MACHINE LEARNING APPROACHES TO AUTOMATIC INTERPRETATION OF EEGS

Iyad Obeid and Joseph Picone

Abstract— An EEG measures electrical activity of the brain from the scalp surface and is the primary clinical tool for diagnosing epilepsy and strokes. Clinical use of EEGs is rapidly increasing as new applications are being developed, including the diagnosis of sleep disorders, head-related trauma injuries and Alzheimer's. Furthermore, with the advent of wireless technology, long-term monitoring occurring over a period of time from several hours to several days has become possible, overwhelming clinicians and rapidly outstripping the resources available to manually interpret such data. The development of a system that can automatically interpret an EEG allows healthcare providers to keep pace with the growing demand for this diagnostic tool and would provide real-time alerting of potentially life-threatening conditions.

Machine learning has made tremendous progress over the past three decades in many fields due to rapid advances in low-cost highly-parallel computational infrastructure, powerful machine learning algorithms, and, most importantly, big data. In this chapter, we describe the steps involved in the development and evaluation of a high-performance machine learning system that automatically identifies key events in an EEG signal at performance levels close to human performance. We describe a unique open source big data resource known as the TUH EEG Corpus that enabled the development of this technology by supporting the application of state of the art statistical modeling. We also describe several related applications including the detection of abnormal EEGs and seizures. The underlying technology common to these systems is based on a combination of hidden Markov models and deep learning.

Keywords— electroencephalography (EEG), machine learning, deep learning, spike detection, abnormal EEG detection, seizure detection.

Manuscript submitted March 6, 2017. I. Obeid and J. Picone are with the Department of Electrical and Computer Engineering at Temple University. The corresponding author is J. Picone, College of Engineering, ENGR 703A, 1947 North 12th Street, Philadelphia, Pennsylvania 19122 USA (phone: 215-204-4841; fax: 215-204-5960; email: joseph.picone@gmail.com).

1. INTRODUCTION

An EEG measures the spontaneous electrical activity of the brain, as shown in Figure 1. A routine EEG typically lasts 20 to 30 minutes, and is used in clinical circumstances where a short measurement is sufficient (e.g. distinguishing epileptic seizures from other types of seizures). Routine EEGs often include standard activation procedures that increase the chance of capturing seizure-like discharges or even seizures (e.g., hyperventilation and photic stimulation). The entire session for a routine EEG, including the time required to affix sensors to a patient's scalp, requires one to two hours. Patients are asked to lie still in a prone position, and are periodically requested to perform limited movements (e.g., breath, blink).

Long-term monitoring (LTM) is useful in a variety of situations in which patients have intermittent disturbances that are difficult to capture during routine EEG sessions. Patients who experience medical conditions such as epilepsy or stroke are often subjected to long-term monitoring in a critical care setting such as an Epilepsy Monitoring Unit (EMU). In such cases recordings can last several hours or several days, generating a large amount of data that needs to be reviewed by a clinician. Advances in digital technology have greatly enhanced the ability to acquire, store and review large amounts of clinical data, which typically now includes electrical signals, video and other vital signs. Clinical practice is struggling to keep pace with the vast amount of patient data being collected. This has created an opportunity for automated computer-based processing and interpretation of EEGs.

The signals measured along the scalp can be correlated with brain activity, which makes it a primary tool for diagnosis of brain-related illnesses (Tatum et al., 2007; Yamada & Meng, 2009). The electrical signals are digitized and presented in a waveform display as shown in Figure 2. EEG specialists review these waveforms and develop a diagnosis. The output of the process is a physician's EEG report, as shown in Figure 3.

It is important to note that a vast majority of all routine EEGs conducted are inconclusive (Smith, 2005): "In healthy adults with no declared history of seizures, the incidence of epileptiform discharge in routine EEG was 0.5%. A slightly higher incidence of 2–4% is found in healthy children and in nonepileptic patients referred to hospital EEG clinics. The incidence increases substantially to 10–30% in

cerebral pathologies such as tumor, prior head injury, cranial surgery, or congenital brain injury." Patients experiencing epilepsy rarely seize during a routine EEG session, even though audio visual stimulation (AVS) such as rhythmic stimulation using light and sound is often applied to induce seizures. In recent years, with the advent of wireless technology, long-term monitoring occurring over a period of days to weeks has become possible. Ambulatory data collections, in which untethered patients are continuously monitored, are becoming increasingly popular due to their ability to capture seizures and other critical infrequently occurring events (Dash et al., 2012). Unfortunately, the data collected under such conditions is often sufficiently noisy and poses serious challenges for automated analysis systems.

EEGs traditionally have been used to diagnose epilepsy and strokes (Tatum et al., 2007) although other common clinical diagnoses include coma (Ardeshna, 2016), encephalopathies (Sutter et al., 2015), brain death (Ercegovac, 2010) and sleep disorders (Rudrashetty et al., 2015). EEGs and other forms of brain imaging such as fMRI are increasingly being used to diagnose conditions such as head-related trauma injuries, Alzheimer's disease, Posterior Reversible Encephalopathy Syndrome (PRES) and Middle Cerebral Artery Infarction (MCA Infarct). Computerized EEG signal processing applications have included predicting dementia in Parkinson's disease patients (Klassen et al., 2011; Al-Qazzaz et al., 2014), stroke volume measurement and outcome prediction (Sheorajpanday et al., 2011; Finnigan & van Putten, 2013), psychosis evaluation in high-risk patients (van Tricht et al., 2014) and assessment of traumatic brain injury severity.

The increasing scope of conditions addressable by an EEG suggest that there is a growing need for expertise to interpret EEGs and, equally importantly, research to understand how various conditions manifest themselves in an EEG signal. Computer-generated EEG interpretation and indentification of critical events can thererefore be expected to significantly increase the quality and efficiency of a neurologist's diagnostic work. Clinical consequences include regularization of reports, real-time feedback and decision-making support to physicians. Computerized EEG assessment can therefore potentially alleviate the bottleneck of inadequate resources to monitor and interpret these tests.

The primary focus of this chapter is to introduce readers to the process of developing machine learning technology for automatic interpretation of EEGs. Readers interested in the fundamentals of the electrophysiology or neuroscience on which an EEG is based are encouraged to read one of many excellent discourses (Ebersole & Pedley, 2014; Tatum et al., 2007) on the topic. Here we focus less on basic EEG science and more on the issues that impact machine learning systems. We also describe the challenges in providing reliable real-time alerts, which is an important capability for long-term monitoring, and a critical gap preventing EEGs from becoming more useful and pervasive. This chapter begins with an overview of the challenges in manual interpretation of an EEG in Section 2. We then introduce the TUH EEG Corpus (TUH-EEG) and discuss the important role big data is playing in the development of automatic interpretation technology in Section 3. Next, in Section 4, we introduce machine learning systems that can detect critical EEG events such as spikes with performance close to human experts. We then conclude with a discussion of emerging directions in high performance classification using deep learning in Section 5. The latter is enabled by the existence of large corpora of labeled data such as that described in Section 3.

2. BIG DATA ISSUES IN MANUAL INTERPRETATION OF EEGS

The information yielded by an EEG channel is essentially the difference of electrical activity between two electrodes. Figure 4 shows a standard electrode mapping for a 10/20 configuration as recommended by the American Clinical Neurophysiology Society (ACNS) (Jurcak, 2007). Because conduction of electricity along the scalp is a nonlinear process, changes in the electrode locations on the scalp often cause significant variations in the signals observed, as does the reference point used to measure scalp voltages. Grounding also plays a critically important role in the quality of the observed signals.

2.1. Waveform Displays Are Still a Primary Visualization Tool

Manual interpretation of an EEG is a subtle process that can require extensive knowledge of the patient's medical history, medication history, and alertness, as well as the duration and morphology of EEG signal events. EEG signals are often analyzed based on their temporal properties, e.g. amplitude, shape and frequency. The latter is typically measured by counting peaks in the time domain. No single

feature or collection of features identify an EEG as normal. It is the overall orderly progression of signal over time that best represents a normal pattern (Ebersole & Pedley, 2014). Essential features of a normal EEG include frequency content, polarity of spikes in the signal, symmetry of transient behavior, and perhaps most importantly, the locality of an event (e.g., whether an event is observed across all channels or only a few channels that correspond to a particular brain region). Frequency content in the alpha, beta, delta, and theta bands, which are not measured directly but can be derived from the channels available in a standard EEG recording, can often be used to detect anomalous behavior. Time-based waveform displays are still the most popular means by which a neurologist interprets an EEG, although frequency domain visualizations have become more popular for rapidly scanning large amounts of data to locate regions of interest (Swisher et al., 2016; Thiess et al., 2016).

Neurologists can often determine whether an EEG is normal or abnormal with high reliability by examining the first few minutes of an EEG session. Not surprisingly, machine learning systems can approach human performance on this task operating on similar amounts of data (Lopez, 2017; Lopez, et al., 2015). Neurologists can also identify what are often referred to as benign variants with high reliability – behaviors that on the surface might be considered anomalous, but that are not indicative of a medical condition or are simply inconclusive. For example, patient movement or eye blinks can often cause spikes in the signal that can be easily misclassified by machine learning systems. These benign variants contribute to the high false alarm rate from which most commercial systems suffer (Scheuer et al., 2017). Therefore, interactive tools used by neurologists typically include many digital signal processing options that accentuate certain behaviors in the signal (van Beelen, 2013; Thiess et al., 2016), including low pass, high pass, band pass and notch filters.

2.2. Signal Conditioning Enhances Interpretation

Because typical electrical voltage ranges for EEG signals are in the tens of microvolts and extremely noisy, EEG signals are typically visualized using a differential view, known as a montage, that consists of signal differences from various pairs of electrodes (e.g., Fp1-F7). ACNS recognizes that there are a great variety of montage styles in use among EEG practitioners, and has proposed guidelines for a minimum set

of montages (Jurcak, 2007). Neurologists are often particular about the specific montage used when interpreting an EEG. At Temple University Hospital (TUH), for example, the Temporal Central Parasagittal (TCP) montage (Acharya et al., 2016) is often used, as it accentuates spike behavior. Despite ACNS guidelines, several voltage reference sites are still used during EEG recordings depending on the purpose of the EEG recording (Harati et al., 2014).

