas, 2008).

In this chapter, we will examine the EEG interpretation process and demonstrate that the interpretation process is quite subjective. EEG interpretation requires extensive clinical training. Board certified neurologists complete a one-year internship and three years of clinical training. During this time, they are required to experience both inpatient (e.g., a hospital) and outpatient (e.g., a clinic) care. These candidates are exposed to patients in intensive care units (ICU), emergency rooms (ER), neuroimaging units (e.g., a department of neurology), child neurology units (e.g., hospitals specializing in pediatric care) and psychiatry (Benbadis, 2014).

Training includes studying EEG signal morphologies observed during subclinical seizures and learning how to integrate clinical correlates and medical histories. Subclinical seizures are observed from the EEG signal, typically visualized using a tool that displays the multichannel signal in -sec segments (Capp et al., 2017). Clinical seizures often require observing a patient’s mental state (e.g., consciousness). Typical examples of subclinical seizures are simple partial seizures (conscious state) or a sedated ICU patient showing no outward physical signs of a seizure. On the other hand, patients with generalized tonic-clonic seizures show extreme physical activity. Patients with complex partial seizures show a lack of consciousness. In this study, we only focus on seizures which can be observed from electrographic data (e.g., subclinical seizures).

The two major types of epilepsy that influence the type of seizure are generalized epilepsy and partial (or focal) epilepsy. The localization of an event often determines the difference between these two. Generalized epilepsy can cause tonic-clonic, absence, myoclonic, clonic, tonic, and atonic seizures. Partial epilepsy, on the other hand, can cause simple partial and complex partial seizures. Secondary generalized epilepsy is the combination of these two types where a focal seizure evolves into a generalized seizure or spreads across the lobes. An example of a generalized epileptic seizure can be seen in . An example of a partial seizure evolving into a generalized seizure can be seen in Figure 4.

The first question that arises is why is the seizure interpretation task difficult? The following reasons are just a few of the many reasons this is a challenging task:

1. In a clinical setting, identification of a seizure event is based on the information collected from the encephalogram as well as clinical sense.
2. Seizure morphologies can vary drastically and, in many cases, are specific to the patients.
3. The EEG event classification problem is similar to an array processing or video processing problem in which temporal and spatial context plays an important role.
4. The signals are extremely noisy due to the low voltage levels of the electrical signals and the presence of artifacts introduced by the ambient recording environment.
5. Poor interrater agreement among the experts make it harder to set hard rules for establishing standards.

A picture containing screenshot

Description automatically generated

**Figure 4.** Secondary generalized seizure

Points 1 and 5 are a challenge to deal with when developing computer-based algorithms. In order to incorporate information collected from clinical reports, video recordings and other environmental settings, sophisticated information processing technologies are required. For example, clinical information about a patient’s medications and medical history is often represented by unstructured text and requires advanced natural language processing (NLP) algorithms to extract (Picone et al., 2018; Chapman et al., 2001). Development of such algorithms requires an abundance of annotated deidentified clinical reports.

On the other hand, there are specific traits which confirm the existence of a seizure or at least ictal discharges. Some of the primary features include:

1. Existence of epileptiform discharges such as a train of spike and wave discharges/complexes;
2. Periodic or rhythmic discharges from specific frequency ranges ( Hz to  Hz) and durations  sec);
3. Evolution in amplitude and frequency;
4. The spatial context based on the EEG event’s locality (such as the frontal or parietal lobes).

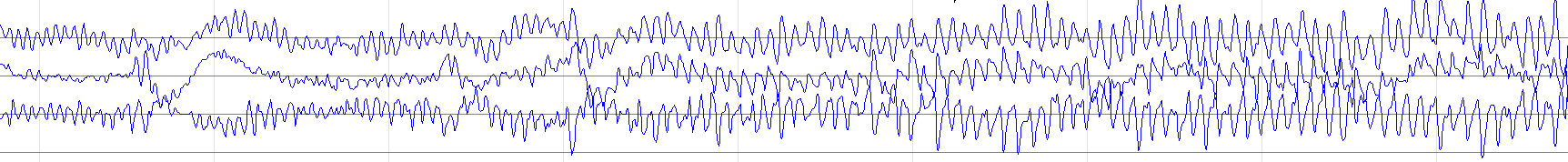
The latter point plays an important role in identification or rejection of cerebral or artifactual activities.

Evolution in frequency is defined as at least consecutive changes in the same direction by at least s (American Clinical Neurophysiology Society, 2012). Development of any EEG event detection algorithm should incorporate such knowledge. In the following sections, we briefly explore the EEG event interpretation process and discuss important edge conditions that make this a challenge.

## Contemporary Visualization Tools

EEG event classification requires characterization of a variety of events including cerebral activities emanating from the surface of the scalp, biological or mechanical/external artifacts and ambient or electrical noise. For our research, we will explicitly discuss epileptiform ictal discharges and artifacts. The most common events in this category include spike and wave discharges, periodic discharges, biological artifacts and mechanical artifacts (Krauss & Fisher, 2011). The evolution of such events with respect to time can give us an indication of a seizure event.

Typical seizure events are comprised of ictal epileptiform discharges such as spike and wave discharge/complexes, sharp and wave discharges/complexes, generalized polyspikes, hypsarrhythmia, and lateralized/generalized periodic discharges. A seizure is indicated based on such discharges evolving with an increasing rate in amplitude and a decreasing rate in frequency.

******

A close up of a womans face

Description automatically generated

**Figure 5.** A typical seizure event: after onset (top) and near offset (bottom)

Two typical seizure waveforms are shown in . The signals are captured at the onset of a seizure (top) and at the offset of a seizure (bottom). shows the spectral properties of a seizure event captured on a subset of channels. The time axis (horizontal) for the spectrogram covers  seconds while the frequency scale (vertical) is truncated to  Hz so that the most meaningful aspects of the seizure event can be observed.

In practice, this information is typically visualized using an approach known as a quantitative EEG (qEEG) (Nuwer, 1997). The qEEG is also known as a trend analysis technique that helps experts rapidly analyze EEG records and localize specific events. Recently there have been multiple studies validating the reliability of qEEG trend analysis methods (Haider et al., 2016; Swisher et al., 2015), and hence, this tool has been rapidly growing in popularity. Tools related to the qEEG include an amplitude integrated EEG (aEEG) (Scheuer et al., 2004), asymmetry index, spectrograms, rhythmicity spectrograms and envelope trends (Hirsch et al., 2010). These tools essentially aggregate information collected from the left and right hemispheres of the brains. Examples of seizures identified using these tools are shown in . These examples came from a study conducted by Haider et al. (2016).

***A screenshot of a cell phone

Description automatically generated***

**Figure 6.** A spectrogram of a clean seizure

Although such tools save time, they are not able to detect subtle seizure events which are extremely focal, small in amplitude or very brief in duration (Haider et al., 2016). Human intervention is often required (Swisher et al., 2015). Consequently, qEEG analysis is considered an auxiliary tool to be used alongside the analysis of the raw EEG signal.

A screenshot of a cell phone

Description automatically generated

**Figure 7.** qEEG analysis tools used for identifying seizures (Haider et al, 2016)

Commercially available seizure detection software, such as an industry-leading product offered by Persyst Development Corporation (2016), perform poorly in clinical settings. Common complaints from users include low sensitivity and large latency (Haider et al., 2016). To understand how to design an efficient automated seizure detection algorithm, we first need to understand what makes the seizure identification problem difficult. We will specifically examine the morphologies that comprise a seizure. Specifically, we will discuss the two most prominent ictal morphologies: spike and wave discharges and lateralized/generalized periodic discharges: We will also discuss two morphologies that contribute most prominently to a high false alarm rate: slowing and artifacts.

### Ictal Epileptiform Features

Epileptic seizures consist of epileptiform events such as spikes, sharps, spike/sharp and wave discharges/complexes, spike and slow wave discharges/complexes, polyspikes and hypsarrhythmia. In our research, we decided to consider all these events variants of spike and wave discharges with the exception of hypsarrhythmia. We exclude hypsarrhythmia because the event morphologies observed during hypsarrhythmia lack any periodic or rhythmic structure. Rather, hypsarrhythmia event morphologies are a chaotic mixture of high amplitude slow waves, multifocal spikes and asynchrony (Scher, 2017). The identification of a seizure event is confirmed based on the evolution of frequency, amplitude, affected (focal or hemispheric) regions and their spread (Stefan & Theodore, 2012). Trains of small duration ictal discharges (less than  msec for spikes and  msec for sharps) are observed during the most seizure events. These discharges with frequencies greater than  Hz and a duration of at least  seconds confirm the existence of a seizure event (American Clinical Neurophysiology Society, 2012). Potential seizure events with parameters outside these ranges require additional information such as a determination of the epilepsy type (i.e. absence seizures). On the other hand, epileptiform discharges that satisfy these criteria are automatically confirmed as seizure events. Other epileptiform events which are not covered by this definition require further evidence (e.g., absence seizures lasting for  seconds or periodic discharges occurring at a frequency of  Hz). Since identification of ictal features is the first step towards confirming a seizure event, in our automated system we add an additional step for detecting epileptiform features on a per-channel basis.

### Periodic Discharges

Periodic discharges such as generalized periodic discharges (GPD) or lateralized periodic discharges (LPD) are abnormalities which occur at a constant interval. If these discharges are confirmed as epileptiform, high frequency trains or evolution of these events can be indicative of a seizure event. However, the periodicity at which the GPDs are considered as epileptic seizures is not clear and highly disputed in the literature (Kubota et al., 2018). This makes identification of these events challenging.

### Slowing

Slowing includes a broad set EEG events which can indicate normal, mildly abnormal and abnormal EEGs depending on a number of factors. We specifically focus on post-ictal slowing events which occur at the offset of a seizure. Annotation of post-ictal slowing is not easy since it can last from a couple seconds to a week (von Weltin et al., 2018). This adds additional difficulty in marking the offset of a seizure event. We have applied a stringent set of rules for determining slowing in the TUSZ Corpus (Ochal et al., 2020; Shah et al., 2018): an offset is annotated when consistent evolution is phased out and small attenuated gaps between the slow waves are found.

### Artifacts

Since scalp EEG recordings are non-invasive, extraneous activities from surface and environmental disturbances can be observed throughout the record. These events are referred to as artifacts. Artifacts are events observed in EEG signals which are not related to actual brain activity. EEG artifacts may resemble activities similar to cerebral abnormal or normal morphologies (e.g., epileptiform discharges and intermittent rhythmic delta activity). While some artifacts resemble actual brain activity, others introduce severe noise into a signal, making interpretation of the record difficult if not impossible.

Clinical settings such as a neurological intensive care unit (NICU) or a cardiac intensive care unit (CICU) contain a great deal of ambient electronic noise. In such environments, the introduction of noise due to movement of personnel and an agitated patient are common (Krauss & Fisher, 2011). Such settings produce intermittent or continuous electromyogram (EMG) high-frequency artifacts which can completely obscure underlying brain activity. These high frequency/amplitude irregular activities can appear in a frequency band ranging from  to  Hz. Other artifacts may include chewing, electrode pop/shake, shivering and eye movements.

## ACNS Standards for Seizure Identification

The American Clinical Neurophysiology Society (ACNS) (2012) has produced a standard set of guidelines for identifying seizure events for intracranial and scalp EEG recordings. Clinicians are encouraged to research, interpret and diagnose patients using this nomenclature and these guidelines.

Spike and wave discharges at  Hz or faster and clearly evolving discharges of any type that reach a frequency of  Hz can be considered as seizures (Hirsch et al., 2013). There is also the  Hz- second rule previously discussed. Although these definitions are straightforward, other morphologies which do not meet these criteria introduce ambiguity. The ACNS guidelines state: “Generalized spike and wave patterns slower than s; and evolving discharges that remain slower than or equal to s does not imply that these patterns are not ictal, but simply that they may or may not be.” Additional information is necessary to make a final decision about a seizure event. Such decisions are usually dependent on a clinician’s experience, institutional conventions and a patient’s condition. IRA among experts on such seizure events is very low (Haider et al., 2016; Swisher et al., 2015).

## Ictal Discharges and EEG Gray Zone Events

Based on the evolution of the epileptiform events, seizure events can be divided into two classes: isomorphic and metamorphic. Isomorphic events show little to no change throughout the phase (i.e. absence seizure) of the event whereas metamorphic events show multiple phases of an event evolving through time (i.e. tonic-clonic seizure). and show easily identifiable examples of isomorphic and metamorphic seizure events respectively. Although the example shown in shows a clear distinction between its onset and offset marks, in many cases, for patients with other epilepsy syndromes such as Lennox-Gastaut syndrome (LGS) or patients showing hypsarrhythmia, there is little to no changes in their electrographic signals. Patients with LGS can show multiple types of epilepsies in different focal regions. As a result, multiple types of seizure events can be observed throughout the record with subtle transitions in the signal morphologies. This makes the onset or offset detection of a seizure event hard. The onset and offset marking for the isomorphic seizures is a big challenge which requires additional insight about the clinical correlates and patient history.

Periodic discharges such as LPDs and GPDs are heterogeneous and multifactorial (attributed variety of causes such as metabolic, infectious or epileptic disorders) which makes some sort of generalized rule for seizure detection difficult (Chong & Hirsch, 2005). These periodic discharges are often observed in an ictal-interictal continuum (IIC) where seizure identification becomes subjective (Sivaraju et al., 2016; Persyst Development Corporation, 2020). In , we can see the ambiguous “gray zone” region associated with the periodic discharges with periodicity less than 3/cycles. Similarly, shows a clear depiction of the uncertainty involved in determining whether periodic discharges due to a cortical injury are determined to be a seizure event. Both these examples often require additional information about the patient and remain a very subjective decision.

***A screenshot of a cell phone

Description automatically generated***

**Figure 8.** The ictal-interictal continuum (IIC) frequency range for periodic discharges (Sivaraju & Gilmore, 2016)

These are a few examples of why the identification of a seizure event from an EEG signal is a challenging problem. There are additional challenges in interpreting scalp EEG recordings in clinical settings that include a lack of a medical history (including medications), a lack of video recordings, mechanical artifacts and other pathological complications such as triphasic waves observed in encephalopathic patients. Also, sedative medications reduce the energy of the cerebral signals and hence the quality of the recordings.

***A close up of text on a black background

Description automatically generated***

**Figure 9.** A continuum of events related to neuronal injury (Chong & Hirsch, 2005)

## Interrater Agreement Among Experts

For many of the reasons previously discussed, and to avoid over-medicating patients (which can cause long-term damage), neurologists tend to annotate EEGs conservatively (Benbadis, 2010). Not surprisingly, there are often disagreements among the experts. Consequently, interrater agreement (IRA) among experts remains quite low (a typical kappa statistic is approximately ) (Haider et al., 2016; Swisher et al., 2015; Ronner et al., 2009). Poor IRA makes annotation of large corpora a challenge because the annotations will not be consistent when multiple annotators are involved. Models developed and evaluated on such databases tend to yield lower performance, especially with harsher scoring metrics.

We typically employ undergraduates to annotate data. These student experts were responsible for creating the TUSZ database (Shah et al., 2018). These students were trained to annotate data based on the principles previously discussed. We conducted an IRA study between the student annotators and expert clinicians to evaluate their consistency for annotating seizures. We established some specific rules for these annotations since the ACNS guidelines are a bit vague in places. Examples of rules we needed to refine address (1) frequency evolution of periodic lateralized discharges (PLDs) and generalized periodic discharges (GPDs) appearing in long bursts, (2) post-status epilepticus stages and “gray zone” Ictal Interictal Continuum (IIC) phases, and (3) low frequency (1-3 Hz) spike and wave discharges lasting for more than 10 seconds.

To solve (1), annotators were required to annotate periodic discharges greater than  Hz in frequency and lasting more than  seconds regardless of their waxing (bursty) and waning (slow) patterns. If the frequency is lower than  Hz, evolution of the discharges was taken into account to confirm a seizure. For (2), we mark the offset of an event when the first suppressed or attenuated background is observed during the post-ictal slow waves. Similar to (1), we incorporate evolution of the discharges prior to making a decision about (3). Despite these very specific guidelines, there are still many cases for which determination of a seizure is ambiguous. To make sure the final annotations are consistent, we submit these cases to group review and achieve a consensus.