Some commonly used reference schemes, which are depicted in Figure 4, include:

- Common Vertex Reference (Cz): uses an electrode in the middle of the head.
- Linked Ears Reference (A1+A2, LE, RE): based on the assumption that sites like the ears and mastoid bone lack electrical activity, often implemented using only one ear.
- The Average Reference (AR): uses the average of a finite number of electrodes as a reference.

The robustness of a state of the art machine learning system that decodes EEG signals depends highly on the ability of the system to maintain its performance when there are variations in the recording conditions. The specific montage of a recording could potentially affect the operation of such systems in a negative way, which constitutes a fundamental problem, given the fact that EEG signals tend to present high variability in clinical settings. In Lopez et al. (2016) it was observed that there are some systematic biases in performance that depend on the source of the data, but these are relatively small compared to the overall problem of detecting seizures, spikes or other such transient phenomena.

2.3. Locality Is an Extremely Important Feature

The spatial locality of an event often plays a major role in its classification. Since each electrode is tied to a particular location on the scalp, the channels in which an event occurs prominently becomes an important key for classification. Frontal lobes, which are defined as the area at the front of the brain behind the forehead, are responsible for voluntary movement, conscious thought, learning and speech (Nolte & Sundsten, 2015). Temporal lobes, which are defined as the areas of the brain at the side of the head above the ears are responsible for memory and emotions. Parietal lobes are the area of the brain at the top of your head behind your frontal lobes and control cognitive functions such as how we process

sensory input, how we judge spatial relationships and coordination. Our ability to read, write and do math is also tied to this region of the brain. The occipital lobes are the area at the back of the brain at the back of your head and are responsible for our sense of sight.

Conditions such as epilepsy are caused by disruptions in normal brain activity. There are several key types of disruptions that can occur. These can be broadly clustered into two classes (Misulis et al., 2014):

- **Partial (Focal) Seizures:** seizures that happen in, and affect, only part of the brain. The signatures of these types of seizures depends on which part of the brain is affected.
- **Generalized Seizures:** seizures that happen in, and affect, both sides of the brain. The patient is often unconscious during this type of seizure and won't remember the seizure itself. The most well-known category for this type of seizure is a tonic clonic (convulsive) seizure.

A seizure causes a change in the EEG, so detecting changes from normal patterns becomes an important first step in the process. Observation of an abnormal EEG does not prove that the patient has epilepsy. EEGs must be used alongside other tests to conclusively diagnose a condition. For example, video of the patient is often examined along with an EEG and MRIs are increasingly being used to confirm diagnoses. The combination of the montage, used to accentuate spike or transient behavior, and the nature of the locality are important cues for manual interpretation of an EEG. Clinicians also visually adapt to the background channel behavior of a patient’s data before they can identify cues such as spikes that lead to the identification of a seizure.

2.4. Annotations Play a Critical Role in Machine Learning Systems

There are generally two approaches to developing machine learning technology to automatically interpret EEGs. The first approach, which relies on expert knowledge, requires a deeper understanding of how EEGs are manually interpreted and the translation of this process into an algorithm description. Low-level events, such as spikes, are detected, and then a higher-level of logic is applied to map sequences of these events to diseases or outcomes. This is analogous to the process of recognizing phonemes in speech recognition and then building word-level transcriptions from these phoneme hypotheses (Deshmukh et al.,

1999). This requires data annotated in such a way that low-level event models can be trained, which, in turn, requires some agreement on a set of low-level labels.

After several iterations with a group of expert neurologists, and following popular approaches found in the literature (Baldassano et al., 2016), we have developed the following 6-way classification for a segment of an EEG, which we refer to as an epoch:

(1) *Spike and/or Sharp Wave (SPSW)*: epileptiform transients that are typically observed in patients with epilepsy.

(2) *Periodic Lateralized Epileptiform Discharges (PLED)*: EEG abnormalities consisting of repetitive spike or sharp wave discharges, which are focal or lateralized over one hemisphere and that recur at almost fixed time intervals.

(3) *Generalized Periodic Epileptiform Discharges (GPED)*: periodic short-interval diffuse discharges, periodic long-interval diffuse discharges and suppression-burst patterns according to the interval between the discharges. Triphasic waves (diffuse and bilaterally synchronous spikes with bifrontal predominance, typically periodic at a rate of 1-2 Hz) are included in this class.

(4) *Artifacts (ARTF)*: recorded electrical activity that is not of cerebral origin, such as those due to equipment or environment.

(5) *Eye Movement (EYEM)*: common events that can often be confused for a spike.

(6) *Background (BCKG): all other signals.*

These classes are very similar to what others have used (Waterstraat et al., 2015; Wulsin et al., 2010) to perform stroke and epilepsy detection. Epochs are usually one second in duration and are further subdivided in time for signal conditioning and analysis.

Examples of the SPSW and EYEM classes are shown in Figure 2. We typically annotate data in a channel-dependent manner since we need to establish the locality of an event. We refer to such annotations as event-based annotations since they identify the start and stop times of evens on specific channels. A summary judgement for each epoch is then made based on the channel-dependent annotations. We refer to these annotations as term-based, following a convention used in other research

communities (Doddington et al., 2000). We generate these automatically using a majority voting scheme that looks across all channels. In cases where the outcome of a majority vote is not clear, we resolve ambiguity manually.

Identification of these six events is important towards making a final classification of a section of data as constituting a seizure event. The first three classes are information bearing in that they describe events that are critical in manual interpretation of an EEG. What primarily distinguishes these three classes is the degree of periodicity and the extent to which these events occur across channels. Neurologists can identify PLEDs and GPEDs with a high degree of accuracy since these events are distinctive because of their long-term repetitive behavior. Accuracy for manually detecting spikes, however, is more problematic (Scheuer et al., 2017).

The last three classes are used to improve our ability to model the background channel. Background modeling is an important part of any machine learning system that attempts to model the temporal evolution of a signal (e.g., hidden Markov models, deep learning). We let the system automatically perform background/non-background classification as part of the modeling process rather than use a heuristic preprocessing algorithm to detect signals of interest. This is described in more detail in Section 4. This follows a very successful approach that we have used in speech recognition (Picone, 1990).

Artifacts, eye blinks and eye-related muscle movements occur frequently enough that they merit separate classes. These events appear as transient events in the signal and to an untrained eye can be misinterpreted as spike behavior. The rest of the events that don't match the first five classes are lumped into the background class. Hence, it is important that the background class model be robust and powerful, because this is the primary way the false alarm rate is minimized. Further, the critical aspects of performance are related to the sensitivity and specificity of the first three classes since these are the events neurologists will key on to interpret a session. Hence, as discussed in Section 2.6, we often adjust our scoring criteria to give proper weight to these classes.

The second machine learning approach, which is embodied in many deep learning-based systems that are so popular today, is to let the system learn the underlying structure of the data. In this approach, we provide manual annotations of seizure events that simply indicate the start and stop time of a seizure, and optionally the type of seizure. For this approach to work, a large archive of data is needed, and hence the focus on big data resources described in Section 3. A waveform display for a typical seizure event, along with the corresponding spectrogram for a selected number of channels, is shown in Figure 5. The TUH EEG Seizure Corpus (Golmohammadi et al., 2017), which is a subset of TUH-EEG, provides a large annotated corpus of seizure events that can be used to develop such technology.

Our seizure event annotations include: start and stop times; localization of a seizure (e.g., focal, generalized) with the appropriate channels marked; type of seizure (e.g., simple partial, complex partial, tonic-clonic, gelastic, absence, atonic); and the nature of the seizure (e.g., convulsive). The non-seizure event annotations include: artifacts which could be confused with seizure-like events such as ventilatory artifacts and lead artifacts; non-epileptiform activity that may resemble epileptiform discharges, such as psychomotor variant, mu, breach rhythms and positive occipital sharp transients of sleep (POSTS); abnormal background which could be confused with seizure-like events (e.g. triphasics); and interictal and postictal states. The types of features are important when manually interpreting an EEG and determining how a seizure manifests itself. We use these finer categorizations of seizures to build models specific to these events, which also helps reduce the false alarm rate.

2.5. Inter-Rater Agreement Is Low

A board-certified EEG specialist currently is required by law to interpret an EEG and produce a diagnosis. It takes several years of additional training post-medical school for a physician to qualify as a clinical specialist. However, despite completing a rigorous training process, there is only moderate inter-rate agreement (IRA) in EEG interpretation (Stroink et al., 2006; van Donselaar et al., 1992) for low-level events such as spikes. In Halford et al. (2015), it was noted that IRA among experts was significantly higher for identification of electrographic seizures compared to periodic discharges. In Swisher et al. (2015), it was noted that even augmenting traditional waveform displays with a display based on

advanced “brain mapping” analytics, known as a quantitative EEG (qEEG), was not adequate as the sole method for reviewing continuous EEG data. Kappa statistics (Nizam et al., 2013) in the range of 0.6 are common for manual interpretations and drop considerably as the data becomes more challenging or is collected under typical clinical conditions.

What makes this problem so challenging is that an EEG signal is a very low voltage signal (e.g., microvolts). The slightest disturbances, such as simply pressing on the electrical connections, cause large deflections in the waveforms. There are many anomalies that produce spike-like behavior in the signal. An example is shown in Figure 2, where we see an SPSW event on the left side of the image and an EYEM event on the right side of the image. Video is often used concurrently with an EEG to characterize paroxysmal clinical events that might be seizures, including grimacing, chewing, or nystagmoid eye movements; abrupt and otherwise unexplained changes in pulse, blood pressure or respiratory pattern; or abrupt deterioration in conscious level. Ideally, video and the EEG signals should be recorded concurrently. Accurate recognition and distinction of benign variants in an EEG are essential to avoid over interpretation. The range of benign variants include highly confusable events such as small sharp spikes (often referred to as BSSS or BSST for benign small sleep transients), 14 and 6 positive spikes, 6-Hz “phantom” spike and wave, and subclinical rhythmic EEG discharge (Britton et al., 2017). These often require additional input beyond EEG waveforms (e.g., video). Therefore, it is not surprising that IRA is fairly low even amongst experts, particularly on clinical data in which patient behavior is not well controlled.