Although student annotators rely heavily on the reports for the annotation process, for this IRA study they were not permitted to use the reports. Their performance was compared to “gold-standard” annotations derived from annotations provided by multiple expert neurologists. Data was drawn from three databases collected at three different hospitals: TUSZ (Shah et al, 2018), the Duke University Seizure Corpus (DUSZ) (Swisher et al, 2015), and the Emory University Seizure Corpus (EUSZ) (Haider et al., 2016). We evaluated the performance of our annotators using Cohen’s kappa statistic (McHugh, 2012) The final annotations delivered by our student annotation workflow yielded a kappa statistic of  for TUSZ, for DUSZ and for EUSZ. This demonstrated the viability of using student annotators for the seizure detection task and enabled the development of a large corpus annotated data necessary for this study.

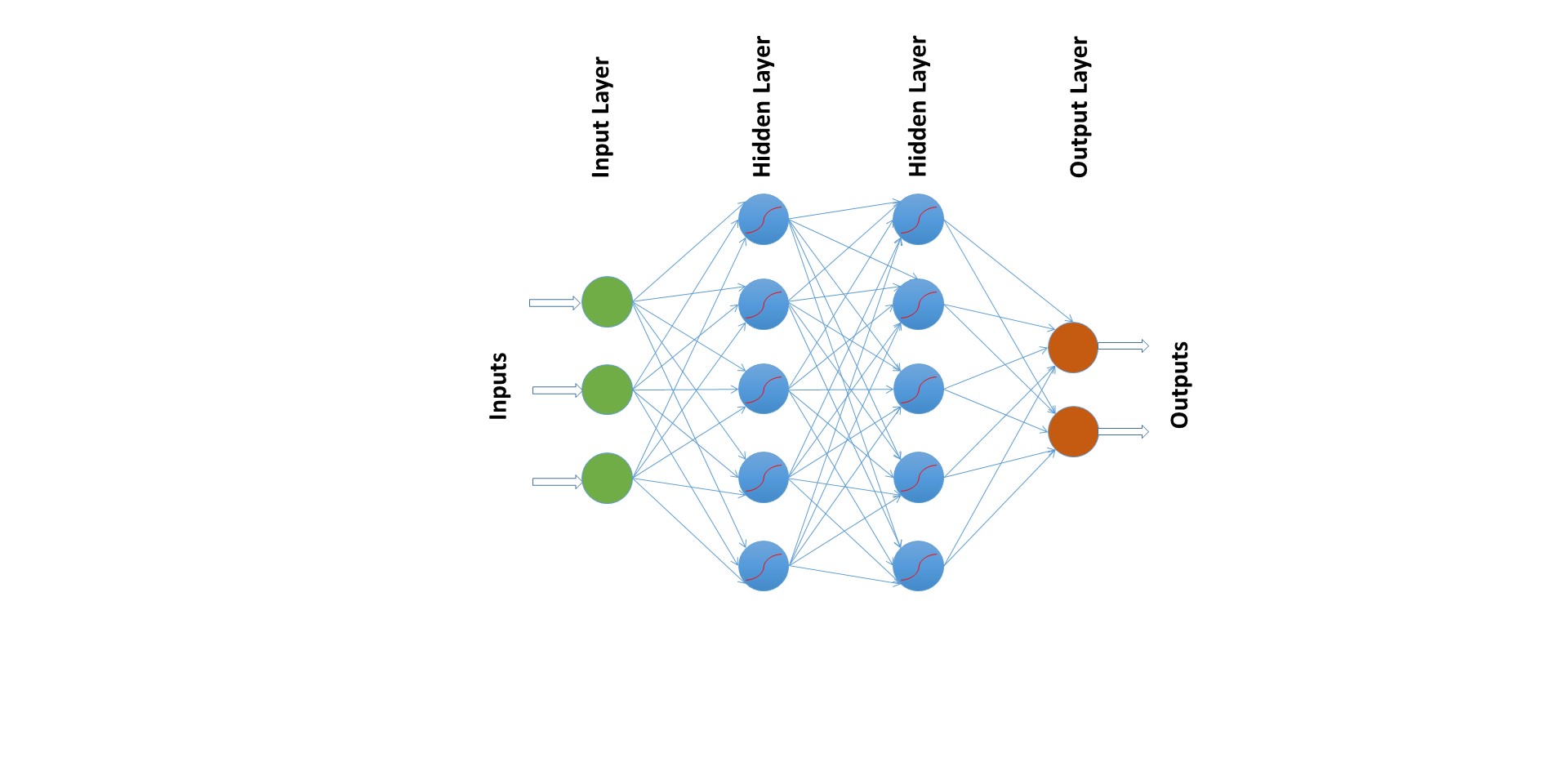
# 

**NEURAL NETWORK MODELS**

Advances in computer technology and big data resources have allowed researchers to develop a new generation of machine learning technology, often referred to as deep learning, that has significantly advanced performance in traditional disciplines like speech, image and text analysis. Multiple machine learning algorithms have been proposed for EEG event classification in the past two decades. Shoeb & Guttag et al. (2011) used support vector machines (SVM). Roy et al. (2019) used k-nearest neighbors (kNN) and convolutional neural networks (CNN) for classification of various types of seizures. Golmohammadi et al. (2018) used various hybrid machine learning architectures to establish baseline performance on TUSZ. The best performing model from this study will serve as a baseline system in this study. In this chapter, we introduce a new two-phase model design for improved segmentation performance.

## Relevant Neural Network Architectures

EEG interpretation requires integrating information from both the temporal and spatial domains. Algorithms that can efficiently process this type of information are ideal candidates for seizure detection. Sequential algorithms such as hidden Markov models (HMM) (Picone, 1990) and long short-term memory (LSTM) neural networks (Hochreiter & Schmidhuber, 1997) are well-known algorithms that can model temporal sequences. Similarly, convolutional neural networks (CNNs) (Goodfellow et al., 2016) are able to learn structural information from the data. Various hybrid architectures involving combinations of these algorithms will be discussed extensively in the following sections. However, prior to discussing such complex networks, it is important to understand the structure of a generic feedforward neural network known as a multilayer perceptron (MLP). This network forms the basis for the more complex networks discussed later in this chapter and serves as an important baseline.



**Figure 10.** Diagram of a multilayer perceptron network

### The Multilayer Perceptron

A multilayer perceptron (MLP) is a fundamental type of neural network. Each layer of the network performs an affine transform on the input data followed by a nonlinear transformation of that result (Goodfellow et al., 2016). An affine transform can be represented as:

1. *,* (1)

where the matrix and vector represent parameters that are usually set by optimizing some objective function. shows diagram of a multilayer perceptron network with a single hidden layer. The number of nodes associated with input (green), two hidden layers (blue) and output layer (red) are , , and respectively. Connections between the nodes, represented by arrows pointing from left to right, are assigned specific weights and are the parameters of the network that must be optimized during training. Note that in a typical MLP network, each neuron in a given layer is connected to all of the neurons in the predecessor layer. This is why an MLP network is often referred as a fully connected or dense network.

An MLP network exhibits powerful learning capabilities through the introduction of a nonlinearity. In order to introduce a nonlinearity, we transform *y* using a nonlinear function, also known as an activation function, such as a rectified linear unit (ReLU), a hyperbolic tangent (Tanh), or a sigmoid (Goodfellow et al., 2016). These nonlinearities control the amount of information transfer to the following sections of the network. The output of a complete MLP network can be represented as:

1. ,(2)

where , which includes the bias term in equation (), represents the weights, or parameters of the network, and represents an activation function. The superscript represents the number of layers of the network. For example, output of the first layer of an MLP network can be represented by . Similarly, can be considered as the input to the first layer.

In many hybrid classifiers, the final layer is an MLP network where the activation function is a sigmoid or softmax function (Goodfellow et al., 2016). A sigmoid function can be represented as:

1. ,(3)

while a softmax function can be defined as:

1. .(4)

Both these functions yield an output in the range . These functions compress the unbounded output of the network to a fixed range so that the output value can be used as a proxy for a posterior probability (Goodfellow et al., 2016). This final layer design is employed by many hybrid networks including the CNNs and LSTMs discussed here.

### Convolutional Neural Networks

A convolutional neural network (Jordan, 2017) is a popular neural network variant which is efficient at learning local correlations in structured data. Unlike the fully connected MLP network discussed in the previous section, CNNs operate on data on a block by block basis using a specific local parameter matrix called a convolutional kernel. Figure 11 shows the basic CNN network topology. The nodes associated with the convolutional layer in this figure are transformed vectors generated by a kernel. That is why they share the same parameters () across all input vectors.

A close up of a map

Description automatically generated

**Figure 11.** CNN network topology (Jordan, 2017)

Convolutional kernels are designed to learn abstract local patterns in data. This is best demonstrated through a simple image classification example. The outputs of the kernels associated with the initial layers of the network are designed to learn low-level properties of the image such as edges, curves and color gradients. The kernels at the later stages learn global patterns and shapes in the data such as whether the edges or curves are associated with specific objects or properties of the image. Figure 12 shows an example of how facial features are learned by the kernels as we move deeper and deeper into the network.

The outputs of the convolutional kernels are called feature maps. This is because during training, kernels attempt to learn the features associated with the data. A channel of a convolutional network, different from an EEG channel, can focus on a specific dimension of the data such as pixel color. For example, a -channel image classification system could have three sets of convolutional neural networks – one associated with each color (e.g., red, green and blue ­– if color was determined to be an important feature for classification.

A close up of a keyboard

Description automatically generated

**Figure 12.** An example of feature maps generated by CNN kernels (Lee et al., 2009)

Mathematically, output of a CNN layer can be defined as:

1. ,(5)

where the operator ‘’ indicates a convolution operation, represents the kernels of layer and represents the bias vectors associated with the kernels. In some of the models developed in this study, we will be using 2D CNN networks. In this case, the output of a 2D CNN network at layer is:

1. ,(6)

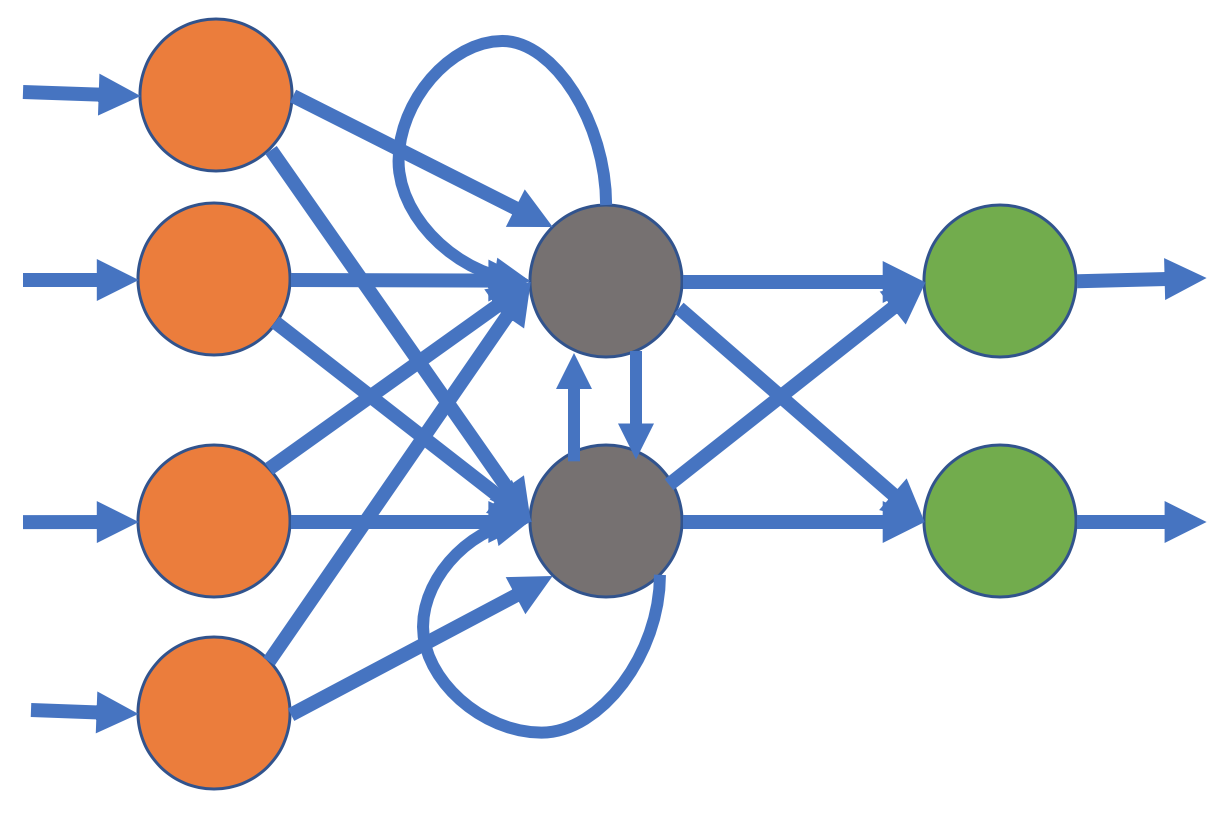
where and refer to the coordinates of a pixel. A CNN operation is performed with a sliding window by iterating over values of  and  until the entire image has been processed. In equation (6), the subscript represents the reverse indexing portion of the convolution operation. The channel and kernel numbers are represented by the variables and respectively. A bias operator, , is added at the end of the convolution operation.

It is important to note that the operations performed in equation (6) are for a single kernel associated with a CNN channel. The outputs from each channel are concatenated and passed to the next layer. This is a simplified version of the core CNN operation. It does not incorporate strides, sliding batches or zero padding operations that are essential for operating efficiently on big data sets (Vesely´ et al., 2013). We can exploit this image processing approach for EEG signals by treating the multichannel EEG signal as a 2D signal that is a function of time (e.g., sample number or frame number) and space (e.g., channel number).

### Long Short-Term Memory Networks

A long short-term memory (LSTM) network is a class of recurrent neural networks (RNN) which is popular due to its ability to learn long-term relationships in sequential data (Hochreiter & Schmidhuber, 1997). A typical network layout for an RNN is shown in . Comparing this to the MLP network shown in , we can observe that along with adjacent layers, nodes of RNNs also have connections within the layers. This suggests that the RNN networks are influenced by a node’s recurrent path (self-loops) and the state of the remaining nodes within the layer.

The architecture of an LSTM cell is shown in . An LSTM network is designed to control information flow within the network by introducing additional states which are able to remember or forget (weight) a sample of a sequence vector. These networks add an additional state vector called a cell state which aids in controlling the emphasis of new information given to its memory. Cell states are computed via following equation:



**Figure 13.** A typical recurrent neural network

1. *=*  ,(7)

1. *g* ,(8)
2. ,(9)

where , , and are internal gates which are used to compute the cell state of the LSTM network. The symbol is the element-wise multiplication operator. Element-wise multiplication, also known as a Hadamard product, performs multiplication on two matrices of the same dimension where the element at position of the first operand is multiplied by the element at position of the second operand. In equations (7)-(9), this operator is used to control the information flowing through the input gate, the output gate and the cell states. We can see from equation (8) that the gate controls the information coming from the previous state and similarly gate controls the information coming from the new input step. This weighting mechanism helps LSTMs learn patterns in longer sequences.