2.6. Evaluation Metrics Are Important

Annotations play a key role in most machine learning applications where supervised training (Picone, 1990) is used. Accurate system evaluation, however, also represents a challenge in itself. Researchers typically report performance in terms of sensitivity and specificity (Japkowicz & Shah, 2011) of epochs in biomedical research applications (Altman & Bland, 1994). Each epoch is considered as a separate testing example even though EEG events can span multiple epochs. The results of the classifier are presented in a

confusion matrix, which gives a very useful overview of performance. For example, for a two-class problem such as seizure detection, a confusion matrix has following categories:

- **true positives (TP):** the number of epochs identified as a seizure in the reference annotations, and were correctly labeled as a seizure;
- **true negatives (TN):** the number of epochs correctly identified as non-seizures;
- **false positives (FP):** the number of epochs incorrectly labeled as seizure;
- **false negatives (FN):** the number of epochs incorrectly labeled as non-seizure.

Sensitivity (TP/TP+FN) and specificity (TN/TN+FP) are derived from these quantities. A precision–recall (PR) curve is an alternative method of scoring (Manning et al., 1999) in which precision is percentage of correctly detected seizure divided by predicted seizure epochs (TP/TP+FP), while recall is called sensitivity.

However, sensitivity can often be increased arbitrarily if one is willing to tolerate a poor specificity or a high false alarm rate. Interviews conducted with many clinicians have indicated that the primary reason commercially available technology is not used in clinical settings is due to the high false alarm rate (Hu, 2015; Obeid & Picone, 2015). This is perhaps the single most important metric today in guiding machine learning research applications in critical care. Critical care units are overwhelmed with the number of false positives that automated event detection equipment generates. To put this in perspective, one false alarm per bed per hour in a 12-bed ICU generates 12 interrupts per hour that must be serviced. This can easily overwhelm healthcare providers. Since many types of automated monitoring equipment are used in an ICU setting, each with significant false alarm issues, the number of false alarms that must be serviced by healthcare providers is overwhelming (Christensen et al., 2014). As a result, clinicians report that in practice they simply ignore these systems.

Of course, one must balance sensitivity, specificity and false alarms. This has been studied extensively in other communities focused on event-spotting technology such as spoken term detection in voice signals (Madal et al., 2014). A measure that we have borrowed from this research community is the

term-weighted value (TWV) (Doddington et al., 2000), which is based on the notion of a Detection Error Tradeoff (DET) curve (Martin et al., 1997). A DET curve is very similar to a Receiver Operating Characteristic originally developed to assess the performance of a communications system (Jacobs & Wozencraft, 1965). TWV essentially assigns an application-dependent reward to each correct detection and a penalty to each incorrect detection. TWV is one minus the weighted sum of the term-weighted probability of missed detection and the term-weighted probability of false alarms. The actual TWV (ATWV) is performance measured for a specific decision threshold – essentially establishing a specific operating point on the DET curve. This measure is useful when it is preferred to compare two systems based on a single number, though it is always better to compare DET curves over a range of operating characteristics. ATWV and DET curves are our recommended way to evaluate EEG interpretation systems.

To use ATWV, however, you need what are referred to as term-based annotations of the data, and you need to tune the weights assigned to various error modalities (Ziyabari et al., 2017; Doddington et al., 2000). Epoch-based annotations are defined as those in which each frame is labeled. Researchers often choose to evaluate their systems on a subset of the epochs available in a database because the overwhelming majority of epochs are assigned to a background, non-seizure, or the equivalent. Since the data is dominated by the presence of events assigned to the background class, performance and evaluation will be biased towards background events if proper normalizations are not considered. In a typical clinical corpus, seizure events account for less than 0.01% of the data (as measured by the cumulative number of seconds seizure events exist). Hence, optimization of a system based on such a metric will focus on non-seizure events, which is not desirable in practice.

Therefore, we annotate the data using term-based annotations, which simply denote the start and stop time of specific events such as a seizure. ATWV allows one to tune the tradeoffs between various types of error classes. It is ideal for these types of applications because it adequately weights false alarms, which is crucial to this application. ATWV ranges from $[-\infty,1]$, with a score greater than 0.5 being indicative of a system that is performing well. Negative ATWV scores are typically indicative of systems with high false

alarm rates. ATWV software is available from the National Institute of Standards and Technology (NIST, 2010) as part of the Open Keyword Search Evaluation package.

2.7. Decision Support Systems Can Enhance Interpretation

Decision support systems in healthcare, which can greatly improve manual interpretation, can leverage vast archives of electronic medical records (EMRs) if high performance automated data wrangling can be achieved (Picone & Obeid, 2016; Harabagiu, 2015; Picone, et al., 2015). EMRs can include unstructured text, temporally constrained measurements (e.g., vital signs), multichannel signal data (e.g., EEGs), and image data (e.g., MRIs). Clinicians who specialize in visual interpretation of data often require second opinions, consult data banks of reference samples, or even reference textbooks for difficult cases that are outside of their normal daily experiences. Medical students specializing in neurology spend years shadowing clinicians while they learn how to read EEGs through experiential training.

One application of automatic interpretation technology that integrates high performance classification with big data is cohort retrieval. This application can positively impact clinical work and medical student training. When observing an event of interest, such as a seizure, it is desirable to locate other similar examples of such signals, either from previous sessions from the same or similar patients. Information from EEG reports and EEG signals can be mined in such a way that database queries to locate such events can be performed on the aggregated data. Clinical consequences include regularization of EEG reports, real-time feedback and decision-making support, and enhanced training for young neurophysiologists.

One of the challenges in this task is that the EEG reports, such as the one shown in Figure 3, are captured as unstructured text in most clinical environments. Therefore, natural language processing is required to identify key medical concepts in the reports. Identification of the type and temporal location of EEG signal events such as spikes or generalized periodic epileptiform discharges in the EEG signal are critical to the interpretation of an EEG. Cohort retrieval systems allow users to query such information using natural language (e.g., “Show me all similar young patients with focal cerebral dysfunction who

were treated with Topamax"). In Figure 6, we show an example of one such system (Picone & Obeid, 2016) based on TUH-EEG (described in Section 3).

2.8. The Unbalanced Data Problem Makes Machine Learning Challenging

Finally, as mentioned previously in Section 2.6, something that makes this problem additionally challenging is the large imbalance between events of interest (e.g., seizures, spikes) and non-events (e.g., background). This is often referred to as the imbalanced data problem (He et al., 2013). The amount of time that a patient experiences a seizure is typically less than 0.01% of the overall data in a clinical setting. Clinicians must sift through vast amounts of data in real-time to diagnose a disorder. For example, a patient is often admitted to an EMU or ICU for continuous monitoring. Seizure events will occur optimistically only a few times per day. Nevertheless, all data must be manually reviewed. Further, patients now can also use ambulatory EEG systems that allow continuous data collection from their normal living environments, further amplifying the amount of data that must be reviewed.

Direct training on this type of data poses several challenges for traditional machine learning approaches. Even the best deep learning systems will ignore such infrequently occurring events in their efforts to optimize a performance metric. In such situations where one class is significantly more probable than another, the obvious solution is to always guess the most probable class. Though overall performance will appear to be good, performance on events of interest is very poor. This is a common problem in this application space, and another reason a weighted metric like ATWV becomes important. As we will see in the next section, big data plays a critical role in this field, because we need large amounts of these infrequently occurring events to train high performance statistical models. Techniques such as cross-validation and boosting (Duda et al., 2001) play an important role in avoiding such machine learning problems.

3. THE TUH EEG CORPUS

The development of TUH-EEG (Obeid & Picone, 2016) was an attempt to build the world's largest publicly available database of clinical EEG data. It currently comprises more than 30,000 EEG records from over 16,000 patients and is growing at a rate of about 3,000 sessions per year. It represents the

collective output from Temple University Hospital's Department of Neurology since 2002. Data collection began in 2013 and is an ongoing effort. We have currently released all data from 2002-2013 in v0.0.6, and will continue to release additional data on an annual basis (v1.0.0 was released in Spring 2017 and contains data through 2015). The data is available from the Neural Engineering Data Consortium (*http://www.nedcdata.org/*). Future releases are expected to include data from other hospitals and metadata from a number of collaborative projects. All work was performed in accordance with the Declaration of Helsinki and with the full approval of the Temple University Institutional Review Board (IRB). All personnel in contact with privileged patient information were fully trained on patient privacy and were certified by the Temple IRB.

Because of the long time horizon of the data collection, the original data exists in many data formats that reflect the evolution of clinical practice and instrumentation. Archival EEG signal data were recovered from CD-ROMs. Files were converted from their native proprietary file format (Nicolet's NicVue) to an open format EDF standard. Data was then rigorously de-identified to conform to the HIPAA Privacy Rule by eliminating 18 potential identifiers including patient names and dates of birth. Patient medical record numbers were replaced with randomized database identifiers, with a key to that mapping being saved to a secure off-line location. An important part of our process was to identify similar patients even though they appeared in the original data with different medical record numbers, name spellings or in some cases name changes. Data de-identification was performed by combining automated custom-designed software tools with manual editing and proofreading. All storage and manipulation of source files was conducted on dedicated non-network connected computers that were physically located within the TUH Department of Neurology.

There are two distinguishing aspects of this data. First and foremost, it cannot be overemphasized that the data was collected in a live clinical setting. TUH is the public hospital for Philadelphia and serves a diverse population. The EEGs were collected from adults ranging in age from 16 to 90+ years old. The data was not collected under carefully controlled research conditions. This becomes apparent when we discuss the challenges that EEG signal events such as patient movements pose in terms of robust pattern

recognition. The second important aspect of this corpus is that it is openly available. Users are not required to be added to an IRB or sign data-sharing agreements. Further, users are not restricted in their use of the data, though they should acknowledge the source of the data in all publications. Both research and commercialization work can be conducted with the data.

3.1. Digitization and Signal Processing

EEG signals in TUH-EEG were recorded using several generations of Natus Medical Incorporated's Nicolet^{TM} EEG recording technology. The raw signals obtained from the studies consist of recordings that vary between 20 and 128 channels sampled at a minimum of 250 Hz using a 16-bit A/D converter. The data is stored in a proprietary format that has been exported to an open standard, the European Data Format (EDF) (Kemp, 2013), with the use of NicVue v5.71.4.2530. These EDF files contain an ASCII header with important metadata information distributed in 24 unique fields that contain the patient's information and the signal's condition. There are additional fields that describe signal conditions, such as the maximum amplitude of the signals, which are stored for every channel. A complete description of this header can be found at the TUH-EEG project website (Bergey & Picone, 2017). The signal data is stored in an uncompressed format using 16 bits per sample. The data is normalized by a minimum and maximum value to maximize precision and minimize quantization effects. These normalization values are stored in the EDF header for each file.