A picture containing clock, sitting, filled, light

Description automatically generated

**Figure 14.** Architecture of an LSTM cell (Hochreiter & Schmidhuber, 1997)

## Features

There are two general types of inputs used in modern machine learning systems: knowledge-based features or sampled data. While it is possible to input sampled data directly into a system, this approach has not shown to be particularly promising for EEG signal classification in our preliminary experiments. In this study, we will focus on knowledge, or model-based, features known as linear frequency cepstral coefficients (LFCCs) (Harati et al., 2015). LFCCs are similar to the mel-scaled features used in speech recognition known as MFCCs (Xu et al., 2004). For EEG signals, our LFCC features use linearly scaled filter bank outputs as opposed to the logarithmic scale used in MFCCs. Both speech signals and EEG signals contain crucial information in the lower frequency range of the spectrum – much lower than the sample frequency. For example, a speech signal is a quasi-periodic signal that includes a fundamental frequency related to a speaker’s vocal tract dimensions (Darwin et al., 2003). The fundamental frequency of a speech signal ranges from about  Hz to  Hz. A common sample frequency for a speech signal is  kHz. A logarithmic spectral representation is used based on knowledge of the human auditory system, which is known to be sensitive to frequency on a logarithmic scale (Picone, 1990).

An EEG signal, on the other hand, as explained previously, includes periodic discharges that have a fundamental frequency in the range of  Hz. EEG signals are most often sampled at  Hz. The frequency range from contains most of the useful information. A linear frequency representation of this range is slightly advantageous over a logarithmic representation (Harati et al., 2015).

In our standard feature extraction process, the first cepstral coefficients excluding the zeroth order coefficient are computed. A frequency domain energy term is added as the feature:

1. *.* (10)

We also include a differential energy term which has been proven effective in identification of spike-like activity that can indicate epileptiform activity. This differential energy term is computed using a sliding window approach:

1. ,(11)

where differential energy is calculated over frames ( seconds). This term attempts to model long-term change in energy of a signal.

Another computation we borrow from a classical speech recognition front end is the use of first and second order differential features (Young et al., 2006; Harati et al., 2015). We augment the base features, which we often refer to as absolute features, with first derivatives of these features, referred to as delta features, and second derivatives, referred to as delta-delta features. Note that the second derivative of the differential energy term is not used. This brings the total dimension of our feature vector to .

Rather than attempt to compute spatial features, we leave the task of learning spatial correlations of the features to the deep learning system. We will also explore better ways to select features, order features, combine features and reduce dimensionality.

## A Multiphase Recognition System

Often algorithms designed to solve the seizure detection problem analyze the signals (or its derived features) in a single iteration using a single stream. The models are expected to learn the traits of a seizure event without any knowledge of important components of these events such as spike and wave discharges. Commercially available technology (Golmohammadi et al., 2018; Persyst Corporation, 2016; Haider et al., 2016) yield extremely low performance on seizure detection tasks sensitivity with false positives per hours). These models self-organize information without direct knowledge of underlying cerebral activities. In this study, we attempt to mimic the human annotation process by splitting the seizure detection problem into two separate phases.

We develop a model which focuses on underlying epileptiform morphologies in the first phase (P1) of the system and localizes events in the second phase (P2). Specifically, the primary model learns temporal context via a channel-specific LSTM model. This is followed by learning the spatial/temporal context using a CNN-LSTM model. The second phase CNN-LSTM model uses features collected from the first phase. The results from this approach suggest that multiphase systems are able to outperform traditionally developed models.

Multiphase models have traditionally been used in many machine learning disciplines such as speech recognition. Automatic speech recognition (ASR) tools such as Kaldi (Povey et al., 2011) use multiphase models (also known as a multi-pass system) to improve system performance by iteratively increasing the complexity of the model in each phase. We follow a similar approach for our seizure detection models. Instead of emphasizing a purely data-driven approach, we separate the problem into two separate event detection tasks motivated by domain knowledge.

A very common indication of a seizure event is the existence of an epileptiform event called a spike and wave discharge (Tatum et al., 2014). These events can be observed throughout all the channels of an EEG for a generalized event or on a specific set of channels for a focal event. Sometimes a focal event can spread to other parts of the brain, which is referred to as a secondary generalized seizure. The evolution of these activities across time helps identify a seizure event. Once these activities are observed, seizures can be confirmed based on their focality, signal energy and polarities across electrodes.

The seizure detection task can be divided into two separate tasks as shown in the block diagram in . A channel-based LSTM model is used to detect epileptiform activities associated with various types of seizures. Since there is no way of confirming a seizure event without examining multiple channels of an EEG, this model only learns the shape and evolution of the discharges associated with the seizures. With this approach, seizures evolving with time (indicated by an increase in signal amplitude) are easier to identify compared to seizures with discharges with a constant amplitude and frequency.

A screenshot of a cell phone

Description automatically generated

**Figure 15.** The proposed multiphase model

This channel-based LSTM is followed by a CNN-LSTM seizure detection model which incorporates temporal and spatial context (EEG channels). This model uses two sets of features: (1) LFCC features described in Section , (2) augmented features collected from the phase 1 model. In other words, the augmented features of the P2 model are created using the posteriors extracted from the P1 model.

### Data Flow Considerations

Both models are optimized to perform a binary classification – seizure or background. This decision is made on a frame by frame basis using a predefined sliding analysis window. The P1 model was developed to identify epileptiform activities. Epileptiform activity can be associated with a seizure event, so we expect our P1 model to be confident for epileptiform morphologies belonging to a seizure but confused when the seizure is absent. We will abbreviate a seizure label as “seiz” and non-seizure label as “bckg” (short for background EEG) even though for the P1 model we are actually detecting epileptiform events.

Training is performed on minibatches (Goodfellow et al., 2017) using a variant of stochastic gradient descent (SGD) algorithm called Nesterov momentum (Sutskever et al., 2013). Nesterov momentum attempts to increase the speed of training by introducing a momentum term based on accumulated gradients of its previous steps and a correction term in the direction of the current gradient. This tends to reduce the amount of overshoot during the optimization.

The models are trained and evaluated using LFCC feature vectors discussed in Section . We use a frame duration, , of secs to represent how often features are computed. We use a window duration, , of secs to denote the amount of data used to compute the features. Windows are usually centered about frames and overlap into the preceding and future frame to provide continuity. We will denote a feature vector as . We train deep learning models using a sliding window approach with a window of duration (in seconds) and frame duration . This is the time interval over which the deep learning system will output a decision. This is set to a constant value of 1 sec for all our experiments. This is equivalent to 10 frames ().

Furthermore, we pass a window W consisting multiple epochs to the models. Window sizes for the P1 model and P2 model are 7 seconds and 11 seconds respectively (= 7 , = 11 ). These windows incorporate sufficient temporal context to learn the differences between the classes.

We can represent the reference annotations, often referred to as ground truth labels, using a categorical vector:

1. ,(12)

where is the indicator function which represents binary values 0/1 based on the index of the label and the variables , correspond to the labels “seiz” and “bckg.” For a seizure event, ; for a background event, .

Once the feature vectors and their labels are created, we can define the training set for the model as:

1. .(13)

Note that the training samples are created with a sliding window approach with a window duration of and frame duration of Each epoch serves as a center of the window (a center aligned window). For example, for the P1 model, training samples are created by a sliding window approach with the window duration of seconds and frame duration of second with label of the epoch is chosen from the epoch (the center epoch of the window). Epochs at the beginning and end of the EEG record lack the samples necessary to fill up a window. In such cases, the windows are zero padded when the values are missing.

The output of the neural network can be considered as an estimated probability mass function (PMF) of the input data distribution. The output of the neural network provides posterior values. The class estimation is done by selecting the maximum posterior values from the output neurons:

1. .(14)

Note that contains the posteriors of the model which should be in the range . Since, the classification is binary, the posterior distribution of training samples simply becomes a Bernoulli distribution.

### Channel-based Long Short-Term Memory Networks

The P1 model of the recognition system uses a -layer LSTM model to learn variants of the spike and wave discharges occurring during a seizure event. There are , and LSTM neurons in three hidden layers of this network. This is followed by two dense neurons at the output layer. The activation function used for each layer is a hyperbolic tangent function, , except for the final layer which uses a softmax nonlinearity. The complete architecture of the model is shown in .

We use a -dimensional LFCC feature vector discussed in Section for this model. This vector is concatenated by adding frames of context on both sides of the frame. The output dimensionality for each frame becomes . The static LSTM cells are used with a fixed batch size of and a window size of seconds. The data is randomly split into subsets where is used for training and is used for cross-validation during optimization. The features are normalized and scaled down to a range of on a file basis, which helps the gradient descent algorithm (and its variants) to converge much faster (Ioffe & Szegedy, 2015). Shuffling was performed on batches to avoid training biases.

The output of a window sequence of the channel-based LSTM model can be represented as:

***A close up of a map

Description automatically generated***

**Figure 16.** A channel-based long short-term memory network architecture

1. .(15)

serves as posterior value for epoch . This per-channel posterior vector is used to create an augmented feature vector for the P2 model.

### Feature Augmentation

Prior to training the P2 model, which can learn spatio-temporal properties of a seizure, we augment the existing feature vectors with the previously trained model’s hypotheses. The P1 model learns ictal discharges on a per-channel basis. The features generated by the previous models’ posteriors can provide more insight into properties of the signals. These posteriors are used to create three types of augmented features. The first type of augmented features are posterior probabilities of the individual classes estimated from equation (15). The second type of augmented features are created from estimated onsets of the detected events. For each frame of the feature vector, we calculate a distance from the onset mark and bound it within the range by applying an exponential function to it. Similar to the second type, the third type of augmented features are created from estimated offsets of the detected events. These features are bound between using the same method.

The last two types of augmented features capture temporal information about the distance from detected event’s onset and offset. These features indicate how far the current epoch/frame is from the previously detected event and how far the current epoch/frame is in the current event. That is why we will refer to these features as “raw history features.” The range in which EEG events last vary drastically. For instance, a seizure event can last from 3 seconds to several days. On the other hand, background events can last indefinitely (days if a patient is not seizing).

Patients with long duration or frequently occurring seizures are diagnosed differently than patients showing brief seizures. This is because longer recurrent seizures are considered more severe to the brain and requires immediate attention. On the other hand, brief seizures not associated with autonomic instability (loss of or nonvoluntary bodily functions) may differ unless a seizure interferes with a patient’s breathing process (Stafstrom & Carmant, 2015).

Recurrent seizures are expected to occur more closely over a very long period of time (i.e. hours or days). In some cases, recurrent seizures can occur every couple of minutes which add up to hundreds of seizures per hours (Yuan et al., 2013). That means when the first seizure is identified, we can expect more seizures in the near future. In other words, recently detected events carry more information than an event detected further back in time. We use our P1 model’s detections to include this information as features. We mark these detections as a binary vector.

1. ,(16)

where is the timestep in the record/file of duration N epochs. The event distance from the previous onset/offset mark is captured via the cumulative function:

1. ,(17)

whererepresents distance from the previous onset or offset mark. We perform this operation for both classes.

The detections are further converted into segments where each segment contains the distance from the last detected onset. If a new onset is not detected, the segment extends until the end of the record. This same approach is used for the offset segments. These segments representing the distance from last detected onsets and offsets are weighted by the equation:

1. *,* (18)

where is the time step of the segment indicating distance of the current, is a constant value that is heuristically estimated for seizure and background classes, is an offset value set to *1e-08* to introduce some denseness to the array.

The value of used for our experiments is for the seizure class and for other events such as artifacts and background event. The offset value, , is introduced because we feed these features to a CNN network in the P2 model. CNNs are designed to work with dense data. The generic CNNs used in P2 are not optimized for sparse data. In the image recognition field, to deal with the sparse representation of the data, preprocessing steps such as adding validity masks (Jaritz et al., 2018) and new variants of CNNs such as sparsity invariant CNNs have been developed (Uhrig et al., 2017). In our case, adding a small offset value avoids degradation in performance.

Equation () is used to scale the onset and offset feature dimensions for both classes. Not only does it scale the value of the feature dimension based on its distance from the event, but also bounds it within the range . This normalization helps the neural network learn faster (Ioffe & Szegedy, 2015). This feature augmentation process is summarized in . We combine three derived history features for the seizure class, three derived history features for the background class and LFCC features discussed in the Section for our P2 model.

### A CNN-LSTM Network for Modeling Context

Spatial information plays an essential role in identification of an EEG event. The P2 model is a hybrid network which consists of a -layer CNN network followed by two bidirectional LSTM layers and a -neuron MLP network. The model design is shown in . A 2D CNN network deals with data in three dimensions: samples representing time-steps of the EEG record, EEG channels, and feature dimensions. The input shape of the model is set so that kernels can stride (slide) through timesteps and EEG channels. The channels of the CNN network are associated with the dimensions of the feature vector (similar to the RGB channels of an image).

Generic CNN networks are incapable of learning translation and rotation invariance (Dieleman et al., 2016). Translation and rotation invariant systems should be able to detect target objects (such as face detection in image recognition) regardless of its location, shift or rotation in its geometry. Electrode positions belong to specific regions of the brain. Each region behaves differently and plays an important role during the interpretation process. We also avoid zero padding in the first layer of the network so it can remember the channel locations more effectively. This is because lower elements of the kernels are never exposed to the upper channels and vice versa. This way, kernels do not generalize over different brain lobes which is beneficial because brain lobes behave differently. Additionally, this also reduces the size of the feature maps generated from the first layer.

***A close up of a map

Description automatically generated***

**Figure 17.** Feature augmentation using the P1 posteriors

In the P2 model, we use the original -dimensional LFCC features augmented with six history features collected from the P1 model, making the input feature vector a ‑dimensional feature vector. The additional six features consist of posteriors, onset and offset history for both the seizure and background classes. Hyperparameters associated with the kernel sizes of the network are selected based on domain knowledge. The input shape of the network is because the input signal contains seconds ( frames) of data, TCP montage channels and features. We use seconds of temporal context following the  Hz- second rule discussed in . We use kernels during the first layer with a size of where kernels operate on samples in the temporal dimension and electrode samples in the spatial dimension. The minimum duration that a seizure can have is seconds (i.e. absence seizures). Since there are feature vectors per second, the temporal dimension is set to . We focus on channels at a time in the spatial dimension to mimic how experts divide channels while visually reviewing an EEG signal.

***A screenshot of a cell phone

Description automatically generated***

**Figure 18.** A CNN-LSTM network architecture for modeling context

The output from the first layer of the 2D-CNN layer is learned using kernels. The feature maps generated by the first layer are learned by the second layer with the same number of kernels but with half the size used in first layer. The second layer uses kernels. Maxpooling (Goodfellow et al., 2017) is performed on the feature map outputs using a kernel size of samples. The third 2D-CNN layer is identical to the second layer but uses convolutional kernels. The fourth layer of the network is a ‑dimensional (1D) CNN network. Output of the third layer is a -dimensional tensor which is reshaped to a -dimensional tensor by concatenating the vectors of last two dimensions. This reshaped data is passed to a 1D CNN layer which has kernels of size . All CNN layers use a ReLU nonlinearity function (Goodfellow et al., 2017). The output of the 1D CNN layer is further reshaped to a 1D array and passed to the LSTM sequential layers. Two LSTM layers with sizes of and are used at the end of the network. The activation used for these sequential layers is a hyperbolic tangent function.

The CNN approach used here is different from conventional approaches used for image classification. We keep the kernel size large in the beginning and avoid maxpooling to allow first two layers to learn all the underlying EEG signal properties without loss of information. Channels of the CNN network are mapped to the dimensional feature vector. Since channels of the CNNs are optimized independently, the set of CNN kernels assigned to each feature are optimized independently during the first layer of the network. From , we can see that there are kernels assigned to the first layer of the network. Because the dimensionality of the feature vector is , we have total of kernels. Each set of kernels is assigned to a dimension of the feature vector. Since we use history features, there are kernels learning solely the history collected from the P1 model’s detections.

Note that the network learns dependencies between the temporal and spatial dimensions via its convolutional kernels. The operations between kernels and input data are linear (i.e. convolution, affine, correlation). On the other hand, the onset-offset features along with each LFCC feature are processed independently in the first layer of the P2 model. Feature maps from the first layer are then combined for all convolutional channels where remaining network competes to better align the segment boundaries. This will be discussed in detail in the results section.

Embedding the history of the previous system’s detection in this way helps the following model highlight the regions of interest and specifically focus on its deficiencies. For example, we know that the model can learn duration and distance from the previously detected event directly from the history features. These augmented features force the P2 model to put more emphasis on the P1 model’s low confidence detections. The P2 model also learns the distribution of the recurrent seizures. Depending on how far the P2 model is from the previously detected event, it should be able to predict seizures occurring in the near future.

## Network Optimization

Optimization of a neural network involves minimization of a loss function, typically by using some variant of a gradient descent algorithm. The optimal set of parameters for a network, denoted, , can be represented as:

1. ,(19)

where is the loss function. We choose our loss function based on cross-entropy. For our categorical labels shown in equation () and equation (), categorical cross-entropy loss becomes:

1. .(20)

The cross-entropy is estimated using the Kullback-Leibler divergence (Goodfellow et al., 2017).

During the training process, learning rate schedules are modified in three stages. The algorithm for doing this is summarized in . During the first stage, we keep the learning rate high at and train the model for iterations. At this stage, model avoids any spurious starting points and explores the error space. The model either escapes the area of convergence or settles near more stable regions of the error surface. During the second stage, we keep the learning rate constant but keep track of the training and cross-validation losses. Assuming that the cross-validation set and training set are similar, variation in the model’s training and cross-validation losses should be within expected ranges, if the model is not overtrained. The expected range is represented as . If the model shows a huge difference between the train and cross-validation set losses, we backtrack and load the previous epoch’s weights. If two consecutive cross-validation losses are close, we can infer that the model’s performance has stagnated for the current learning rate, and we move to the third stage of training known as the annealing stage. In this stage, the learning rate, , is reduced by half after each iteration until it reaches to some preset minimum value (i.e. ). Note that, this annealing stage is different from the “learning rate annealing” method where each layer of the network is frozen and optimized with a different learning rate (Brock et al., 2017).

This technique was slightly modified for P2 models since there was no cross-validation set used. We skip stage 2 of the algorithm and simply move to the annealing stage after exploring the error space for *4* iterations.

**Algorithm 1:** An annealing learning rate with backoff

**Input**: data *trX, cvX* , size *seqLen × featDim*

**Stage1:** Exploration of the error space

**Initialize** initLR, minValError , annealFactor

**for**  *i = 1*, *2, 3, 4*

scoreHistory <= train w. high LR

**end for**

**Stage 2:** Stagnate w. Backoff

**repeat**

scoreHistory <= train & crossvalidate w. high initLR

relativeTrLoss <= TrLoss[-2] – TrLoss[-1]

relativeCvLoss <= CvLoss[-2] – CvLoss[-1]

**if**  CvLoss >

Load previous epoch weights & Shuffle

**end if**

**until** relativeCvLoss > minValError

**Stage 3:** Anneal

**repeat**

scoreHistory <= train w. learningRate

learningRate \*= annealFactor

**until**  *learningRate == ∈*

**Figure 19.** The learning rate schedule

## Postprocessing Steps

It is well known in machine learning research that a good heuristic can be very powerful tool to filter out statistical model errors. We apply a series of heuristics, summarized in , to improve the system performance. These heuristics are very important in reducing the false alarm rate to an acceptable level. The first heuristic we apply is a popular method that focuses on a model’s confidence in its output. Probabilistic filters are implemented to only consider target events which are detected above a specified probability threshold. This method tends to suppress spurious long duration events (e.g., slowing) and extremely short duration events (e.g., muscle artifacts). This decision function is applied on the seizure (target) labels only. We compare each seizure label’s posterior with the threshold value. If the posterior is above the threshold, the label is kept as is; otherwise, it is changed to the non-seizure label, which we denote as “background.”

Our second heuristic was developed after performing extensive error analysis. The most common types of errors we observed were false detections of background events as seizures (FPs) which tend to occur in bursts. Usually these erroneous bursts occur for a very small duration of time (e.g., to seconds). To suppress these, any seizure event whose duration is below a specified threshold is automatically considered as a non-seizure, or background event.

***A screenshot of a cell phone

Description automatically generated***

**Figure 20.** Postprocessing steps

Finally, we also implement a smoothing method that collapses sequences of two seizure events separated by a background event into one long seizure event. This is typically used to eliminate spurious background events. If seizures are observed in clusters separated by small intervals of time classified as background events, these isolated events are most likely part of one longer seizure event. In this method, we apply a nonlinear function that computes a pad time to extend the duration of an isolated event. If the modified endpoint of that event overlaps with another seizure event, the intervening background event is eliminated. We used a simple regression approach to derive a quadratic function that produces a padding factor:

 ,(21)

where 𝑑 is the duration of the event. This method tends to reduce isolated background events when they are surrounding by seizure events, thereby increasing the specificity.

The combination of these three postprocessing methods tends to decrease sensitivity slightly and reduce false alarms by two orders of magnitude, so their impact is significant. The ordering in which these methods is applied is important. We apply them in the order described above to achieve optimal performance.

# 

**EVALUATION METRICS**

Machine learning model optimization is typically performed in four steps: (1) feature extraction, (2) training, (3) decoding, and (4) scoring. Feature extraction is done by designing well-engineered, and in most cases reduced, quantities of variables which can still effectively represent data for the classification problem at hand. In machine learning, training is done in supervised and unsupervised fashion. Supervised learning requires ground truth labels (i.e. support vector machine) while unsupervised learning doesn’t (i.e. k-nearest neighbors). Decoding is performed to classify unknown data samples. Finally, scoring step evaluates the performance of the model against the data typically not used for training.

There are, in general, two ways to score or evaluate a machine learning technology: user acceptance testing (von Goethem & Hambling, 2013; Banchs et al., 2006) and objective performance metrics based on annotated reference data (Picone et al., 1990; Michel et al., 2017). User acceptance testing is slow, time-consuming and expensive. It has never been a practical way to guide technology development because algorithm developers need rapid turnaround times on evaluations. Hence evaluations using objective performance metrics, such as sensitivity and specificity, are common in the machine learning field. With this approach, it is very important to have a rich evaluation dataset and a performance metric that correlates well with user and application needs. The metric must have a certain level of granularity so that small differences in algorithms can be investigated and parameter optimizations can be evaluated.

## Fundamental Scores and Derived Measures

In machine learning research, performance of a system is assessed using four fundamental quantities: true positives (), true negatives (), false positives () and false negatives (). The quantities and represent the correct detections and rejections whereas and represent errors (false alarms and misses). These fundamental scores can be used to calculate further derived measures such as sensitivity, specificity, accuracy, F-scores, etc.

Most supervised machine learning algorithms output different types of bounded (class probability) or unbounded (log-likelihood) confidence measures for the hypotheses. It is possible to put a constraint on one of the performance measures, such as , and evaluate it against the other (e.g., ). A Receiver Operating Characteristic (ROC) (Hajian-Tilaki, 2013) or a Detection Error Trade-off (DET) curve (Martin et al., 1997) calculates and respectively by sweeping through a range of operating points. A ROC or DET curve provides a complete view of performance.

In subsequent sections, we will discuss various metrics which are traditionally used for scoring a temporal sequence. We have adapted these to the EEG problem and discuss their relative merits. We introduce a new metric, Time Aligned Event Scoring, that is specifically designed measure the accuracy of EEG events by carefully considering the accuracy of the segmentation information.

## Evaluation Metrics

Researchers design algorithms to optimize fundamental metrics such as , , and . For example, evaluation of voice keyword search technology was carefully studied in the Spoken Term Detection (STD) evaluations conducted by NIST (Wegmann et al., 2013; Fiscus et al., 2006). On the other hand, EEG research groups report their performance by considering an EEG event as a single target or by splitting the time axis into fixed sized frames where each frame is considered an individual target (Vidyaratne & Iftekharuddin, 2017; Knowles & Ghahramani, 2007). We will discuss five different scoring metrics two of which are adopted from speech recognition community; two which are adopted from the EEG research community; and one which was introduced to address the topic of this study. The metrics are briefly described below:

1. NIST Actual Term-Weighted Value (ATWV): based on NIST’s popular scoring package (F4DE v3.3.1), this metric, originally developed for the NIST 2006 Spoken Term Detection evaluation, uses an objective function that accounts for temporal overlap between the reference and hypothesis using the detection scores assigned by the system.
2. Dynamic Programming Alignment (DPALIGN): similar to the NIST package known as SCLite (Fiscus, 2017), this metric uses a dynamic programming algorithm to align terms. It is most often used in a mode in which the time alignments produced by the system are ignored.
3. Epoch-Based Sampling (EPOCH): treats the reference and hypothesis as temporal signals, samples each at a fixed epoch duration, and counts errors accordingly.
4. Any-Overlap (OVLP): assesses the overlap in time between a reference and hypothesis event, and counts errors using binary scores for each event.
5. Time-Aligned Event Scoring (TAES): similar to (4) but considers the percentage overlap between the two events and weights errors accordingly.

We now briefly describe each of these approaches and provide several examples that illustrate their strengths and weaknesses. These examples are drawn on a compressed timescale for illustrative purposes and were carefully selected because they demonstrate the strengths and weaknesses of the algorithms we are evaluating.

### NIST Actual Term Weighted Value (ATWV)

ATWV is a measure that balances sensitivity and FA rate. ATWV essentially assigns an application-dependent reward to each correct detection and a penalty to each incorrect detection. A perfect system results in an ATWV of 1.0, while a system with no output results in an ATWV of 0. It is possible for ATWV to be less than zero if a system is doing very poorly (for example a high rate). Experiments in voice keyword search have shown that an ATWV greater than 0.5 typically indicates a promising or usable system for information retrieval by voice applications. We believe a similar range is applicable to EEG analysis.

The metric accepts as input a list of N-tuples representing the hypotheses for the system being evaluated. Each of these N-tuples consists of a start time, end time and system detection score. These entries are matched to the reference annotations using an objective function that accounts for both temporal overlap between the reference and hypotheses and the detection scores assigned by the system being evaluated. These detection scores are often likelihood or confidence scores (Wegmann et al., 2013). The probabilities of miss and errors at a detection threshold θ are computed using:

1. (22)
2. (23)

where is the number of correct detections of terms with a detection score greater than or equal to , is the number of incorrect detections of terms with a detection score greater than or equal to , and is number of non-target trials for the term in the data. The number of non-target trials for a term is related to the total duration of source signal in seconds,, and is computed as

A screenshot of a cell phone

Description automatically generated

**Figure 21.** ATWV scores this segment as 1 TP and 5 FPs.

A close up of a person

Description automatically generated

**Figure 22.** ATWV scores this segment as 0 TP and 3 FN events.

A term-weighted value is then computed that specifies a trade-off between misses and . ATWV is defined as the value of TWV at the system’s chosen detection threshold. Using a predefined constant, , that was optimized experimentally () (Fiscus et al., 2007), ATWV is computed using:

1. (24)

A standard implementation of this approach is available at (Fiscus, 2017). This metric has been widely used throughout the human language technology community for almost 20 years.

To demonstrate the features of this approach, consider the case shown in . The hypothesis for this segment consists of several short seizure events while the reference consists of one long event. The ATWV metric will assign a score of because the midpoint of the first event in the hypothesis annotation is mapped to the long seizure event in the reference annotation. This is somewhat generous given that of the event was not detected. The remaining events in the hypothesis annotation are counted as false positives. The ATWV metric is relatively insensitive to the duration of the reference event, though the 5 false positives will lower the overall performance of the system. The important issue here is that the hypothesis correctly detected about of the seizure event, and yet because of the large number of false positives, it will be penalized heavily.

In we demonstrate a similar case in which the metric penalizes the hypothesis for missing three seizure events in the reference. Approximately of the segment is correctly identified. This type of scoring penalizing repeated events that are part of a larger event in the reference might make sense in an application like voice keyword search because in human language each word hypothesis serves a unique purpose in the overall understanding of the signal. However, for a two-class event detection problem such as seizure detection, such scoring too heavily penalizes the hypothesis for splitting a long event into a series of short events.

### Dynamic Programming Alignment (DPALIGN)

The DPALIGN metric essentially performs a minimization of an edit distance (the Levenshtein distance) (Picone et al., 1990) to map the hypothesis onto the reference. DPALIGN determines the minimum number of edits required to transform the hypothesis string into the reference string. Given two strings, the source string of length , and target string of length , we define , which is the edit distance between the substring and , as:

1. (25)

The quantities being measured here are often referred to as substitution (sub), insertion (ins) and deletion (del) penalties. For this study, these three penalties are assigned equal weights of . A dynamic programming algorithm is used to find the optimal alignment between the reference and hypothesis based on these weights. Though there are versions of this metric that perform time-aligned scoring in which both the reference and hypothesis must include start and end times, this metric is most commonly used without time alignment information.

**Ref: bckg seiz SEIZ SEIZ bckg seiz bckg**

**Hyp: bckg seiz BCKG \*\*\*\* bckg seiz \*\*\*\***

**(Hits: 4 Sub: 1 Ins: 0 Del: 2 Total Errors: 3)**

**Ref: bckg seiz BCKG \*\*\*\* bckg seiz \*\*\*\***

**Hyp: bckg seiz SEIZ SEIZ bckg seiz bckg**

**(Hits: 4 Sub: 1 Ins: 2 Del: 0 Total Errors: 3)**

**Figure 23.** DPALIGN aligns symbol sequences based on edit distance, ignoring the actual time alignments present in the reference annotation and the system output.

The metric is best demonstrated using the two examples shown in . In the first example, the reference annotation has a series of events, while the hypothesis contains events. The hypothesis substitutes background for the second seizure event, omits the third seizure event and the last background event. Hence, there are a total of three errors: two deletions and one substitution. In the second example, the reference annotation and hypothesis have been swapped to demonstrate the symmetry of the error calculations. The hypothesis generated two insertions and one substitution.

### Epoch-Based Sampling (EPOCH)

A screenshot of a cell phone

Description automatically generated

**Figure 24.** EPOCH scoring directly measures the similarity of the time-aligned annotations. TP, FN and FP are , and respectively.

Epoch-based scoring uses a metric that treats the reference and hypothesis as signals. These signals are sampled at a fixed epoch duration. The corresponding label in the reference is compared to the hypothesis. This process is depicted in . Epoch-based scoring requires that the entire signal be annotated (every second of the signal must be accounted for in the reference and hypothesis annotations), which is normally the case for sequential decoding evaluations. It attempts to account for the amount of time the two annotations overlap, so it directly addresses the inconsistencies demonstrated in and .

One important parameter to be tweaked in this algorithm is the frequency with which we sample the two annotations, which we refer to as the scoring epoch duration. It is ideally set to an amount of time smaller than the unit of time used by the classification system to make decisions. For example, the hypothesis in contains decisions made for every sec of data. The scoring epoch duration should be set less than sec. We set this parameter to  sec for most of our work because our analysis system epoch duration is typically sec. We find in situations like this the results are not overly sensitive to the choice of the epoch duration as long as it is below sec. This parameter simply controls the precision used to assess the accuracy of segment boundaries.

Because EPOCH scoring samples the annotations at fixed time intervals, it is inherently biased to weigh long seizure events more heavily. For example, if a signal contains one extremely long seizure event (e.g., secs) and two short events (e.g., each secs in duration), the accuracy with which the first event is detected will dominate the overall scoring. Since seizure events can vary dramatically in duration, this is a cause for concern.

### Any-Overlap Method (OVLP)

The OVLP metric is a popular choice in the neuroengineering community (Gotman et al., 1997; Wilson et al., 2003) due to difficulties associated with the accurate annotations of targets. OVLP is a more permissive metric that tends to produce much higher sensitivities. If an event is detected in close proximity to a reference event, the reference event is considered correctly detected. If a long event in the reference annotation is detected as multiple shorter events in the hypothesis, the reference event is also considered correctly detected. Multiple events in the hypothesis annotation corresponding to the same event in the reference annotation are counted as detected for the overlapping reference event. Duration of the events or partial overlaps are ignored during scoring which tends to yield higher sensitivities.

The OVLP scoring method is demonstrated in . It has one significant tunable parameter ­– a guard band that controls the degree to which a misalignment is still considered as a correct match. In this study, we use a guard band of zero. The guard band needs to be tuned based on the needs of the application. Sensitivity generally increases as the guard band is increased.