The large variability among EEG channels and montages utilized in clinical EEGs is not usually something that is represented in data collected under controlled research conditions. However, this is an important practical issue present in clinical data. For example, in TUH-EEG, there are over 40 different channel configurations and at least four different types of reference points used in the EEGs administered. One example underscoring the importance of data diversity is Lopez et al. (2016), which studied the impact of sensor configuration on classification performance. It was determined there was a statistically significant degradation in performance when reference channel conditions were mismatched (e.g., training on Average Reference data and evaluating on Linked Ears data). Attempts to mitigate this using standard approaches such as Cepstral Mean Subtraction (CMS) (Huang et al., 2001) were not successful,

indicating that further study of this problem is necessary. It is unclear that whether this data can be modeled by a single statistical model, or whether special measures must be taken to account for this variability. Research fields such as speech recognition have dealt with this problem for many years using technologies such as speaker and channel adaptation (Huang et al., 2001), but these technologies have yet to be explored in EEG research.

Although there are many unique sensor configurations, about 50% of the data follows the 10/20 convention shown in Figure 4 closely. The remaining 50% of the data can be mapped onto this configuration using a combination of channel selection and spatial interpolation. We have developed signal processing software to abstract these details from the user so that the data can be easily processed using typical machine learning paradigms. In our preliminary experiments described in Section 4, we will focus on data adequately modeled using a 10/20 configuration since this is the most prevalent configuration for an EEG. Neurologists can manually interpret EEGs collected using this configuration with relatively high accuracy, and there is no compelling evidence that higher resolution EEGs improve human performance. However, higher resolution EEGs are still useful for techniques such as localization of seizures and brain mapping (Michel & Murray, 2012).

The EEG data archived by a hospital is, unfortunately, not the entire signal of record. Because these are multichannel signals, the amount of data storage required for a single EEG would exceed the capacity of a DVD, which are still the primary way this data is archived. For example, a 22-channel signal digitized at 16 bits/sample for 20 minutes totals about 132 Mbytes. A long-term monitoring EEG, which can last for 72 hours or more, can grow in size to several Gbytes of data. At the time these digital systems were designed, this was deemed excessive. Therefore, during a session, technicians mark sections of interest in the signal. When the data is archived to disk, only these sections are retained. This is a process referred to as EEG pruning (LaRoche, 2013). The EEGs in TUH-EEG are all pruned. When these EEGs are exported to an EDF file, they are split into multiple files corresponding to each segment of interest. Start and stop times of these segments relative to the original signal are retained in the EDF file so that some parts of the original timeline can be reconstructed.

3.2. EEG Report Pairing

For every EEG, there is also a report, such as the one shown in Figure 3, which was generated by a board-certified neurologist. This report contains a summary of the physician's findings (e.g., clinical correlation sections) as well as information such as the patient's history and medications. These reports are generated by the neurologist after analyzing the EEG scan and are the official hospital summary of the clinical impression. They are generated anywhere from a few hours to a few days after the patient has been treated depending on the particular workflow for the neurologist and the hospital. These reports are comprised of unstructured text that describes the patient, relevant history, medications, and clinical impression. The report also includes information about the location of the session (e.g., inpatient or outpatient), the type of EEG test (e.g., long-term monitoring or standard) and the protocol invoked for the test (e.g., the type of stimulation used).

These reports typically have four sections: Clinical History, Medications, Introduction, Description of the Record and Clinical Correlations. The last section is perhaps the most important for machine learning research since it contains a summary of the findings. Not surprisingly, the language used in these reports is intentionally measured, which often makes it difficult to automatically interpret using natural language processing techniques. Due to the limitations of information technology (IT) systems at many hospitals, reports are not often easily matched with the EEG signals. Reports are often stored in a separate IT system, and their connection to the actual EEG data is lost over time. Therefore, we manually paired each retrieved EEG with its corresponding clinician report. Reports were mined from the hospital's central electronic medical records archives and typically consisted of image scans of printed reports. Various levels of image processing were employed to improve the image quality before applying optical character recognition (OCR) to convert the images into text. A combination of software and manual editing was used to scrub protected health information (PHI) from the reports and to correct errors in OCR transcription. Only sessions with both an EEG and a corresponding clinician report were included in the final corpus. The unpaired data is still available, and is useful for studies where the lack of an EEG report

or a summary of the findings is not an issue (e.g., clustering, unsupervised training, or self-organizing data analysis).

The pairing process can often be challenging due to a number of complicating factors. A patient can often have multiple medical record numbers (MRNs) either due to practical issues such as clerical errors (e.g., misspelled names or a typo in an MRN), a change in medical insurance, the use of another patient's medical insurance (not uncommon in public hospitals) and changes in marital status which trigger a name change. The timestamp is also a valuable key in pairing reports, but even that can be misleading. A patient can receive multiple EEGs in the same day in some cases, or often receives a standard 20-minute baseline EEG before beginning a long-term monitoring (LTM) session. We were able to resolve these issues using a combination of automated software that detects potential conflicts and manual review. In extreme cases the EEG report must be consulted to make sure the patient's medical history matches demographic data collected by the technician at the time the EEG is administered. The latter is logged as part of the data collection system and is available in our private, unreleased version of the database which we refer to when we need to disambiguate data. We have been able to manually pair over 95% of the data with reports using a combination of MRNs, timestamps and patient demographic information.

3.3. Deidentification of the Data

The EEG reports in TUH-EEG have been manually deidentified so that a patient's identity remains anonymous. This is extremely important because HIPAA compliance (Brezinski, 2016) requires a patient's identity be kept anonymous. Patient information appears in several places in the original data: the EDF header, annotation channels where technicians make comments during the session, EEG reports and some auxiliary files generated by the NicVue software to document the recording session (e.g., the Impedance Report). Deidentification, also referred to as anonymization, requires removal of this information. We decided to release only two types of data – an EEG report as a plain text file and the EEG signals as a collection of pruned EDF files.

The EDF files were redacted in order to ensure the patients' anonymity. This process included modifying the medical record numbers, names, exact dates of birth and study numbers in the ASCII EDF

header. Patients were assigned a randomized ID which can be mapped back to the original MRN through a mapping file stored in a secure, off-line location. Technician comments, which are time-aligned with the data and show up in a binary format as additional channels of the signal, were stripped since these tended to be unproductive for our research needs. These can at times contain patient names or other sensitive information. The manual effort required to redact these was not considered cost-effective relative to our machine learning research goals.

Redaction of personal information in the EEG reports was a much more sensitive issue and required intensive amounts of manual review. Information relevant to the outcome and interpretation of the EEGs, such as gender, age, medical history and medications, was retained. Selected fields from this header that contain important metadata are shown below in Table 1.

Though the EEG reports appear to have a semantic structure, the reports are created typically in Microsoft Word as flat unstructured files. Extracting useful information from these reports can only be done using natural language processing techniques. More research is needed on the organization and representation of these reports, but we are making progress in parsing and representing this data in a related research project (Harabagiu et al., 2016; Obeid et al., 2016) and plan to release this metadata soon (see *www.nedcdata.org* for further details on its release). The Word versions of the EEG reports have been manually redacted, and then automatically converted to flat text files. This data was reviewed multiple times by several data entry specialists and run through a series of text filters designed to spot thousands of special cases indicating incorrect redaction.

3.4. Basic Descriptive Statistics

The first release of TUH-EEG, referred to a v0.0.6, was made in 2015. Approximately 75% of the sessions are standard EEGs less than one hour in duration, while the remaining 25% are from LTM sessions. A distribution of the number of records per year is presented in the upper right of Figure 7. To put the size of this corpus in perspective, the total EEG signal data collected thus far, including all unreleased data, requires over 2 TBytes of storage with a median file size of 20 Mbytes. Though the EEG signal data is pruned, the amount of data is staggering. For example, if we treat each channel of data as an

independent signal, there is over 1 Billion seconds of data. Though this might seem huge at first, the events we are interested in are relatively rare, often occupying less than 0.01% of the recording duration. The number of patients experiencing seizures during a session is on the order of several hundred. When these sessions are cross-referenced by patient medical histories, even this huge amount of data appears small.

The completed v0.0.6 corpus comprises 16,986 sessions from 10,874 unique subjects. Each of these sessions contains at least one EDF file (more in the case of LTM sessions that were broken into multiple files) and one physician report. Corpus metrics are summarized in Figure 7. Subjects were 51% female and ranged in age from less than one year to over 90 (average 51.6, stdev 55.9; see Figure 7 bottom left). The average number of sessions per patient was 1.56, although as many as 37 EEGs were recorded for a single patient over an eight-month period (Figure 7 top left). The number of sessions per year varies from approximately 1,000-2,500 (except for years 2000-2002, and 2005, in which limited numbers of complete reports were found in the various electronic medical record archives; see Figure 7 top right).

There was a substantial degree of variability with respect to the number of channels included in the corpus (see Figure 7 bottom right). The corpus is relatively evenly split between AR reference data (51.5%) and LE referenced data (48.5%). (There are a very small number of reference schemes that don't conform to these two standards). EDF files typically contained both EEG-specific channels as well as supplementary channels such as detected bursts, EKG, EMG, and photic stimuli. The most common number of EEG-only channels per EDF file was 31, although there were cases with as few as 20. A majority of the EEG data was sampled at 250Hz (87%) with the remaining data being sampled at 256Hz (8.3%), 400Hz (3.8%), and 512Hz (1%).

An overview of the distribution of signal events described in Section 2.4 is shown in Table 2 for a subset of the data we have used to develop baseline technology. We see that all classes except SPSW occur frequently enough that robust statistical models can be built. A subset of the SPSW events will ultimately correspond to seizures, so we see how infrequently seizure events actually occur. We estimate that about 5% of the sessions contain actual seizure events, and less than 0.01% of the recorded data

contains an actual seizure event. PLEDs and GPEDs can be located with relative ease. ARTF and EYEM events also can be relatively easily identified using a variety of heuristic or statistical methods.

An initial analysis of the physician reports reveals a wide range of medications and medical conditions. Unsurprisingly, the most common listed medications were anti-convulsants such as Keppra and Dilantin, as well as blood thinners such as Lovenox and heparin. Approximately 87% of the reports included the text string ‘epilep’, and about 12% included ‘stroke’. Only 48 total reports included the string ‘concus’. The EEG reports contain a total of 3.5 M words, which makes it an interesting corpus for natural language processing research.

3.5. The Structure of the Released Data

The TUH EEG Corpus v0.0.6 has been released and is freely available online at *www.nedcdata.org*. Users must register with a valid email address. The uncompressed EDF files and reports together comprise 572 GB. For convenience, the website stores all data from each patient as individual gzip files with a median file size of 4.1 MB; all 10,874 gzips together comprise 330 GB. Users wanting to access the entire database are encouraged to physically mail a USB hard drive to the authors in order to avoid the downloading process.

The corpus was defined with a hierarchical Unix-style file tree structure. The top folder, edf, contains 109 numbered folders, each of which contain numbered folders for up to 100 patients. Each of these patient folders contains sub-folders that correspond to individual recording sessions. Those folder names reflect the session number and date of recording. Finally, each session folder includes one or more EEG (.edf) data files as well as the clinician report in .txt format. Figure 8 summarizes the corpus file structure and gives examples of text and signal data.

We also have released subsets of the data of interest to the community. TUH EEG Epilepsy Corpus (v0.0.1) is a subset of the TUH EEG Corpus that contains 100 subjects with and without epilepsy. The TUH EEG Seizure Corpus (Golmohammadi et al., 2017) is a subset designed to support seizure detection experiments. This corpus contains a 50-patient evaluation set and a 250-patient training set. The evaluation set has been manually annotated for seizure events by a panel of expert neurologists. The

training set has been annotated by a team of experienced annotators and validated using a variety of commercial systems. The TUH EEG Abnormal EEG Corpus is a subset in which each EEG has been manually classified as normal or abnormal by our expert annotation team (Lopez et al., 2015). In addition to these subsets, we have released automatically generated alignments of the six signal event classes described in Section 2.4 for all the data in TUH-EEG. Over the next few years we also expect to expand the corpus to include data from other hospitals.

4. AUTOMATIC INTERPRETATION OF EEGS

Automatic interpretation of EEGs essentially involves detecting epileptiforms and then classifying their nature. For example, if they are persistent, an event will be classified as a GPED. If it is an isolated even, it is classified as SPSW. Disambiguating spikes from background noise is a significant problem, especially for signals such as EEGs where many types of artifacts produce spike like behavior. Most spike detection systems use one of two approaches: (1) heuristic waveform detection techniques or (2) static frame classification techniques. The performance of commercial tools, which are often based on heuristics, has proven to be unacceptable for clinical use (Scheuer et al., 2017; Swisher et al., 2015).

4.1. Machine Learning Approaches

There have been historically three general machine-learning approaches to the problem of EEG classification: (1) a single classification of an entire EEG, (2) segment-level classification and (3) sequential decoding. Single classification of an entire signal represents the simplest approach in which an entire EEG segment is classified by learning a mapping of a single aggregate feature vector (Chandran, 2012; Wang et al., 2011). Lopez et al. (2015) demonstrated classification of abnormal EEGs using this approach. Nonlinear statistical models such as neural networks (NNs), support vector machines (SVMs), and even visual pattern recognition techniques are typically used. Such systems often make a binary decision (e.g., normal/abnormal), but do not identify critical events within the signal.

A second variation on this approach is to segment each channel of the multi-channel signal into frames, analyze and classify each frame, and then perform some overall classification of a segment or an entire file based on these event probabilities (Wulsin at al., 2011; Bao et al., 2009). These two-level

hierarchical systems can be viewed as static classifiers since each frame of data is judged independently of the others. These approaches often use feature extraction techniques based on wavelets (Lin et al., 2011; Rosso et al., 2006) or other similar time/frequency representations of the signal, and often concatenate feature vectors in time to increase temporal context. They do not make use of the temporal sequence of EEG events.

Both these approaches can be viewed as a bottom-up approach because the statistical models that transform segments of the signal to probabilities are not generally informed by preceding or following context. When the low-level signal is highly ambiguous, as is the case in speech recognition or EEG event detection, these techniques fail because the event probabilities generated are not adequately distinct for high performance inference. This was a lesson learned many years ago in speech recognition when expert systems (Lowerre, 1980) and event spotting (Erman et al., 1980) were abandoned in favor of statistical models based on Bayes' rule and exhaustive search (Lee & Hon, 1989).

The third approach, segment decoding, has been popular in applications such as speech recognition, where language provides an overall structure to the signal. We can exploit sequential relationships between words and phonemes to improve the accuracy of this low-level event detection (known as acoustic decoding in speech recognition) (Rabiner, 1989; Picone, 1990). Feature extraction and the transformations used to postprocess features, which we refer to as signal models, learn context-dependent statistics (Picone, 1993) and use these to better calibrate the frequency domain and time domain predictions of the signal. This type of top-down processing is critical to achieving high performance in speech recognition, but has yet to be applied to EEG analysis, because the decision tree or logic behind classifying an EEG based on expert knowledge has not been adequately codified.

Most previous studies have focused on small numbers of patients (typically less than 20) and have not demonstrated robust performance. For example, the Seizure Detection Challenge (*https://www.kaggle.com/c/seizure-detection*) (American Epilepsy Society et al., 2014) was based on "prolonged intracranial EEG recorded from four dogs with naturally occurring epilepsy and from 8 patients with medication resistant seizures during evaluation for epilepsy surgery." Most research focused

on EEG classification has focused on static classification of the data. An example of this type of approach was used by Bao et al. (2009), in which 94% classification accuracy was achieved for epilepsy detection on 6 normal and 6 epileptic patients. The signal was segmented into 20-sec. non-overlapping intervals (4096 samples at 200 Hz), and converted to a feature vector using a collection of heterogeneous features that included power spectral intensity, fractal dimension and other measures of nonlinearity. A probabilistic neural network (PNN) was applied to each channel, and the individual channel outputs were combined using a voting system.

Although the feature extraction process was manually optimized by using bandpass filters and increasing the segment length to 40 seconds, automated optimization could have been exploited if more data were available and discriminative training of features could be employed. Our experience in other classification problems, such as speech recognition (May et al., 2008) is that feature extraction plays a relatively insignificant role in overall system performance if the higher-level classification system is suitably powerful. In recent years, deep learning systems are circumventing a model-based feature extraction process (e.g., cepstral features) and operating directly on samples of the signal (Bengio et al., 2013) using an approach called representation learning.

The drawbacks of many of the static classification approaches being used are that the spectral behavior of the signal is coarsely quantized, and hence the channel-specific classifiers do not account for the temporal behavior of the signal. The analysis windows typically used are on the order of one to two seconds, which give poor temporal resolution. In addition, the sample size is typically too small to achieve true statistical significance. In such limited studies, there is always the potential for the classifier keying in on artifacts of the data, such as the background noise or the quality of the transducer conduction.

4.2. Typical Static Classification Systems

As machine learning research has evolved, interest in detecting low-level events has been growing. For example, Wulsin et al. (2010) defined a detection problem based on the six signal events described in Section 2.4. Their recall (0.22) and precision (0.19) rates were respectable, but their study was conducted

on only 11 patients. A small portion of the data was selected that contained significant EEG events, and this data was hand-labeled by a clinical epileptologist. For feature extraction, temporal features, such as mean energy, zero crossings and average peak/valley amplitude, and spectral features, such as frequency band power and wavelet energy, were combined. A variety of classifiers were evaluated including support vector machines (SVMs) and deep belief networks (DBNs) (a particular form of a deep learning-based system).

False positive rates were in the range of 4 to 25 false alarms per hour (Wulsin et al., 2011). A mean F-score was used as an overall performance metric. Best performance was obtained using a principal component analysis of the features followed by a decision-tree (DT) classifier. More interestingly, the performance of DTs, SVMs and DBNs was comparable and slightly better than k-nearest neighbor (kNN). We conjecture that this is a byproduct of insufficient training data and less mature deep learning technology. Head to head comparisons of deep learning systems on speech recognition tasks with large amounts of training data (Xiong et al., 2017; Hinton, et al., 2012) have shown two things: (1) for the same amount of training data, deep learning systems can exceed the performance of more conventional hidden Markov models (HMMs) if properly configured, and (2) the performance of HMMs can often approach the performance of deep learning systems if additional training data is provided. Bengio et al. (2013) also demonstrated that the feature extraction process could be automatically learned using a deep learning-based system.

We have replicated the Wulsin results on a publicly available seizure detection task (Goldberger et al., 2000) using a relatively simple standard HMM-based (Lu & Picone, 2013; Huang & Picone, 2002) approach as a proof of concept. We achieved a sensitivity of 96.5% and a false alarm rate of 3.8/hr. It was possible to build this system quickly and efficiently because time-aligned markers for the seizure events were provided with the data. It has been more challenging to reach our goal of 95% sensitivity on TUH-EEG due to the lack of time-aligned markers for events. This has necessitated investigation into active learning approaches that can be used to bootstrap systems from small amounts of data (Yang et al., 2016).

4.3. Feature Extraction

There are two general approaches to contemporary signal processing – functional or model-based features (e.g., cepstral coefficients) or statistically-based features that are embedded in some larger machine learning system. The latter appear to work well when there are ample amounts of data and offer the benefit of integrating key algorithmic enhancements such as discriminative training (Povey et al., 2008); the former is more traditional and has worked well in a variety of recognition applications (e.g., speech recognition) where there is significant subject matter expertise. Model-based features tend to work well for EEG analysis, although statistically-based features are slowly emerging as more data becomes available. In this section, we describe a standard model-based approach.

Neurologists most often review EEGs in 10 sec windows when doing fine-grained analysis. Larger time intervals are used to triage data and locate sections of interest, but detailed interpretations are most often performed on 10 sec windows. The time resolution used in these assessments is often on the order of 1 second. Therefore, we decompose the signal into fixed intervals typically of 1 to 2 seconds, which we refer to as an epoch. We have experimentally adjusted these parameters and found that a 1 second epoch is an appropriate tradeoff between time resolution, computational complexity and performance (Harati et al., 2015). We further subdivide this interval into 0.1 sec frames and use an overlapping window approach to compute features, as shown in Figure 9.

We use a standard cepstral coefficient-based feature extraction approach similar to the Linear Frequency Cepstral Coefficients (LFCCs) used in speech recognition (Picone, 1993; Davis & Mermelstein, 1980). Though popular alternatives to LFCCs in EEG processing include wavelets, which are used by many commercial systems, our experiments with such features have shown very little advantage over LFCCs on TUH-EEG. Unlike speech recognition which uses a mel scale for reasons related to speech perception, we use a linear frequency scale for EEGs, since there is no physiological evidence that a log scale is meaningful (Ebersole, 2014).