### Time-Aligned Event Scoring (TAES)

A screenshot of a cell phone

Description automatically generated

**Figure 25.** OVLP scoring is very permissive about the degree of overlap between the reference and hypothesis. The TP score for Example 1 is 1 with no false alarms. In Example 2, the system detects 2 out of 3 seizure events, so the TP and FN scores are 2 and 1 respectively.

Though EPOCH scoring directly measures the amount of overlap between the annotations, there is a possibility that this metric also too heavily weighs single long events. Seizure events can vary in duration from a few seconds to an hour. In some applications, correctly detecting the number of events is as important as their duration. Hence, the TAES metric was designed as a compromise between these competing constraints. The essential parameters for calculation of sensitivity and specificity such as , and for the TAES scoring metric are defined as follows:

1. (26)
2. (27)
3. (28)

where and represent the reference and hypothesis events respectively, and represents the duration of the reference events.

TAES gives equal weight to each event, but it calculates a partial score for each event based on the amount of overlap. The score is the total duration of a detected term divided by the total duration of the reference term. The score is the fraction of the time the reference term was missed divided by the total duration of the reference term. The score is the total duration of the inserted term divided by total amount of time this inserted term was incorrect according to the reference annotation. are limited to a maximum of per event. Therefore, like and , a single event contributes only a fractional amount to the overall score if it correctly detects a portion of the same event in the reference annotation (partial overlap). Moreover, if multiple reference events are detected by a single long hypothesis event, all but the first detection are considered as . These properties of the metric help manage the tradeoff between sensitivity and by balancing the contributions from short and long duration events. An example of TAES scoring is depicted in .

## A Comparison of Metrics

A simple example of how these metrics compare on a specific segment of a signal is shown in . A -sec section of an EEG signal is shown subdivided into -sec segments. The reference has three isolated events. The system being evaluated outputs one hypothesis that starts in the middle of the first event and continues through the remaining two events.

A screenshot of a cell phone

Description automatically generated

**Figure 26.** TAES. scoring accounts for the amount of overlap between the reference and hypothesis. TAES scores Example 1 as 0.71 TP, 0.29 FN and 0.14 FP. Example 2 is scored as 1 TP, 1 FN and 1 FP.

A screenshot of a cell phone

Description automatically generated

**Figure 27.** An example that summarizes the differences between scoring metrics

ATWV scores the system as   and since it assigns the extended hypothesis event to the center reference event and leaves the other two undetected. The ATWV score is for seizure events, 0.25 for background events, resulting in an average ATWV of . The sensitivity and FA rates for seizure events for this metric are and per hrs. respectively.

DPALIGN scores the system the same way since time alignments are ignored and the first event in each annotation are matched together, leaving the other two events undetected.

The EPOCH method scores the alignment , and using a -sec epoch duration because there are epochs for which the annotations do not agree and epochs where they agree. The sensitivity is and the rate per hrs. is very high because of the .

The OVLP method scores the segment as and because detected events have partial to full overlap with all the reference events, giving a sensitivity of with an rate of . TAES scores this segment as and because the first event is only correct and there are errors for the 5th to 7th and 9th epochs (an example of multiple overlapping reference events), giving a sensitivity of and a corresponding high rate.

It is difficult to conclude from this example which of these measures are most appropriate for EEG analysis. However, we see that ATWV and DPALIGN generally produce similar results. The EPOCH metric produces larger counts because it samples time rather than events. OVLP produces a high sensitivity while TAES produces a low sensitivity but a relatively higher FA rate. In the following section, we will evaluate the performance of various hybrid machine learning architectures and try to infer the behavior of these models without performing any extensive error analysis.

After evaluating each metrics’ performance on a single segment, we can understand the generic behavior of each of these metrics by comparing and contrasting their results. Comparing these metrics’ performance on a real database can provide us an empirical view about their traits. In this section, we score five hybrid architectures developed by Golmohammadi et al. (2018) and compare their results which were decoded on the files of TUSZ v1.1.1 dev-test set. We tabulate these scores in where performance is evaluated based on sensitivity, specificity and false alarm rate per hours.

**Table 2.** Performance of machine learning architectures according to derived metrics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Metric** | **Measure** | **HMM/ SdA** | **HMM/ LSTM** | **IPCA/ LSTM** | **CNN/ MLP** | **CNN/ LSTM** |
| **ATWV** | Sensitivity | 30.35% | 26.73% | 24.73% | 29.52% | 30.34% |
| Specificity | 61.38% | 68.93% | 64.51% | 65.87% | 93.15% |
| FA/24 hr. | 98.65 | 75.59 | 94.41 | 94.25 | 12.78 |
| ATWV | -0.8392 | -0.8469 | -0.4628 | -0.7971 | 0.1737 |
| **DPALIGN** | Sensitivity | 44.11% | 33.77% | 35.77% | 43.35% | 32.46% |
| Specificity | 66.87% | 72.99% | 69.59% | 71.49% | 95.17% |
| FA/24 hr. | 86.15 | 66.98 | 81.17 | 77.67 | 10.19 |
| **EPOCH** | Sensitivity | 20.71% | 50.46% | 51.02% | 65.03% | 9.784% |
| Specificity | 98.22% | 94.82% | 94.09 | 91.55% | 99.84% |
| FA/24 hr. | 1418.02 | 4133.34 | 4711.58 | 6738.82 | 125.79 |
| **OVLP** | Sensitivity | 35.35% | 30.05% | 32.97% | 39.09% | 30.83% |
| Specificity | 73.35% | 80.53% | 77.57% | 76.84% | 96.86% |
| FA/24 hr. | 77.39 | 60.92 | 73.52 | 77.19 | 6.75 |
| **TAES** | Sensitivity | 17.29% | 22.84% | 22.12% | 31.58% | 12.48% |
| Specificity | 66.04% | 70.41% | 66.64% | 64.75% | 95.24% |
| FA/24 hr. | 82.26 | 68.31 | 83.01 | 91.53 | 7.54 |

Comparing results across these five metrics can provide useful diagnostic information and provide insight into the system’s behavior. First and foremost, we tried to postprocess results of each model so that they operate around the same operating point according to OVLP. That is why the sensitivity (according to OVLP) of each model is around . Regardless of having our OVLP results within this range, other metrics show drastically different results which reflects on different properties of the metrics and models.

For example, the IPCA/LSTM and HMM/LSTM systems have relatively higher sensitivities according to the EPOCH metric, indicating that these systems tend to detect longer seizure events. Conversely, since the CNN/LSTM system has relatively low sensitivities according to the TAES and EPOCH metrics, it can be inferred that this system misses longer seizure events. Similarly, if the sensitivity was relatively high for TAES and relatively low for EPOCH, it would indicate that the system tends to detect a majority of smaller to moderate events precisely regardless of the duration of an event. A comparison of ATWV scores with other metrics gives diagnostic information such as whether a system accurately detects the onset and end of an event or whether the system splits long events into multiple short events.

# 

**EXPERIMENTAL RESULTS**

In Chapter 4, we introduced the multiphase model. In this chapter we evaluate performance of the multiphase model for seizure detection and segmentation. We will demonstrate that the P2 model (CNN-LSTM) is able to learn deficiencies of the P1 model and improve performance by delivering better segmentation. We will support this claim by providing a numerical comparison to our previous state-of-the-art system (Golmohammadi et al., 2018). Visualizations of these results will support the results of our quantitative analysis. Performance will be analyzed in terms of sensitivity and specificity using two scoring metrics, OVLP and TAES. Receiver Operating Curve (ROC) plots will also be used to confirm our findings.

**Table 3.** TUSZ v1.2.1 statistics

|  |  |  |  |
| --- | --- | --- | --- |
| **Description** | **Train** | **Devtest** | **Blindeval** |
| Patients | 264 | 50 | 50 |
| Sessions | 581 | 239 | 229 |
| Files | 1989 | 1015 | 985 |
| #Seizure events | 1254 | 679 | -- |
| Seizure Dur (sec.) | 78,838 | 58,322 | -- |
| Total Dur (sec.) | 1,188,313 | 617,102 | 647,948 |

## Database

We will use the TUH EEG Seizure Corpus (TUSZ) v1.2.1 (Shah et al., 2018) to evaluate performance. The corpus is partitioned into training (train), development test data (dev) and blind evaluation data (eval), which makes it ideal for machine learning research. A summary of the statistics is shown in . The number of seizures and their total duration for the blindeval set is not provided to avoid biases towards the dataset. The database consists a variety of seizure types including absence, tonic-clonic and myoclonic. There are two types of annotations available – channel-based and term-based. Term-based annotations are formed by making an aggregate decision for a segment based on the channel-based annotations (Ferrell et al., 2019). We used channel-based annotations to develop the P1 model, which includes channel-based LSTM decoders, and term-based annotations to develop the P2 model.

## Performance Assessment using Objective Evaluation Metrics

The overall seizure detection system uses three distinct types of processing: P1, P2 and a final heuristic postprocessor (P3). A block diagram of the multiphase model is shown in along with the performance of each state of processing. Since the P1 and P2 models are different network designs, averaging posteriors and pooling their information allows detection of a more diverse set of seizure morphologies, building on the strengths of each model. According to the OVLP metric, the multiphase model yields a sensitivity with per hours compared to the previous CNN/LSTM model with a sensitivity of with per hours. At the same operating point, according to the TAES metric, multiphase model yields a sensitivity with compared to the previous CNN/LSTM model ( sensitivity with per hours).

**Table 4.** A comparison of performance on the TUSZ dev set

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **Measure** | **CNN/LSTM Baseline** | **P1: Channel-Based LSTM** | **P2: CNN/LSTM** | **P3: Heuristic Postprocessor** |
| **OVLP** | Sensitivity | 30.83% | 39.46% | 41.16% | 40.29% |
| Specificity | 97.10% | 95.20% | 95.29% | 97.56% |
| FAs/24 hrs | 6.74 | 11.62 | 11.69 | 5.77 |
| **TAES** | Sensitivity | 11.33% | 32.57% | 32.87% | 32.59% |
| Specificity | 95.58% | 85.48% | 89.28% | 90.72% |
| FAs/24 hrs | 7.62 | 27.44 | 20.85 | 17.03 |

These two systems were also evaluated on two different blind evaluation sets: TUSZ and DUSZ. The performance of the two systems is tabulated in and . EEGs from these databases were collected using different instrumentation and environmental conditions. The TUSZ data was collected with several generations of the Natus Medical Incorporated EEG equipment (Natus, 2019) while DUSZ was collected using Nihon Kohden equipment (Nihon Kohden, 2019). These two corpora were selected to evaluate the robustness and reliability of the system. The models were not previously exposed to DUSZ training data. The baseline system’s performance is lower, yielding a very high false alarm rate of per hours and per hours for a similar level of sensitivity. The multiphase system, on the other hand, maintains an rate of while yielding a sensitivity greater than . This demonstrates the robustness of the multiphase system design.

**Table 5.** Performance on the TUSZ blind evaluation set

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **Measure** | **CNN/LSTM Baseline** | **P1: Channel-Based LSTM** | **P2: CNN/LSTM** | **P3: Heuristic Postprocessor** |
| **OVLP** | Sensitivity | 33.11% | 36.39% | 42.02% | 42.96% |
| Specificity | 92.54% | 96.55% | 96.09% | 95.53% |
| FAs/24 hrs | 19.89 | 8.74 | 10.02 | 11.45 |
| **TAES** | Sensitivity | 5.01% | 30.97% | 30.27% | 35.55% |
| Specificity | 90.43% | 93.05% | 93.87% | 91.79% |
| FAs/24 hrs | 19.89 | 13.98 | 12.75 | 17.22 |

**Table 6.** Performance on the DUSZ blind evaluation set

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Metric** | **Measure** | **CNN/LSTM Baseline** | **P1: Channel-Based LSTM** | **P2: CNN/LSTM** | **P3: Heuristic Postprocessor** |
| **OVLP** | Sensitivity | 33.71% | 42.32% | 41.75% | 43.75% |
| Specificity | 70.72% | 86.93% | 85.71% | 91.01% |
| FAs/24 hrs | 40.4 | 14.26 | 16.4 | 9.33 |
| **TAES** | Sensitivity | 19.77% | 36.47% | 35.46% | 37.13% |
| Specificity | 45.46% | 46.43% | 64.57% | 54.03% |
| FAs/24 hrs | 43.75 | 33.03 | 22.45 | 26.92 |

Performance according to the TAES metric shows the extent to which the seizure events are detected. In all three sets, the baseline system shows significant under-detection of the events. On the other hand, multiphase systems tend to maintain sensitivities around the same range but with increased false alarm rates. For example, from we can see that OVLP score of the baseline model is sensitivity with per hours. The same operating point evaluated by the TAES metric is sensitivity with per hours. On the other hand, the multiphase model suffers a decrease in sensitivity by absolute ( to ) with an increase in by per hours.

shows an extreme case for both models where the baseline system yields a much higher rate (approx. 5x). On the other hand, the multiphase system has approximately double the rate (e.g., vs. ) but with a much higher sensitivity (). TAES scores for the baseline model suffers from a decrease of in sensitivity () at a similar FA rate () whereas the multiphase model suffers from a decrease in sensitivity () with a increase in rate (). For a complete view of performance, ROC curves are analyzed next.

## Receiver Operating Characteristic (ROC) Analysis

ROC plots for the baseline system and multiphase system are shown in and for the TUSZ Corpus. In , ROC curves are shown for the DUSZ evaluation set. Though performance of all three systems in seems similar, if we focus on the region where are low (e.g., ), we see that the multiphase system generally outperforms all other systems. These results are confirmed in .

***A close up of a map

Description automatically generated***

**Figure 28.** ROC curves on the TUSZ dev set

***A close up of a map

Description automatically generated***

**Figure 29.** ROC curves on the TUSZ eval set

In , we see the difference is more pronounced on the TUSZ blind evaluation set. In , we see that the multiphase system is clearly superior over a much wider range of the ROC curve. Since neither of these systems had been previously exposed to DUSZ data, this is a strong indication that the multiphase system generalizes better. These findings are consistent with the tabulated results of through .

***A close up of a map

Description automatically generated***

**Figure 30.** ROC curves on the DUSZ Corpus

ROC curves for the TAES metric are shown in to . The TAES metric places more emphasis on the accuracy of segment boundaries. Performance of the baseline model is below the multiphase system for the entire range of the curve. The difference between P2 and P3 is very close throughout the entire FA range. The best operating point is observed around sensitivity with an FA rate of .

This trend is further supported by the results on the blind evaluation sets in and . Performance of the multiphase model is relatively low for DUSZ, which suggests that this database consists multiple prolonged isomorphic or subtle events. The

***A close up of a map

Description automatically generated***

**Figure 31.** ROC curves on the TUSZ dev set using the TAES metric

***A close up of a map

Description automatically generated***

**Figure 32.** ROC curves on the TUSZ eval set using the TAES metric

recognition systems are detecting prolonged events as multiple short events. However, the multiphase model consistently outperforms the baseline system.

## Segmentation Analysis

Although the TAES metric can provide a gross approximation of the amount of overlap between the reference and hypotheses events, certain aspects of this metric limit its ability to precisely characterize differences in performance. For instance, the TAES metric penalizes multiple overlapping events in a hypothesis as miss events. Once the duration of the detected event exceeds the reference event, the additional duration of the hypothesis event is ignored. Also, our miss rate or rate cannot exceed a value of because the metric calculates per event scores.

Accurate determination of onset and offset behavior of a seizure event pose different challenges and bear closer examination. The TAES metric, by design, does not clearly analyze the amount of over/under-detections of hypothesis events at the event boundaries. To gain better insight into the differences between these algorithms, we need alternative approaches to the analysis of segmentation performance.

***A close up of a map

Description automatically generated***

**Figure 33.** ROC curves on the DUSZ eval set using the TAES metric

We will visualize the performance of the systems as histograms where the onset and offset of an event will be analyzed independently. We will call detection an “over-detection” if it falls outside the boundary of a seizure event (reference annotation) and an “under-detection” if it falls inside the reference annotation boundary. shows an example of this analysis. The first row of the histogram shows performance based on the onset and offset boundaries. This row is the aggregate performance of detected events based on the detected event’s distance with respect to the duration of the event.

***A screenshot of a cell phone

Description automatically generated***

**Figure 34.** Baseline model segmentation for the TUSZ dev set is shown. Over-detection is shown on the left while under-detection is shown on the right. Onset boundary behavior is shown in red while offset boundary behavior is shown in green.

For example, let’s consider a reference event with a duration of seconds. If a detected event’s onset is observed outside the reference segment’s onset by one second, the over-detection value is calculated as (distance from the onset mark / duration of the reference segment). Rows 2 and 3 separate performance into onsets and offsets. Over- and under-detections are separated in the left and right side of the histograms respectively. For example, row 2 and column 1 indicates the over-detection for the onset boundary and row 3 and column 1 indicates the over-detection of the offset boundary. Finally, total under and over-detections are shown numerically for each model in the last (4th) row.

Analysis of the segment boundaries for the baseline system is shown in . It clearly indicates that the model tends to detect very small portions of the seizures with only over-detected and under-detected event boundaries. Overall, the model is off in terms of onset boundary detections and off in terms of offset boundary detections. This is clear from the first row in , where the mean values are and respectively. The smaller standard deviation values of and for onset and offset boundaries suggest that most of these hypothesis events are closer to these mean values. Higher mean value suggest that the event fall farther from the event boundaries. Specifically, under-detection of the model is far from the edge of the events which can be observed from the right columns of row 2 and 3 of the plot. The under-detected events of the model have a mean value of and from the onset and offset boundaries (with small standard deviation of 0.21 and 0.2) respectively. This suggests that the model detects a very small portion of the event when a seizure event is present, and the detections seem to be around the middle of the event segment (far from the segment boundaries).

shows performance of the baseline model on the TUSZ evaluation set. The model yields under-detected seizure events. These under-detected segments are off by 33% for onsets and 39% for offsets. The variance of these distributions are 20% and 25%, which are lower than the variances shown in for the dev set, suggesting the behavior is even more conservative in detecting seizure events.

This trend is a bit different for the final blind set, DUSZ, where the baseline model manages to over-detect and under-detect seizure events. This is shown in . The seizure events which are over-detections have mean values of and with a variance of and . This suggests most detected events are off by - of the total duration of the event. However, the higher variance suggests that these detections are separated from the mean of the hypothesis boundaries. On the other hand, under-detections have the mean values of for the onset and for the offset of the segment boundaries.

***A screenshot of a cell phone

Description automatically generated***

**Figure 35.** Baseline model segmentation for the TUSZ eval set

Since P2 is designed to optimize the segment boundaries, we expect it to show better performance. shows performance of the P2 model on the dev set. This model seems to detect over- and under-detected events comparably. The total number of under-detections are and over-detections are . However, the mean over-detection performance of this model is significantly larger (e.g., more than 200% of the duration of seizure events). The variance of the model onset over-detection is also very high which suggests that there is a significant spread in the errors. Onset boundary over-detections, shown in row 2 column 1 in red, have a maximum mean () and variance () suggesting that some of the events are detected too far in advance. The under-detections on the other hand are very close to the edges of the reference events (row 2 and 3 of column 2) yielding mean values of only and with standard deviation of only and for the onset and offset boundaries respectively. The reasons for the poor performance of the model for over-detections is the detection of extreme duration events such as very short and very long events.

***A screenshot of a cell phone

Description automatically generated***

**Figure 36.** Baseline model segmentation for DUSZ set

***A screenshot of a cell phone

Description automatically generated***

**Figure 37.** P2 model segmentation for the TUSZ dev set

shows performance of the P2 model on the TUSZ evaluation set. The number of over- and under-detected events are and respectively. Overall, the means of the onset and offset boundary detections are off by and . For the onset and offset events separately (row 2 and 3), the mean values are in-between 14-20% of the event duration with a small standard deviation of 0.17 and 0.16 for the under detection (column 2) and 0.20 and 0.40 for the over-detection (column 1). This suggests that P2 model detects seizure segment boundaries very close to the actual reference boundaries.

Finally, we evaluate performance of the P2 model on the second blind set, DUSZ, in . The performance on the DUSZ set shows similar trend observed in the datasets. The onset over-detections are off by a mean value of from the reference event’s start time with a high variance. Mean values of all the other over- and under-detections are very close to the segment boundaries ( for offset over-detection; and and for the under-detections for the onset and offsets respectively) with a very low standard deviation. The total number of over and under-detections are very close – and respectively. The overall mean of the onset/offset detections shown in row one (in blue) are well centered around the zero point, which is consistent throughout all the previous datasets.

***A screenshot of a cell phone

Description automatically generated***

**Figure 38.** P2 model segmentation for the TUSZ eval set

## Causes of Poor Segmentation

***A screenshot of a cell phone

Description automatically generated***

**Figure 39.** P2 model segmentation for the DUSZ eval set

It is important to determine the underlying causes of over- and under-detections. The potential reasons for the drastic differences lie within the annotation process/standards and the duration of the event. For example, if a seizure  seconds in duration has an estimated event boundary that is off by second, then the performance is penalized by . If this was an event  seconds in duration, the event is penalized by only . For this reason, it is important to incorporate the duration of the detected events into the evaluation process. In this section, we analyze the detection rates of the target events within specific durations and their sensitivities.

The seizure event durations are separated into different sets: (1) less than seconds, (2) - seconds, (3) - seconds, (4) - seconds and (5)  seconds and above. Other error measures such as and are not incorporated into this analysis because the segmentation quality is only analyzed for the detected events.

Performance of the baseline model and the P2 model on the seizure events based on their duration can be seen in and respectively. Note that the baseline model demonstrates poor performance on shorter events. Similarly, the performance of the model on extremely long seizure events, shown in the last row of these tables, is relatively low as well. Comparing these sensitivities with the P2 model’s performance shows that both models have similar tendencies to detect poorly on shorter seizure events. On the other hand, the P2 model demonstrates much higher performances on the longer duration seizure events. For the subset of seizures with duration greater than seconds, P2 performs at on the dev set compared to baseline model’s performance of . On the TUSZ evaluation set, the P2 model’s performance is compared to the baseline model’s performance of . On the DUSZ set, the P2 model’s performance is compared to the baseline model’s performance of .

It is important to notice that performance of the P2 model on short duration seizure events for the TUSZ dev set () contributes heavily to the segment over-detection errors observed in . Similarly, longer events are usually slowly evolving seizure events which contribute to the uncertainty in the marking of onset and offset of the seizure events which is discussed in . Seizure detection model is expected to perform poorly on these subtle events due to uncertainties involved with the slow isomorphic borderline seizures.

Higher seizure event sensitivity within this set of seizures also contributes to the increased segmentation error rates of the P2 model. For example, a -second seizure detected for only seconds or in clusters of seconds adds significantly to the miss rate. Additionally, the higher sensitivity of the P2 model suggests that the system tends to over-detect the events. Slowly evolving isomorphic events pose a real challenge for the system.

**Table 7.** Sensitivity of the baseline model as a function of event duration

|  |  |  |  |
| --- | --- | --- | --- |
| **Score/Duration (sec.)** | **TUSZ Dev Sens.** | **TUSZ Eval Sens.** | **DUSZ Sens.** |
| **0 – 10** | 10.71% | 9.75% | 0.00% |
| **10 - 30** | 28.49% | 16.83% | 15.13% |
| **30 - 120** | 32.09% | 45.87% | 34.76% |
| **120 - 300** | 56.25% | 43.39% | 59.30% |
| **> 300** | 37.93% | 54.54% | 53.84% |

**Table 8.** Sensitivity of the P2 model

|  |  |  |  |
| --- | --- | --- | --- |
| **Score/Duration (sec.)** | **TUSZ Dev Sens.** | **TUSZ Eval Sens.** | **DUSZ Sens.** |
| **0 – 10** | 27.82% | 0.00% | 0.00% |
| **10 - 30** | 27.31% | 14.28% | 16.44% |
| **30 - 120** | 53.42% | 58.62% | 44.46% |
| **120 - 300** | 48.57% | 69.23% | 69.76% |
| **> 300** | 51.72% | 83.33% | 69.23% |

## Seizure Boundary Detections Within Segment Guard Bands

Finally, we evaluate the segmentation performance based on the margin within which the seizure boundaries are estimated. We calculate number of seizure boundaries closer to the reference boundary marks. Quantitatively, we measure the distance from which the hypothesis boundaries are detected and check if that distance falls within a specific range, which we refer to as a guard band. A Guard Band is defined as the acceptable distance from the hypothesis to the reference boundary. For instance, a guard band of indicates that a detection falls within a margin of  second of the reference annotation. This measure calculates the number of seizures detected in close proximity to the reference annotation.

We evaluate performance of the detected events based on different values of guard bands, as shown in . The main comparison presented is between the baseline model and P2 (blue vs. orange). Not surprisingly, the bar plots show a clear increase in performance as the guard band duration is increased. But the rate at which the performance increases for each model is drastically different. P2 consistently outperforms the baseline model in all cases. For example, for a guard band of seconds on the dev set, the P2 model detects events compared to only events for the baseline model. This difference becomes even more dramatic for the evaluation set where the P2 model detects events as opposed to only events for the baseline system. This trend holds for all other values of the guard bands. Interestingly, differences between the models are not as great on the DUSZ data.

## Analysis of the P2 Kernels Associated with the History Features

As described in Chapter 4, P2 utilizes P1 model’s posteriors and their derived augmented features to learn its deficiencies for seizure detection and segmentation. In the P2 model, each feature dimension of the augmented feature vector is assigned to a channel in a CNN. Since each CNN channel is processed independently during the first layer of the model, we can visualize their feature maps to observe what a trained kernel has learned. The total number of channels for P2 is which includes LFCC-based features and

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

**Figure 40.** Event detections within the guard bands of the reference annotation boundaries

 posterior-derived features. Therefore, there are types of kernels learning CNN features independently. In this section, we visualize the feature maps generated by the CNN kernels.

However, prior to visualizing such data, we focus on a few seizures that were not learned by the same CNN-LSTM model designs when only LFCC features were used. Two sets of experiments were conducted for this study. The first set of models were only trained using conventional -dimensional LFCC feature vectors. The second set used the dimensional augmented feature vector which is used in our P2 model. The model with only LFCC features performs at a sensitivity of with per  hours whereas the P2 model performs at a sensitivity of within the same range. For the visualization purposes we will use examples of the seizure events which were only detected by the P2 model and see how posterior based kernels behave on such seizure events.

After the P2 model is optimized, we pass two different examples of seizure events through the model. These examples were the seizure events that were only detected when posterior features were used. Feature maps generated by posterior kernels for the first seizure example are shown in and . These generalized seizure morphologies last for about - seconds and are recurrent every - seconds suggesting that patient was suffering a status epilepticus at the time of recording. The spectrogram-like plots from contain kernel outputs where each kernel is associated with the index of the feature dimension. Kernels are abbreviated with letter which is followed by an index of the feature dimension. For example, indicates the kernel associated with feature which is the energy of the signal. Kernel is associated with the first cepstral coefficient. Similarly, the last six dimensions (from to ) are associated with the posterior-derived history features. It is clear from that both types of kernels, the ones associated with the LFCC features and the ones with the posterior features, learn seizure morphologies quite well and they also align well with the actual seizure onset and offset marks. This is apparent by examining the vertical strips of feature maps created by each kernel. These vertical strips can be observed in the feature maps from sample  to .

***A close up of a computer

Description automatically generated***

**Figure 41.** Layer-1 CNN kernels for recurring seizures (during the event)

shows the same recurring seizure events with their feature maps using a -second window. There are two seizure events in this plot. The onset and offset are indicated by the red and green arrows respectively. The signal waveform and their feature maps for spectral features (-), posteriors ( and ), onset/offset history features (-) are well aligned with both seizure events. There is a clear difference between the background and seizure events. The transformed feature maps show a gradual increase in the probability of the seizure (viewed as wider vertical strips) which is consistent with the posterior values observed in the hypothesis generated by the P1 model. It is important to note that such seizures are easy to segment visually because there are clear abrupt changes in the morphologies.

A picture containing computer

Description automatically generated

**Figure 42.** Layer-1 CNN kernels for recurring seizures (between the event)

The seizure morphology shown in represents a subtle seizure which poses challenge for annotators as well as seizure detection systems. This plot is made on the same scale ( seconds) as . This seizure is observed on only EEG channels and very slowly transforms to a seizure (the onset is represented by red arrows) from a background state. The feature maps shown in show no apparent patterns learned by their LFCC kernels (rows -). Although, only the first three dimensions of the feature vector are shown, this is consistent for all LFCC features. In contrast, posterior kernels - show a clear gradual change in their feature maps.

It is important to notice the misalignments created by the posterior kernels. Such misalignments convey the confusion inherent in the P1 model. Compared to the seizures observed in , strips observed from the feature maps created by the onset versus offset kernels (e.g., versus ) in are sparse (focal to only channels) and dispersed in the temporal domain. These misaligned feature maps behave as intermediate features of the network which forces the following layers to further fine tune the segment boundaries. Such misalignments are observed during subtle seizures.

The results presented in this chapter validate our hypothesis that the posteriors and their derived features carry important temporal information and can be exploited to deliver improved segmentation performance. Such a design alleviates the need for complicated models that focus on long-term dependencies. When the variance of event durations is high, embedding the posterior information into a feature vector tends to overcome the issues associated with long-term dependencies. Additionally, the multiphase model tends to fine tune the segment boundaries of the seizures by learning the deficiencies of the earlier stages of the models.

A screenshot of a computer

Description automatically generated

**Figure 43.** Layer-1 CNN kernels for subtle seizures

# 

**FUTURE PLANS**

The focus of this dissertation proposal is the underappreciated problem of segmentation of seizure events. This can be directly attributed to the difficulties associated with the annotation process and the poor performance of automated seizure detection algorithms. We separate the seizure detection problem into two phases. In the first phase, we develop a channel-specific model to learn epileptiform activity associated with seizures. New features are derived based on event posteriors and added to the traditional feature vector. These posterior features contain information about previously detected events and the model’s confidence of a detected event.

These augmented features are used in the second phase to integrate spatial and temporal context. We show that the P2 model is able to improve the seizure detection performance and fine tune the segment boundaries detected by the P1 model. This proposed method is able to achieve performance of sensitivity with false alarms / hours compared to baseline model’s performance of sensitivity with false alarms / hours. This proposed method detects segment boundaries within a two second margin of error compared to segments for the baseline model.

## Remaining Research Tasks

Models which use posterior derived features from the P1 model perform at much higher sensitivities () compared to models without these features (). Feature maps of the CNN kernels associated with the posterior features show that the kernels are able to learn history about the previously detected seizures. The activation patterns in the network are spread far apart for a slowly evolving (subtle) seizure and stay closely clustered for the easier seizures. We will explore how the remainder of the network contributes to better segmentation. A simple diagnostic test can be performed by checking the importance of features. Fisher et al. (2018) studies the importance of features for neural network algorithms. The authors introduce a permutation feature importance algorithm in which a trained model is fed features with shuffled indices.

Instead of permuting all features, we will evaluate the performance of a subset of features which include onset/offset features and posteriors. By simply ignoring these features or changing their indices and plotting the corresponding feature maps, we will gain insight into what the remainder of the network is learning. This is a preferred option to training separate models for different sets of features (Jean-Paul et al., 2019) because the decoding process is much faster than training.