It is common in the LFCC approach to compute cepstral coefficients by computing a high resolution fast Fourier Transform, downsampling this representation using an oversampling approach based on a set

of overlapping bandpass filters, and transforming the output into the cepstral domain using a discrete cosine transform (Huang et al., 2001; Picone, 1993). The zeroth-order cepstral term is typically discarded and replaced with an energy term as described below.

There are two types of energy terms that are often used: time domain and frequency domain. Time domain energy is a straightforward computation using the log of the sum of the squares of the windowed signal:

$$E_t = \log\left(\frac{1}{N}\sum_{n=0}^{N-1}|x(n)|^2\right). \quad (1)$$

We use an overlapping analysis window (a 50% overlap was used here) to ensure a smooth trajectory of this feature. The energy of the signal can also be computed in the frequency domain by computing the sum of squares of the oversampled filter bank outputs after they are downsampled:

$$E_f = \log\left(\sum_{k=0}^{N-1}|X(k)|^2\right). \quad (2)$$

This form of energy is commonly used in speech recognition systems because it provides a smoother, more stable estimate of the energy that leverages the cepstral representation of the signal. However, the virtue of this approach has not been extensively studied for EEG processing.

To improve differentiation between transient pulse-like events (e.g., SPSW events) and stationary background noise, we have introduced a differential energy term that attempts to model the long-term change in energy. This term examines energy over a range of M frames centered about the current frame, and computes the difference between the maximum and minimum over this interval:

$$E_d = \max_m\left(E_f(m)\right) - \min_m\left(E_f(m)\right). \quad (3)$$

We typically use a 0.9 sec window for this calculation. This simple feature has proven to be surprisingly effective.

The final step to note in our feature extraction process is the familiar method for computing derivatives of features using a regression approach (Huang et al., 2001; Picone, 1993; Furui, 1986):

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2}, \quad (4)$$

where d_t is a delta coefficient, from frame t computed in terms of the static coefficients c_{t+n} to c_{t-n}. A typical value for N is 9 (corresponding to 0.9 secs) for the first derivative in EEG processing, and 3 for the second derivative. These features, which are often called deltas because they measure the change in the features over times, are one of the most well-known features in speech recognition (Huang et al., 2001). We typically use this approach to compute the derivatives of the features and then apply this approach again to those derivatives to obtain an estimate of the second derivatives of the features, generating what are often called delta-deltas. This triples the size of the feature vector (adding deltas and delta-deltas), but is well-known to deliver improved performance. This approach has not been extensively evaluated in EEG processing.

Dimensionality is something we must always pay attention to in classification systems since our ability to model features is directly related to the amount of training data available. The use of differential features raises the dimension of a typical feature vector from 9 (e.g., 7 cepstral coefficients, frequency domain energy and differential energy) to 27. There must be sufficient training data to support this increase in dimensionality or any improvements in the feature extraction process will be masked by poor estimates of the model parameters (e.g., Gaussian means and covariances). The impact of differential energy is shown in Figure 10.

We have used a subset of TUH-EEG that has been manually labeled for the six types of events described in Section 2.4 to tune feature extraction (Harati et al., 2015). We use the hidden Markov model system described in Section 4.4 as the classification system. The training set contains segments from 359 sessions while the evaluation set was drawn from 159 sessions. No patient appears more than once in the entire subset, which we refer to as the TUH EEG Short Set. The training set was designed to provide a sufficient number of examples to train statistical models. We refer to the 6 classes shown in Section 2.4 as the 6-way classification problem. This is not necessarily the most informative performance metric. It makes more sense to collapse the 3 background classes into one category. We refer to this second evaluation paradigm as a 4-way classification task: SPSW, GPED, PLED and BACKG. The latter class contains an enumeration of the 3 background classes. Finally, in order that we can produce a DET

curve (Martin et al., 1997), we also report a 2-way classification task in which we collapse the data into a target class (TARG) and a background class (BCKG).

DET curves are generated by varying a threshold typically applied to likelihoods to evaluate the tradeoff between detection rates and false alarms. However, it is also instructive to look at specific numbers in table form. Therefore, all experiments reported in the tables use a scoring penalty of 0, which essentially means we are evaluating the raw likelihoods returned from the classification system. In virtually all cases, the trends shown in these tables hold up for the full range of the DET curve.

The first series of experiments was run on a simple combination of features. A summary of these experiments is shown in Table 3, where error rates are given on the 6-way, 4-way and 2-way classification tasks described above. Cepstral-only features were compared with several energy estimation algorithms. It is clear that the combination of frequency domain energy and differential energy, system no. 5 in Table 3, provides a substantial reduction in performance. However, note that differential energy by itself (system no. 4) produces a noticeable degradation in performance. Frequency domain energy clearly provides information that complements differential energy. The improvements produced by system no. 5 hold for all three classification tasks. Though this approach increases the dimensionality of the feature vector by one element, the value of that additional element is significant and not replicated by simply adding other types of signal features (Harati et al., 2015).

A second set of experiments was run to evaluate the benefit of using differential features. These experiments are summarized in Table 4, where we again show 6-way, 4-way and 2-way classification error rates as described above. The addition of the first derivative adds about 7% absolute in performance (e.g., system no. 6 vs. system no. 1). However, when differential energy is introduced, the improvement in performance drops to only 4% absolute.

The story is somewhat mixed for the use of second derivatives. On the base cepstral feature vector, second derivatives reduce the error rate on the 6-way task by 4% absolute (systems nos. 6 and 11). However, the improvement for a system using differential energy is much less pronounced (system no. 5 in Table 3, systems nos. 10 and 15 in Table 4). In fact, it appears that differential energy and derivatives

do something very similar. Therefore, we evaluated a system that eliminates the second derivative for differential energy. This system is labeled no. 16 in Table 4. We obtained a small but significant improvement in performance over system no. 10. The improvement on 4-way classification was larger, which indicates more of an impact on differentiating between PLEDs, GPEDs and SPSW vs. background. This is satisfying since this feature was designed to address this problem.

The results shown in Tables Table 3 and Table 4 hold up under DET curve analysis as well. DET curves for systems nos. 1, 5, 10, and 15 are shown in Figure 11. We can see that the relative ranking of the systems is comparable over the range of the DET curves. First derivatives deliver a measurable improvement over absolute features (system no. 10 vs. no. 5). Second derivatives do not provide as significant an improvement (system no. 15 vs. no. 10). Differential energy provides a substantial improvement over the base cepstral features.

4.4. Hidden Markov Models

An overview of a generic system to automatically interpret EEGs is shown in Figure 12. This system incorporates a signal event detector that operates on each channel using channel independent models, and two stages of postprocessing to produce epoch labels. An N-channel EEG is transformed into N independent feature streams using a standard sliding window based approach. These features are then transformed into EEG signal event hypotheses using a standard hidden Markov Model (HMM) recognition system (Huang et al., 2002). These hypotheses are postprocessed by examining temporal and spatial context to produce epoch labels. The system detects three events of clinical interest (SPSW, PLED and GPED) and three events that map to background (ARTF, EYEM and BCKG) as described in Section 2.4.

A multichannel EEG signal is input to the system, typically as a European Data Format (EDF) file. A subset of the channels corresponding to a standard 10/20 EEG are selected. The signal is converted to a sequence of feature vectors as previously described. A group of frames are classified into an event on a per-channel basis using a hidden Markov model-based (Rabiner, 1991; Picone, 1990) classifier. This approach, which we borrow heavily from speech recognition (Huang et al., 2002; Deshmukh et al., 1999),

uses a left-to-right HMM topology (Picone, 1990) to encode the temporal evolution of the signal. Though there is no direct physiological or neurological motivation for this topology, experiments on alternate topologies have not proven to result in a significant gain on EEG or speech recognition experiments. The standard three-state model works surprisingly well across a wide range of applications.

HMMs are trained for each of the six classes using data for all channels (channel-specific HMMs have not proven to provide a significant gain in performance). Each incoming epoch for each channel is processed through the system, resulting in a likelihood vector that models the probability that the epoch could have been generated from the corresponding model. The event label for a channel is selected based on the most probable class – a forced-choice hypothesis test.

The second level of the system essentially examines multiple adjacent epochs in time, which we refer to as temporal context, and multiple channels, which we refer to as spatial context since each channel is associated with a location of an electrode on a patient's head. There are a wide variety of algorithms that can be used to produce a decision from these inputs. For example, early work in EEG interpretation used a majority vote (Ahangi et al., 2013). We have explored three approaches: (1) a simple heuristic mapping that makes decisions based on a predefined order of preference (e.g. SPWS > PLED > GPED > ARTF > EYEM > BCKG); (2) application of a random forest classification tree approach (Breiman, 2001) that we have used successfully for a number of other applications; and (3) a stacked denoising autoencoder (Bengio et al., 2007; Vincent et al., 2008) that has been successfully used in many deep learning systems. Performance of these three approaches is summarized in Table 5.

The detection rate (DET) is defined as the percentage of correct recognitions for the classes (SPSW, GPED, PLED). The false alarm rate (FA) is defined as the percentage of incorrect recognitions for (BCKG, ARTF, EYEM) – the number of times these classes are detected as one of the three non-background classes (SSW, GPED, PLED) divided by the number of times they occur. The error rate (ERR) is defined as the number of times any class is incorrectly detected divided by the total number of epochs. Applying a machine learning component as a postprocessor to the event detection level achieves our goal of maximizing the DET rate and minimizing the FA rate.

Further, the errors that the current system makes are not "mission critical." For example, it is not critical that the system detect every spike accurately. Distinguishing spikes from background signal is very hard even for a trained neurologist. However, alerting a neurologist that an EEG has spikes in it, and showing the approximate locations of these spikes, is of great value. The neurologist can then manually review only the areas of the signal that have events of interest such as spikes.

One can often argue that it is relatively straightforward to build an automatic labeling system if fully annotated data is available. This a great oversimplification since most realistic applications require a significant amount of engineering. For the EEG problem, the situation is more complex as the agreement between transcribers is relatively low. We have developed a highly effective active learning approach to training models that requires a very small amount of transcribed data (Yang et al., 2016). An overview of the process is shown in Figure 13. We use a small amount of manually transcribed data to seed the models. We then use these models to automatically label the data and produce likelihoods that each epoch could have been one of the six classes. We continue by using these new labels to sort the data and retrain the models. We only consider labels for data for which we are confident that the labels are correct, and for data of interest such as SPSW events. We typically set a confidence threshold of about 80%. We continue iterating on this process until classification performance on the training data and/or development test data is adequate. In practice, convergence occurs quickly and even after only three iterations models have improved significantly.