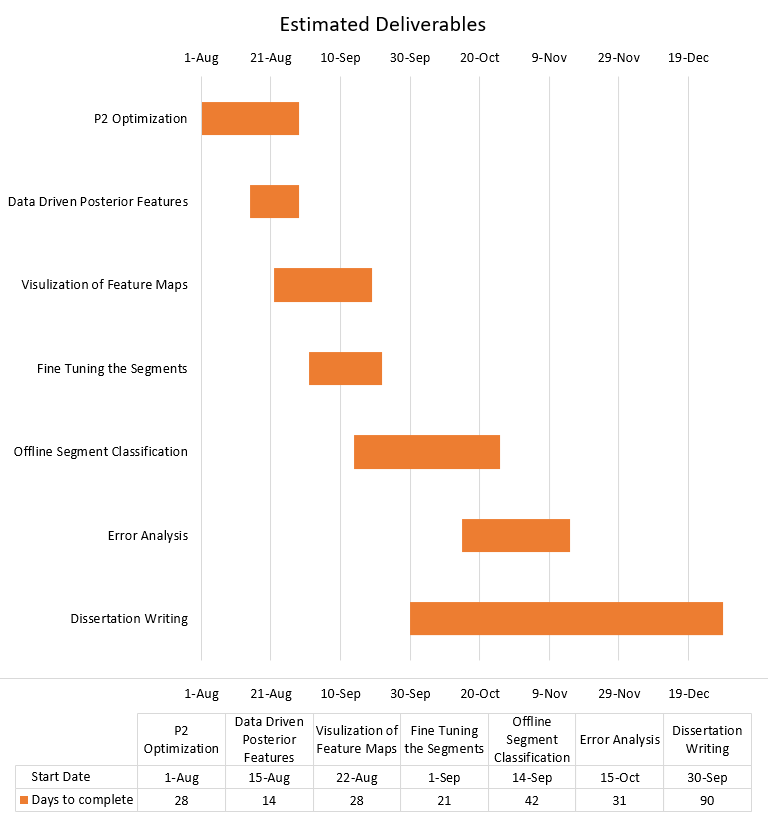
From , we can observe that segment boundaries are detected within a five-second margin of error. These models will be optimized using a ‑second context around the detected boundary. We will first develop a crude model that extracts the energy of the signal from each second of data and estimates the maximum difference in energy. This model will serve as a baseline model. We will train LSTM models on the original data to find the optimal threshold for this differential energy term. These models will be used as a postprocessing step after a seizure has been detected.

Models developed in this study are trained using a sliding window approach. The window size is a fixed value (typically around seconds) which hinders a model’s ability to incorporate the wider context of a seizure event (e.g., a seizure that lasts for minutes). For example, the P2 model operates on a window size of seconds cannot learn the evolution of a -second seizure. We will implement an off-line seizure detection algorithm which can be trained on an entire seizure segment regardless of its duration. This will help the system model the complete evolution of a seizure from its onset until the offset. This model will be implemented as a postprocessing step and is expected to reduce the false alarm rate further.

Finally, the seizure events which are detected by the multiphase model but segmented very poorly convey some important information about the deficiencies of the model. We will investigate the signal morphologies, seizure types, and clinical correlates (information collected from the report) during the recording for these events. We will investigate if this information correlates with a patient’s underlying pathologies (e.g., kidney or liver failure).

## Outline of Future Plans

shows a Gantt chart for the remaining part of the dissertation research. The dissertation proposal material constitutes about 60% of the final dissertation. The remaining work can be broken down into three major activities: error analysis, experimentation and writing. We expect the final dissertation document to be completed by December 2020 with the defense occurring in early Spring 2021.



**Figure 44.** Dissertation completion plan

## Publications

Below we list published work and publications under development.

### Work Published or Accepted for Publication

**Book Chapter(s):**

* Golmohammadi, M., Shah, V., Obeid, I., & Picone, J. (2020). Deep Learning Approaches for Automatic Seizure Detection from Scalp Electroencephalograms. In I. Obeid, I. Selesnick, & J. Picone (Eds.), Signal Processing in Medicine and Biology: Emerging Trends in Research and Applications (1st ed., pp. 233–274). New York, New York, USA: Springer. *https://doi.org/10.1007/978-3-030-36844-9*.

**Refereed Journal Publication(s):**

* Shah, V., von Weltin, E., Lopez, S., McHugh, J. R., Veloso, L., Golmohammadi, M., … Picone, J. (2018). The Temple University Hospital Seizure Detection Corpus. *Frontiers in Neuroinformatics*, *12*, 1–6. *https://doi.org/10.3389/fninf.2018.00083*.

**Peer-Reviewed Conference Paper(s):**

* Golmohammadi, M., Ziyabari, S., Shah, V., Obeid, I., & Picone, J. (2018). Deep Architectures for Spatio-Temporal Modeling: Automated Seizure Detection in Scalp EEGs. In *Proceedings of the IEEE International Conference on Machine Learning and Applications* (ICMLA) (pp. 745–750). Orlando, Florida, USA. *https://doi.org/10.1109/ICMLA.2018.00118*.
* Shah, V., Golmohammadi, M., Ziyabari, S., von Weltin, E., Obeid, I., & Picone, J. (2017). Optimizing Channel Selection for Seizure Detection. In I. Obeid & J. Picone (Eds.), *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium* (pp. 1–5). Philadelphia, Pennsylvania, USA: IEEE. *https://doi.org/10.1109/SPMB.2017.8257019*.
* Golmohammadi, M., Ziyabari, S., Shah, V., Obeid, I., & Picone, J. (2017). Gated Recurrent Networks for Seizure Detection. In I. Obeid & J. Picone (Eds.), *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium* (pp. 1–5). Philadelphia, Pennsylvania, USA: IEEE. *https://doi.org/10.1109/SPMB.2017.8257020*.
* von Weltin, E., Ahsan, T., Shah, V., Jamshed, D., Golmohammadi, M., Obeid, I., & Picone, J. (2017). Electroencephalographic Slowing: A Primary Source of Error in Automatic Seizure Detection. In I. Obeid & J. Picone (Eds.), *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium* (pp. 1–5). Philadelphia, Pennsylvania, USA: IEEE. *https://doi.org/10.1109/SPMB.2017.8257018*.

**Peer-Reviewed Conference Abstracts:**

* Shah, V., Anstotz, R., Obeid, I., & Picone, J. (2018). Adapting an Automatic Speech Recognition System to Event Classification of Electroencephalograms. In I. Obeid & J. Picone (Eds.), *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium* (SPMB) (p. 1). Philadelphia, Pennsylvania, USA. h*ttps://doi.org/10.1109/SPMB.2018.8615625*.

**Other:**

* Lin, R., Marquez, D., Jacobson, M., Castaldi, H., Buckman, S., Shah, V., & Picone, J. (2020). Accuracy of Automated Machine Learning Software in Identifying EEGs with Prolonged Seizures. In Annual Meeting of the American Academy of Neurology (AAN) (p. P6.002). Toronto, Canada. *https://www.isip.piconepress.com/publications/conference\_presentations/2020/aan/prolonged\_seizures/*.
* Ziyabari, S., Shah, V., Obeid, I., & Picone, J. (2018). Curriculum Learning Based on Sample Selection Using Posterior Probabilities. In 26th Conference on Intelligent Systems for Molecular Biology (p. 1). Chicago, Illinois, USA. *https://www.isip.piconepress.com/publications/conference\_presentations/2018/ismb/cl/*.

### Publications Under Development

**Book Chapter(s):**

* Shah, V., Golmohammadi, M., Obeid, I., & Picone, J. (2021). Objective Evaluation Metrics for Automatic Classification of EEG Events. In I. Obeid, I. Selesnick, & J. Picone (Eds.), *Signal Processing in Medicine and Biology: Emerging Trends in Research and Applications* (1st ed., pp. 1–26). New York City, New York, USA: Spring 2021.

**Refereed Journal Publication(s):**

* Shah, V., von Weltin, E., Ahsan, T., Obeid, I., & Picone, J. (2021). On the Use of Non-Experts for Generation of High-Quality Annotations of Seizure Events. *Clinical EEG and Neuroscience*.
* Shah, V., Campbell, C., Obeid, I., & Picone, J. (2021). Improved Spatio-Temporal Modeling in Automated Seizure Detection using Channel-Dependent Posteriors. *Neurocomputing*.
* Shah, V., Obeid, I., & Picone, J. (2021) Artifact Reduction Techniques for Improving the Performance of Seizure Detection Software. *Informatics in Medicine Unlocked*.

**Peer-Reviewed Conference Paper(s):**

* Shah, V., Iskander, R., Roy, Y., Obeid, I., & Picone, J. (2020). Validation of Temporal Scoring Metrics for Automated Seizure Detection. *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium* (SPMB), December 2020.
* Shah, V., Obeid, I., & Picone, J. (2021) Analysis of Spectral Features for EEG Seizure Detection. *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium* (SPMB), December 2021.

# REFERENCES

Albuja AC, Murphy PB. Absence Seizure. [Updated 2020 Jul 1]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan-. *https://www.ncbi.nlm.nih.gov/books/NBK499867/*.

American Academy of Neurology (2020). Policy & Guidelines for EPILEPSY. Retrieved July 1, 2020, from *https://www.aan.com/Guidelines/home/ByTopic?topicId=23*.

American Clinical Neurophysiology Society. (2012). ACNS Standardized ICU EEG Nomenclature v. 2012. Retrieved December 4, 2018, from *https://www.acns.org/pdf/guidelines/Guideline-14-pocket-version.pdf*.

Banchs, R., Bonafonte, A., & Perez, J. (2006). Acceptance Testing of a Spoken Language Translation System. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)* (p. 106). Genoa, Italy. *http://www.lrec-conf.org/proceedings/lrec2006/pdf/60\_pdf.pdf*.

Benbadis, S. (2014). EEG Training for Neurologists Is Inadequate. Retrieved July 1, 2020, from Epilepsy Foundation website: *https://www.epilepsy.com/article/2014/12/eeg-training-neurologists-inadequate*.

Benbadis, S. R. (2010). The tragedy of over-read EEGs and wrong diagnoses of epilepsy. *Expert Review of Neurotherapeutics*, *10*(3), 343–346. *https://doi.org/10.1586/ern.09.157*.

Brain Electrical Source Analysis (BESA Research 7.0). (2020). Retrieved July 1, 2020, from *https://www.besa.de/*.

Brock, A., Lim, T., Ritchie, J., & Weston, N. (2017). FreezeOut: Accelerate Training by Progressively Freezing Layers. *ArXiv.Org*. *https://arxiv.org/abs/1706.04983*.

Capp, N., Krome, E., Obeid, I., & Picone, J. (2017). Facilitating the Annotation of Seizure Events Through an Extensible Visualization Tool. In I. Obeid & J. Picone (Eds.), *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium* (p. 1). *https://doi.org/10.1109/SPMB.2017.8257043*.

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., & Buchanan, B. G. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, *34*(5), 301–310. *https://doi.org/10.1006/jbin.2001.1029*.

Chong, D. J., & Hirsch, L. J. (2005). Which EEG Patterns Warrant Treatment in the Critically Ill? Reviewing the Evidence for Treatment of Periodic Epileptiform Discharges and Related Patterns. *Journal of Clinical Neurophysiology*, *22*(2), 79–91. *https://doi.org/10.1097/01.WNP.0000158699.78529.AF*.

Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *Journal of the Acoustical Society of America, 114*(5), 2913–2922. *https://doi.org/10.1121/1.1616924*.

Dieleman, S., De Fauw, J., & Kavukcuoglu, K. (2016). Exploiting Cyclic Symmetry in Convolutional Neural Networks. In *arXiv*. *https://arxiv.org/pdf/1602.02660.pdf*.

Encevis Solutions. (2020). NeuroTrend: Trending for continuous EEG in critical care. Retrieved July 1, 2020 from *https://www.encevis.com/solutions/eeg-trending-for-icu/*.

Ferrell, S., Mathew, V., Ahsan, T., & Picone, J. (2020). The Temple University Hospital EEG Corpus: Electrode Location and Channel Labels. Philadelphia, Pennsylvania, USA. *https://www.isip.piconepress.com/publications/reports/2020/tuh\_eeg/electrodes*.

Fiscus, J. G. (2017). Speech Recognition Scoring Toolkit. *https://doi.org/https://github.com/usnistgov/SCTK*.

Fiscus, J., Ajot, J., Garofolo, J., & Doddingtion, G. (2007). Results of the 2006 Spoken Term Detection Evaluation. In *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR) Workshop “Searching Spontaneous Conversational Speech”* (pp. 45–50). Amsterdam, Netherlands. *https://www.nist.gov/publications/results-2006-spoken-term-detection-evaluation*.

Fisher, A., Rudin, C., & Dominici, F. (2018). All Models are Wrong, but Many are Useful: Learning a Variable’s Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81. *https://arxiv.org/abs/1801.01489*.

Golmohammadi, M., Ziyabari, S., Shah, V., Obeid, I., & Picone, J. (2018). Deep Architectures for Spatio-Temporal Modeling: Automated Seizure Detection in Scalp EEGs. In *Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 745–750). Orlando, Florida, USA. *https://doi.org/10.1109/ICMLA.2018.00118*.

Gotman, J., Flanagan, D., Zhang, J., & Rosenblatt, B. (1997). Automatic seizure detection in the newborn: Methods and initial evaluation. *Electroencephalography and Clinical Neurophysiology*, *103*(3), 356–362. *https://doi.org/10.1016/S0013-4694(97)00003-9*.

Gross, A., Lopez, S., Obeid, I., & Picone, J. (2016). Statistical Comparison of REF vs. LE Referenced EEG Signals. In *Temple University College of Engineering Research Experience for Undergraduates Conference* (p. 1). Philadelphia, Pennsylvania, USA. *https://www.isip.piconepress.com/publications/unpublished/conferences/2016/summer\_of\_code/features/*.

Grozea, C., Voinescu, C. D., & Fazli, S. (2011). Bristle-sensors - Low-cost flexible passive dry EEG electrodes for neurofeedback and BCI applications. In Journal of Neural Engineering. *https://doi.org/10.1088/1741-2560/8/2/025008*.

Haider, H. A., Esteller, R. D., Hahn, C. B., Westover, M. J., Halford, J. W., Lee, J. M., & LaRoche, S. (2016). Sensitivity of quantitative EEG for seizure identification in the intensive care unit. *Neurology*, *87*(9), 935–944. *https://doi.org/10.1212/WNL.0000000000003034*.

Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian Journal of Internal Medicine*, *4*(2), 627–635. *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/*.

Halford, J. J., Shiau, D., Desrochers, J. A., Kolls, B. J., Dean, B. C., Waters, C. G., … LaRoche, S. M. (2015). Inter-Rater Agreement on Identification of Electrographic Seizures and Periodic Discharges in ICU EEG Recordings. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *126*(9), 1661–1669. *https://doi.org/10.1016/j.clinph.2014.11.008*.

Harati, A., Golmohammadi, M., Lopez, S., Obeid, I., & Picone, J. (2015). Improved EEG Event Classification Using Differential Energy. In I. Obeid & J. Picone (Eds.), *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium* (pp. 1–4). Philadelphia, Pennsylvania, USA. *https://doi.org/10.1109/SPMB.2015.7405421*.

Hirsch L., J., Laroche S., M., Gaspard N., T., Gerard E., F., Svoronos A., B., Herman S., W., … Drislane F., W. (2013). American Clinical Neurophysiology Society’s Standardized Critical Care EEG Terminology: 2012 version. *Journal of Clinical Neurophysiology*, *30*(1), 1–27. *https://doi.org/10.1097/WNP.0b013e3182784729*.

Hirsch, L. J. (2010). Urgent Continuous EEG (cEEG) Monitoring Leads to Changes in Treatment in Half of Cases. *Epilepsy Currents*, *10*(4), 82–85. *https://doi.org/10.1111/j.1535-7511.2010.01364.x*.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. *https://doi.org/10.1162/neco.1997.9.8.1735*.

Hosseini, M.-P., Hajisami, A., & Pompili, D. (2016). Real-Time Epileptic Seizure Detection from EEG Signals via Random Subspace Ensemble Learning. In *Proceedings of the IEEE International Conference on Autonomic Computing (ICAC)* (pp. 209–218). Wurzburg, Germany. *https://doi.org/10.1109/ICAC.2016.57*.

Ian Goodfellow, Bengio, Y., & Courville, A. (2016). *Deep Learning*. *Nature Methods* (1st ed.). Cambridge, Massachusetts, USA: MIT Press. *https://mitpress.mit.edu/books/deep-learning*.

Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 448–456). Lille, France. *https://arxiv.org/abs/1502.03167*.