4.5. Normal / Abnormal Detection

A similar version of this system can be used to do normal/abnormal classification (Lopez, 2017). The automated classification of an EEG record as normal or abnormal represents a significant step for the reduction of the visual bias intrinsic to the subjectivity of the record's interpretation. The main characteristics of an adult normal EEG are (Ebersole, 2014):

- Reactivity: Response to certain physiological changes or provocations.
- Alpha Rhythm: Waves originated in the occipital lobe (predominantly), between 8-13 Hz and 15 to 45 μV.

- Mu Rhythm: Central rhythm of alpha activity commonly between 8-10 Hz visible in 17% to 19% of adults.
- Beta Activity: Activities in the frequency bands of 18-25 Hz, 14-16 Hz and 35-40 Hz.
- Theta Activity: Traces of 5-7 Hz activity present in the frontal or frontocentral regions of the brain.

Neurologists follow procedures similar to the one summarized in Figure 14 and can usually make this determination by examining the first few minutes of a recording. Hence, we focused on examining the first 60 secs of an EEG to calibrate the difficulty of the task.

We selected a demographically balanced subset of TUH-EEG through manual review that consisted of 202 normal EEGs and 200 abnormal EEGs. These sets were further partitioned into a training set (102 normal/100 abnormal), development test set (50 normal/50 abnormal) and an evaluation set (50 normal/50 abnormal). To create an appropriate experimental paradigm, only one EEG channel was selected for consideration. Examination of manual interpretation techniques practiced by experts revealed that the most promising channel to explore was the differential measurement T5-O1, which is part of the popular TCP montage. This channel represents the difference between two electrodes located in the left temporal and occipital lobes. The first 60 seconds of each recording were used to extract signal features using the process described in Section 4.3. The feature vectors (60 sec x 10 frames/sec = 600 vectors) were concatenated into a supervector of dimension 600 x 27 = 16,200. The dimensionality of the supervector was reduced using class-dependent Principal Components Analysis (PCA) in which we retained the N most significant eigenvectors of the covariance matrix.

Two standard algorithms were explored: k-Nearest Neighbor (kNN) (Duda et al., 2001) and Random Forest Ensemble Learning (RF) (Breiman, 2001). We conducted additional searches for an optimal set of parameters for each system. In Table 6, we compare performance of these two systems to a baseline based on random guessing. The first system is random guessing based on priors. The second system is kNN with k = 20 and a PCA dimension of 86. The third system is RF with N_t = 50 and a PCA dimension of 86. The tuned kNN and RF systems outperform random guessing based on priors, which is a promising outcome for these experiments. However, there is a high confusion rate for normal EEGs. The dominant

error is a normal EEG classified as abnormal. This could be explained by the presence of benign variants, or electroencephalographic patterns that resemble abnormalities, but do not qualify as events that would be of significance for the abnormal classification of a record.

Next, we developed a variant of the HMM system depicted in Figure 12. We used standard three-state HMMs with three Gaussian mixture components per state as before. Feature vectors were computed every 0.1 secs, generating a total of 600 feature vectors over the first 60 secs of the EEG signal. Two approaches were followed. First, as discussed before, we built a supervector for each frame consisting of a concatenation of feature vectors for each channel. This supervector was reduced to a dimension of 20 using PCA. This system, shown in Table 6 as PCA-HMM, reduced the error rate from 32% to 26%. The goal in this experiment was to evaluate whether PCA could adequately model localization of the event since it has access to data from all channels.

The second approach selected a single channel and applied the raw feature vector to the same three-state HMM. This system, referred to as GMM-HMM, further reduced the error rate to 17%, which is approaching human performance on this task (when neurologists are constrained to look at the first 60 seconds of data). In Table 7, we explore performance as a function of the channel used. We verified that performance for T5-O1 was, in fact, optimal, as predicted from neuroscience considerations.

4.6. Seizure Detection

We have also trained the system depicted in Figure 12, using a basic three-state HMM topology and SdA postprocessing, to perform a seizure/no-seizure binary decision. Here, we use a subset of TUH-EEG that was specifically labeled using term-based labels for seizures by a series of experts (Golmohammadi, 2017). The data was carefully annotated by a team of students who have been trained to annotate seizures. Their work has been evaluated against expert neurologists who marked a portion of the same data and shown to have an IRA that exceeds that for the expert neurologists. Each event in the evaluation data has been reviewed by at least five different annotators.

Performance was as follows: Sensitivity – 29.2%; Specificity – 66.7%; False Alarms/24 hrs – 78; ATWV – -0.42. This is the first significant benchmark on seizure detection for TUH-EEG, and the first

benchmark in this application space that uses ATWV. Commercially available tools tend to perform at a sensitivity of around 30% (Scheuer, 2017) with a high false alarm rate on this task based on other clinical evaluations. Our internal evaluations of these commercial systems indicate that our HMM baseline system delivered performance competitive to these systems. However, the ATWV score is extremely poor for these systems largely due to the large emphasis this metric places on false alarms. Hence, it is understandable that these systems fail to perform well in live clinical settings. The main challenge on this task is to accurately identify short seizures and the start time of slowly-evolving seizures.

5. SUMMARY AND FUTURE DIRECTIONS

Biomedicine is entering a new age of data-driven discovery driven by ubiquitous computing power, a machine learning revolution, and high speed internet connections. Access to massive quantities of properly curated data is now the critical bottleneck to advancement in many areas of biomedical research. Ironically, clinicians generate enormous quantities of data daily, but that information is almost exclusively sequestered in secure archives where it cannot be used for research by the biomedical research community. The quantity, quality, and variability of such data represent a significant unrealized potential, which is doubly unfortunate considering that the cost of generating that data has already been borne.

In this chapter, we have introduced the problem of automatic interpretation of EEGs and described a paradigm that can be used to develop high performance technology. We began by articulating the problem in terms that machine learning research can understand. We described the development of a big data corpus, TUH-EEG, that is enabling the application of state of the art statistical models. We introduced a baseline system that integrates data-driven model-based parameterizations and subject matter expertise to achieve high performance in EEG signal event detection. This system was based on hidden Markov models because of their proven ability to model sequential data.

Future research is now focused on applying deep learning methodologies to these problems. There are a wide variety of deep learning technologies available that support the integration of temporal and spatial constraints and can learn autonomously from the data. Initial experiments using networks based on long

short-term memory (LSTM) (Graves et al., 2013; Hochreiter & Schmidhuber, 1997) and convolutional neural networks (CNN) (Abdel-Hamid et al., 2012) have shown modest improvements in performance in terms of sensitivity and specificity, but comparable false alarm rates. Experiments in which cepstral-based features are replaced by several additional layers of a deep learning system are also showing comparable performance but no significant gains yet. This is more than likely due to a lack of a large amount of annotated data, as well as the lack of maturity of our pre-training and unsupervised learning approaches in deep learning. Such systems more than likely will require one or two orders of magnitude more data than existing HMM-based systems, and such data resources are under development at the Neural Engineering Data Consortium (*www.nedcdata.org*).

However, if we have significantly more data, there are some fundamental challenges that need to be addressed. Incorporating better features into the system will be critical. These features need to expose the deep learning systems to similar information that neurologists use to interpret EEGs, so both spatial and temporal context is required. This greatly increases the complexity of the networks and raises some computational issues. It also further underscores the importance of more data, since the dimensionality of these systems gets quite large. Similarly, the imbalanced nature of the data must be dealt with in a fundamental manner. The high false alarm rate is also a fundamental barrier to the acceptance of the technology. Future research will address these problems using a range of approaches that include transfer learning (Taylor, 2009), confidence measures (Yu et al., 2011) and discriminative training (Manohar et al., 2015). Our ultimate research goal is the prediction of seizures 30 to 60 minutes before they occur, and research is underway to address this challenge as well.

ACKNOWLEDGMENTS

Research reported in this publication was most recently supported by the National Human Genome Research Institute of the National Institutes of Health under award number U01HG008468. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This material is also based in part upon work supported by the National Science Foundation under Grant No. IIP-1622765. Any opinions, findings, and conclusions or

recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The TUH EEG Corpus work was funded by (1) the Defense Advanced Research Projects Agency (DARPA) MTO under the auspices of Dr. Doug Weber through the Contract No. D13AP00065, (2) Temple University's College of Engineering and (3) Temple University's Office of the Senior Vice-Provost for Research.

In any project of this nature, there are many people who have contributed knowledge, data, and technology. We owe an enormous debt to Dr. Mercedes Jacobson, MD, Professor of Neurology, Lewis Katz School of Medicine, has been our colleague and mentor, and was responsible for our access to Temple Hospital's data. We are particularly grateful to Dr. Steven Tobochnik, MD, currently with New York-Presbyterian Hospital-Columbia University for his willingness to review data and answer many questions about the science. We have been fortunate to benefit from a group of over 135 neurologists that have advised us on some portion of this project.

Without the dedication of many graduate and undergraduate students at the Neural Engineering Data Consortium, none of these resources would exist. Amir Harati, Meysam Golmohammadi and Silvia Lopez were the lead developers of the technology. Vinit Shah, Eva von Weltsin and James Riley McHugh were primarily responsible for the data development. Saeedeh Ziyabari was responsible for the development of the evaluation software. Scott Yang developed the active learning technology. Elliott Krome was responsible for the development of interactive demonstration tools. Over 20 undergraduates have participated in data entry work for the TUH EEG Corpus.

REFERENCES

Acharya, J., Hani, A., Thirumala, P., & Tsuchida, T. (2016). American Clinical Neurophysiology Society Guideline 3: A Proposal for Standard Montages to Be Used in Clinical EEG. *Journal of Clinical Neurophysiology: Official Publication of the American Electroencephalographic Society*, 33(4).

Ahangi, A., Karamnejad, M., Mohammadi, N., Ebrahimpour, R., & Bagheri, N. (2013). Multiple classifier system for EEG signal classification with application to brain–computer interfaces. *Neural Computing and Applications*, *23*(5), 1319–1327.

Al-Qazzaz, N. K., Ali, S. H. B. M., Ahmad, S. A., Chellappan, K., Islam, M. S., & Escudero, J. (2014). Role of EEG as biomarker in the early detection and classification of dementia. *The Scientific World Journal*.