Jaritz, M., de Charette, R., Wirbel, E., Perrotton, X., & Nashashibi, F. (2018). Sparse and Dense Data with CNNs: Depth Completion and Semantic Segmentation. *ArXiv*. *Retrieved from https://arxiv.org/abs/1808.00769*.

Jean-Paul, S., Elseify, T., Obeid, I., & Picone, J. (2019). Issues in the Reproducibility of Deep Learning Results. In I. Obeid & J. Picone (Eds.), *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* (pp. 1–4). Philadelphia, Pennsylvania, USA: IEEE. *https://doi.org/10.1109/SPMB47826.2019.9037840*.

Jeremy Jordan. (2017). Convolutional neural networks. Retrieved July 1, 2020, from *https://www.jeremyjordan.me/convolutional-neural-networks/*.

Jing, J., D’Angremont, E., Zafar, S., Rosenthal, E. S., Tabaeizadeh, M., Ebrahim, S., … M. Brandon Westover. (2018). Rapid Annotation of Seizures and Interictal-ictal Continuum EEG patterns. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Honolulu, Hawaii, USA. *https://doi.org/10.1109/EMBC.2018.8513059*.

Kesavadas C, Thomas B. Clinical applications of functional MRI in epilepsy. Indian J Radiol Imaging 2008;18:210-7 retrieved July 25, 2020 from *http://www.ijri.org/text.asp?2008/18/3/210/41829*.

Klatchko, A., Raviv, G., Webber, W., & Lesser, R. (1998). Enhancing the detection of seizures with a clustering algorithm. *Electroencephalography and Clinical Neurophysiology*, *106*(1), 52–63. *https://doi.org/10.1016/S0013-4694(97)00092-8*.

Knowles, D., & Ghahramani, Z. (2007). Infinite Sparse Factor Analysis and Infinite Independent Components Analysis. In M. Davies, C. James, S. Abdallah, & M. Plumbley (Eds.), *Independent Component Analysis and Signal Separation: Proceedings of the 7th International Conference on Independent Component Analysis and Signal Separation* (pp. 381–388). Berlin, Germany: Springer Berlin Heidelberg. *https://doi.org/10.1007/978-3-540-74494-8\_48*.

Wyllie, E. (2015). *Wyllie’s treatment of epilepsy: principles and practice*. (E. Wyllie, B. Gidal, H. Goodkin, T. Loddenkemper, & J. Sirven, Eds.), *Wolters Kluwer* (6th ed.). China: Wolters Kluwer. https://doi.org/10.1016/j.yebeh.2015.08.030

Krauss, G. L., & Fisher, R. S. (2011). *The Johns Hopkins Atlas of Digital EEG: An Interactive Training Guide* (2nd ed.). Baltimore, Maryland, USA: Johns Hopkins University Press. *https://www.amazon.com/Johns-Hopkins-Atlas-Digital-EEG/dp/0801897335*.

Kubota, Y., Nakamoto, H., Egawa, S., & Kawamata, T. (2018). Continuous EEG monitoring in ICU. Journal of Intensive Care, 6(1), 39. *https://doi.org/10.1186/s40560-018-0310-z*.

Lasitha S. Vidyaratne, & Khan M. Iftekharuddin. (2017). Real-Time Epileptic Seizure Detection Using EEG. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *25*(11), 2146–2156. *https://doi.org/10.1109/TNSRE.2017.2697920*.

Liu, J., Dai, T., Elster, T., Sahgal, V., Brown, R., & Yue, G. (2000). Simultaneous measurement of human joint force, surface electromyograms, and functional MRI-measured brain activation. *Journal of Neuroscience Methods*, *101*(1), 49–57. *https://doi.org/10.1016/S0165-0270(00)00252-1*.

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations. In *International Conference on Machine Learning (ICML)* (pp. 609–616). Montreal, Canada. *https://doi.org/10.1145/1553374.1553453*.

Lopez, S. (2017). Automated Identification of Abnormal Adult Electroencephalograms. Temple University. *https://www.isip.piconepress.com/publications/ms\_theses/2017/abnormal/thesis/*.

Lopez, S., Golmohammadi, M., Obeid, I., & Picone, J. (2016). An analysis of two common reference points for EEGs. In I. Obeid & J. Picone (Eds.), *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium* (pp. 1–4). Philadelphia, Pennsylvania, USA. *https://doi.org/10.1109/SPMB.2016.7846854*.

Nuwer, M. (1997). Assessment of digital EEG, quantitative EEG, and EEG brain mapping: Report of the American Academy of Neurology and the American Clinical Neurophysiology Society. *Neurology*, *49*(1), 277–292. *https://doi.org/10.1212/WNL.49.1.277*.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The DET curve in assessment of detection task performance. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)* (pp. 1895–1898). Rhodes, Greece. *http://www.isca-speech.org/archive/eurospeech\_1997/e97\_1895.html*.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282. *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/*.

Michel, M., Joy, D., Fiscus, J. G., Manohar, V., Ajot, J., & Barr, B. (2017). Framework for Detection Evaluation (F4DE). Retrieved May 16, 2017, from *https://github.com/usnistgov/F4DE*.

Natus Medical. (2018). Nicolet® NicVue Connectivity Solution. August 13, 2018. Retrieved from *https://neuro.natus.com/products-services/nicolet-nicvue-connectivity-solution*.

Nihon Kohden Corporation. July 1, 2020. Retrieved from *https://us.nihonkohden.com/products/eeg-1200*.

Obeid, I., & Picone, J. (2016). The Temple University Hospital EEG Data Corpus. In M. A. Lebedev (Ed.), *Augmentation of Brain Function: Facts, Fiction and Controversy. Volume I: Brain-Machine Interfaces* (1st ed., Vol. 10, pp. 394–398). Lausanne, Switzerland: Frontiers Media S.A. *https://doi.org/10.3389/fnins.2016.00196*.

Ochal, D., Rahman, S., Ferrell, S., Elseify, T., Obeid, I., & Picone, J. (2020). The Temple University Hospital EEG Corpus: Annotation Guidelines. *https://www.isip.piconepress.com/publications/reports/2020/tuh\_eeg/annotations/*.

Orosco, L., Correa, A. G., Diez, P., & Laciar, E. (2016). Patient non-specific algorithm for seizures detection in scalp EEG. *Computers in Biology and Medicine*, *71*, 128‑134. *https://doi.org/10.1016/j.compbiomed.2016.02.016*.

Pathirana, S., Asirvatham, D., & Johar, G. (2018). A Critical Evaluation on Low-Cost Consumer-Grade Electroencephalographic Devices. In 2018 2nd International Conference on BioSignal Analysis, Processing and Systems (ICBAPS) (pp. 160‑165). *https://doi.org/10.1109/ICBAPS.2018.8527413*.

Persyst Development Corporation (2020). Seizure Detection: P14 Seizure Detection is shifting the curve. Retrieved June 18, 2020, from *https://www.persyst.com/technology/seizure-detection/*.

Persyst Development Corporation. (2017). Seizure Detection (P13 rev. B). Retrieved January 1, 2017, from *https://www.persyst.com/ technology/seizure-detection/*.

Picone, J. (1990). Continuous Speech Recognition Using Hidden Markov Models. *IEEE Acoustics, Speech, and Signal Processing Society (ASSP)*, *7*(3), 26–41. *https://doi.org/10.1109/53.54527*.

Picone, J., Obeid, I., & Harabagiu, S. (2018). Automated Cohort Retrieval from EEG Medical Records. *26th Conference on Intelligent Systems for Molecular Biology*, 1‑7. *https://www.isip.piconepress.com/publications/conference\_presentations/2018/ismb/cohort\_retrieval/*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., … Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. *Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, 1–4. Retrieved from *http://www.researchgate.net/publication/228828379\_The\_Kaldi\_speech\_recognition\_toolkit/file/79e4150743dc6ce65c.pdf*.

Ronner, H. E. E., Ponten, S. C. C., Stam, C. J. J., & Uitdehaag, B. M. J. M. J. (2009). Inter-observer variability of the EEG diagnosis of seizures in comatose patients. *Seizure*, *18*(4), 257–263. *https://doi.org/10.1016/j.seizure.2008.10.010*.

Roy, S., Asif, U., Tang, J., & Harrer, S. (2019). Machine Learning for Seizure Type Classification: Setting the benchmark. *ArXiv*, 1–5. *https://arxiv.org/abs/1902.01012*.

Asif, U., Roy, S., Tang, J., & Harrer, S. (2020). SeizureNet: Multi-Spectral Deep Feature Learning for Seizure Type Classification. *ArXiv*. Retrieved from *https://arxiv.org/abs/1903.03232*.

Scher, M. S. (2017). Pediatric Neurophysiologic Evaluation. In K. Swaiman, S. Ashwal, D. Ferriero, N. Schor, R. Finkel, A. Gropman, … M. Shevell (Eds.), *Swaiman’s Pediatric Neurology (sixth edition)* (6th ed., pp. 1309–1403). China: Elsevier. *https://doi.org/10.1016/B978-0-323-37101-8.00013-8*.

Scheuer, L, M., & Wilson, B, S. (2004). Data Analysis for Continuous EEG Monitoring in the ICU: Seeing the Forest and the Trees. *Journal of Clinical Neurophysiology*, *21*(5), 353–378. *https://doi.org/10.1097/01.WNP.0000147078.82085.D3*.

Scheuer, M. L., Wilson, S. B., Antony, A., Ghearing, G., Urban, A., & Bagic, A. I. (2020). Seizure Detection: Interreader Agreement and Detection Algorithm Assessments Using a Large Dataset. *Journal of Clinical Neurophysiology*. *https://journals.lww.com/clinicalneurophys/Fulltext/9000/Seizure\_Detection\_\_Interreader\_Agreement\_and.99407.aspx*. (in publication).

Shah, V., von Weltin, E., Ahsan, T., Obeid, I., & Picone, J. (2020). On the Use of Non-Experts for Generation of High-Quality Annotations of Seizure Events. *Journal of Clinical Neurophysiology*. *https://www.isip.piconepress.com/publications/unpublished/journals/2019/elsevier\_cn/ira/*.

Shah, V., von Weltin, E., Lopez, S., McHugh, J. R., Veloso, L., Golmohammadi, M., … Picone, J. (2018). The Temple University Hospital Seizure Detection Corpus. *Frontiers in Neuroinformatics*, *12*, 1–6. *https://doi.org/10.3389/fninf.2018.00083*.

Shoeb, A. H., & Guttag, J. V. (2010). Application of machine learning to epileptic seizure detection. *Proceedings of the International Conference on Machine Learning (ICML)*, 975–982. *https://doi.org/10.1016/j.jneumeth.2010.05.020*.

Shoeb, A., Kharbouch, A., Soegaard, J., Schachter, S., & Guttag, J. (2011). An algorithm for detecting seizure termination in scalp EEG. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 1443–1446). *https://doi.org/10.1109/IEMBS.2011.6090357*.

Sivaraju, Adithya, J., & Gilmore, Emily, J. (2016). Understanding and Managing the Ictal-Interictal Continuum in Neurocritical Care. *Current Treatment Options in Neurology*, 18(2), 1–13. *https://doi.org/10.1007/s11940-015-0391-0*.

Stafstrom, C. E., & Carmant, L. (2015). Seizures and epilepsy: an overview for neuroscientists. Cold Spring Harbor Perspectives in Medicine, 5(6): a022426. *https://doi.org/10.1101/cshperspect.a022426*.

Stefan, H., & William Theodore. (2012). Epilepsy, Part II, Treatment, Volume 108. In H. Stefan & T. William (Eds.) (p. 608). *https://www.elsevier.com/books/epilepsy-part-ii-treatment/stefan/978-0-444-52899-5*.

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the Importance of Initialization and Momentum in Deep Learning. In *Proceedings of the International Conference on Machine Learning (ICML)* (pp. 1139–1147). *http://proceedings.mlr.press/v28/sutskever13.html*.

Swisher, C. B., White, C. R., Mace, B. E., & Dombrowski, K. E. (2015). Diagnostic Accuracy of Electrographic Seizure Detection by Neurophysiologists and Non-Neurophysiologists in the Adult ICU Using a Panel of Quantitative EEG Trends. *Journal of Clinical Neurophysiology*, *32*(4), 324–330. *https://doi.org/10.1097/WNP.0000000000000144*.

Szaflarski, J. P., Gloss, D., Binder, J. R., Gaillard, W. D., Golby, A. J., Holland, S. K., … Theodore, W. H. (2017). Practice guideline summary: Use of fMRI in the presurgical evaluation of patients with epilepsy. *Neurology*, *88*(4), 395–402. *https://doi.org/10.1212/WNL.0000000000003532*.

Tatum, W. (2014). *Handbook of EEG Interpretation*. (W. Tatum, A. M. Husain, S. R. Benbadis, & P. W. Kaplan, Eds.) (2nd ed.). New York City, New York, USA: Demos Medical Publishing. Retrieved from *https://books.google.com/books/about/Handbook\_of\_EEG\_Interpretation\_Second\_Ed.html?id=BLsiAwAAQBAJ*.

Uhrig, J., Schneider, N., Schneider, L., Franke, U., Brox, T., & Geiger, A. (2017). Sparsity Invariant CNNs. In *arXiv*. Retrieved from *https://arxiv.org/abs/1708.*06500.

University of Freiburg. (2020). Seizure Prediction Project Freiburg. Retrieved July 1, 2020, from *https://epilepsy.uni-freiburg.de/freiburg-seizure-prediction-project/eeg-database*.

Vesely´, K., Ghoshal, A., Burget, L., & Povey, D. (2013). Sequence-discriminative training of deep neural networks. In *INTERSPEECH* (pp. 2345–2349). Lyon, France. *https://www.danielpovey.com/files/2013\_interspeech\_dnn.pdf*.

von Goethem, P., & Hambling, B. (2013). *User Acceptance Testing: A step-by-step guide*. (P. van Goethem, Ed.), *User Acceptance Testing*. Swindon, United Kingdom: BCS Learning & Development Limited. Retrieved from *https://www.oreilly.com/library/view/user-acceptance-testing/9781780171678/*.

von Weltin, E., Ahsan, T., Shah, V., Jamshed, D., Golmohammadi, M., Obeid, I., & Picone, J. (2017). Electroencephalographic Slowing: A Primary Source of Error in Automatic Seizure Detection. In I. Obeid & J. Picone (Eds.), *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium* (pp. 1–5). Philadelphia, Pennsylvania, USA: IEEE. *https://doi.org/10.1109/SPMB.2017.8257018*.

Wegmann, S., Faria, A., Janin, A., Riedhammer, K., & Morgan, N. (2013). The TAO of ATWV: Probing the mysteries of keyword search performance. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 192–197). Olomouc, Czech Republic: IEEE. *https://doi.org/10.1109/ASRU.2013.6707728*.

Wilson, S. B., Scheuer, M. L., Plummer, C., Young, B., & Pacia, S. (2003). Seizure detection: Correlation of human experts. *Clinical Neurophysiology*, *114(11)*, 2156–2164. *https://doi.org/10.1016/S1388-2457(03)00212-8*.

Xu, M., Duan, L.-Y., Cai, J., Chia, L.-T., Xu, C., & Tian, Q. (2004). HMM-Based Audio Keyword Generation. In *Advances in Multimedia Information Processing – PCM 2004: 5th Pacific Rim Conference on Multimedia* (pp. 566–574). Retrieved from *https://web.archive.org/web/20070510193153/*.

Ye, Y., Xun, G., Ma, F., Suo, Q., Xue, H., Jia, K., & Zhang, A. (2018). A novel channel-aware attention framework for multi-channel EEG seizure detection via multi-view deep learning. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)* (pp. 206–209). *https://doi.org/10.1109/BHI.2018.8333405*.

Yuan, G.-Q., Gao, D.-D., Jun Lin, S. H., & Lv, B.-C. (2013). Treatment of recurrent epileptic seizures in patients with neurological disorders. *Experimental and Therapeutic Medicine*, *5*(1), 267–270. *https://doi.org/10.3892/etm.2012.788*.