Altman, D. G., & Bland, J. M. (1994). Diagnostic Tests 1: Sensitivity And Specificity. *BMJ: British Medical Journal*. England: British Medical Association.

Ardeshna, N. I. (2016). EEG and Coma. *The Neurodiagnostic Journal*, *56*(1), 1–16. http://doi.org/10.1080/21646821.2015.1114879

Baldassano, S., Wulsin, D., Ung, H., Blevins, T., Brown, M.-G., Fox, E., & Litt, B. (2016). A novel seizure detection algorithm informed by hidden Markov model event states. *Journal of Neural Engineering*, *13*(3), 36011.

Bao, F. S., Gao, J.-M., Hu, J., Lie, D. Y.-C., Zhang, Y., & Oommen, K. J. (2009). Automated epilepsy diagnosis using interictal scalp EEG. In *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 6603–6607). Minneapolis, Minnesota, USA.

Beelen, T. van. (2013). *EDFbrowser*. (Retrieved from http://www.teuniz.net/edfbrowser/.)

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing System* (pp. 153–160). Vancouver, B.C., Canada.

Bergey, S., & Picone, J. (2017). A Description of the EDF Header. (Retrieved March 7, 2017, from *https://www.isip.piconepress.com/projects/tuh_eeg/doc/edf/*.)

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Britton, J. W., Frey, L. C., Hopp, J. L., Korb, P., Koubeissi, M., Lievens, W., … St. Louis, E. K. (2016). *Electroencephalography (EEG): An Introductory Text and Atlas of Normal and Abnormal Findings in Adults, Children, and Infants [Internet]*. (E. K. St. Louis & L. C. Frey, Eds.) *American Epilepsy Society* (1st ed.). Chicago, Illinois, USA: National Institutes of Health. (Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK390352/.)

Brzezinski, R. (2016). *HIPAA Privacy and Security Compliance - Simplified: Practical Guide for Healthcare Providers and Managers 2016 Edition* (3rd ed.). Seattle, Washington, USA: CreateSpace Independent Publishing Platform.

Chandran, V. (2012). Time-varying bispectral analysis of visually evoked multi-channel EEG. *EURASIP Journal on Advances in Signal Processing*, 2012(1), 1–22.

Christensen, M., Dodds, A., Sauer, J., & Watts, N. (2014). Alarm setting for the critically ill patient: a descriptive pilot survey of nurses' perceptions of current practice in an Australian Regional Critical Care Unit. *Intensive & Critical Care Nursing*. Netherlands: Elsevier B.V.

Dash, D., Hernandez-Ronquillo, L., Moien-Afshari, F., & Tellez-Zenteno, J. (2012). Ambulatory EEG: a cost-effective alternative to inpatient video-EEG in adult patients. *Epileptic Disorders*, *14*(3), 290–297.

Davis, S., & Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*(4), 357–366.

Deshmukh, N., Ganapathiraju, A., & Picone, J. (1999). Hierarchical search for large-vocabulary conversational speech recognition: working toward a solution to the decoding problem. *IEEE Signal Processing Magazine*, *16*(5), 84–107.

Doddington, G. R., Przybocki, M. A., Martin, A. F., & Reynolds, D. A. (2000). The NIST speaker recognition evaluation – Overview, methodology, systems, results, perspective. *Speech Communication*, *31*(2), 225–254. http://doi.org/10.1016/S0167-6393(99)00080-1

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern Classification*. (2nd, Ed.). New York City, New York, USA: John Wiley & Sons, Inc.

Ebersole, J. S., & Pedley, T. A. (2014). *Current practice of clinical electroencephalography* (4th ed.). Philadelphia, Pennsylvania, USA: Wolters Kluwer.

Electroencephalography. (2017, April 19). In Wikipedia, The Free Encyclopedia. Retrieved 16:57, April 16, 2017, from *https://en.wikipedia.org/w/index.php?title=Electroencephalography&oldid=774613669*.

Ercegovac, M. (2010). 33. Importance of EEG in brain death diagnosis. *Clinical Neurophysiology*. Elsevier Ireland Ltd.

Erman, L. D., Hayes-Roth, F., Lesser, V. R., & Reddy, D. R. (1980). The Hearsay-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty. *ACM Comput. Surv.*, *12*(2), 213–253.

Finnigan, S., & van Putten, M. J. a M. (2013). EEG in ischaemic stroke: quantitative EEG can uniquely inform (sub-)acute prognoses and clinical management. Clinical Neurophysiology: *Official Journal of the International Federation of Clinical Neurophysiology*, 124(1), 10–19.

Furui, S. (1986). Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *34*(1), 52–59.

Golmohammadi, M., Shah, V., Lopez, S., Ziyabari, S., Yang, S., Camaratta, J., … Picone, J. (2017). The TUH EEG Seizure Corpus. In *American Clinical Neurophysiology Society* (p. 1). Phoenix, Arizona, USA.

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech Recognition With Deep Recurrent Neural Networks. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (3), 6645–6649.

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*, *126*(9), 1661–1669. http://doi.org/10.1016/j.clinph.2014.11.008.

Harabagiu, S. (2016). Active Deep Learning-Based Annotation of Electroencephalography Reports for Patient Cohort Identification. In M. Dunn (Ed.), *The BD2K Guide to the Fundamentals of Data Science* (p. 1). Bethesda, Maryland, USA: National Institutes of Health. (Retrieved from http://www.bigdatau.org/data-science-seminars.)

Harabagiu, S., Goodwin, T., Maldonado, R., & Taylor, S. (2016). Active Deep Learning-Based Annotation of Electroencephalography Reports for Patient Cohort Retrieval. In *Big Data to Knowledge All Hands Grantee Meeting* (p. 1). Bethesda, Maryland, USA: National Institutes of Health.

Harati, A., Golmohammadi, M., Lopez, S., Obeid, I., & Picone, J. (2015). Improved EEG event classification using differential energy. In *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* (pp. 1–4). Philadelphia: IEEE.

Harati, A., Lopez, S., Obeid, I., Jacobson, M., Tobochnik, S., & Picone, J. (2014). The TUH EEG Corpus: A Big Data Resource for Automated EEG Interpretation. In *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium* (pp. 1–5). Philadelphia, Pennsylvania, USA.

He, H., Ma, Y., & Obooks, W. O. L. U. A. (2013). *Imbalanced learning: foundations, algorithms, and applications* (Vol. 1). Hoboken, New Jersey: John Wiley & Sons, Inc.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–80.

Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Upper Saddle River, New Jersey, USA: Prentice Hall.

Jacobs, I. M., & Wozencraft, J. M. (1965). *Principles of communication engineering* (1st ed.). Long Grove, Illinois USA: Waveland Pr Inc.

Japkowicz, N., & Shah, M. (2011). *Evaluating Learning Algorithms: a classification perspective*. Cambridge; New York; Cambridge University Press.

Jurcak, V., Tsuzuki, D., & Dan, I. (2007). 10/20, 10/10, and 10/5 systems revisited: Their validity as relative head-surface-based positioning systems. *NeuroImage*, 34(4), 1600–1611.

Kemp, R. (2013). European Data Format. (Retrieved March 7, 2017, from *http://doi.org/http://www.edfplus.info*.)

Klassen, B. T., Hentz, J. G., Shill, H. A., Driver-Dunckley, E., Evidente, V. G. H., Sabbagh, M. N., … Caviness, J. N. (2011). Quantitative EEG as a predictive biomarker for Parkinson disease dementia. *Neurology*, 77(2), 118–124.

LaRoche, S., & Collection, E. A. C. S. (2013). *Handbook of ICU EEG monitoring*. New York, New York, USA: Demos Medical Pub.

Lee, K.-F., & Hon, H.-W. (1989). Speaker-Independent Phone Recognition Using Hidden Markov Models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *37*(11), 1641–1648.

Lin, E.-B., & Shen, X. (2011). Wavelet analysis of EEG signals. In *Proceedings of the IEEE National Aerospace and Electronics Conference (NAECON)* (pp. 105–110). Dayton, Ohio, USA.

Lopez, S. (2017). *Automated Identification of Abnormal EEGs*. Temple University. (Available at http://www.isip.piconepress.com/publications/ms_theses/2017/abnormal.)

Lopez, S., Gross, A., Yang, S., Golmohammadi, M., Obeid, I., & Picone, J. (2016). An Analysis of Two Common Reference Points for EEGs. In *IEEE Signal Processing in Medicine and Biology Symposium* (pp. 1–4). Philadelphia, Pennsylvania, USA.

Lopez, S., Suarez, G., Jungries, D., Obeid, I., & Picone, J. (2015). Automated Identification of Abnormal EEGs. In *IEEE Signal Processing in Medicine and Biology Symposium* (pp. 1–4). Philadelphia, Pennsylvania, USA.

Lowerre, B. (1980). The HARPY Speech Understanding System. In W. Lea (Ed.), *Trends in Speech Recognition* (pp. 576–586). Englewood Cliffs, New Jersey, USA: Prentice Hall.

Mandal, A., Prasanna Kumar, K. R., & Mitra, P. (2014). Recent developments in spoken term detection: a survey. *International Journal of Speech Technology*, *17*(2), 183–198.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press.

Manohar, V., Povey, D., & Khudanpur, S. (2015). Semi-supervised maximum mutual information training of deep neural network acoustic models. In *Proceedings of INTERSPEECH* (pp. 2630–2634). Dresden, Germany: ISCA.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The DET curve in assessment of detection task performance. In *Proceedings of Eurospeech* (pp. 1895–1898). Rhodes, Greece.

May, D., Srinivasan, S., Ma, T., Lazarou, G., & Picone, J. (2008). Continuous Speech Recognition Using Nonlinear Dynamical Invariants. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Las Vegas, Nevada, USA.

Michel, C. M., & Murray, M. M. (2012). Towards the utilization of EEG as a brain imaging tool. *NeuroImage*, 61(2), 371.

Misulis, K. E., & Abou-Khalil, B. (2014). *Atlas of EEG and seizure semiology and management* (2nd ed., p. 384). Oxford: Oxford University Press.

NIST. (2010). (Retrieved from *http://www.itl.nist.gov/iad/mig//tools/*.)

Nizam, A., Chen, S., & Wong, S. (2013). Best-Case Kappa Scores Calculated Retrospectively From EEG Report Databases. *Journal of Clinical Neurophysiology*. United States: Copyright American Clinical Neurophysiology Society.

Nolte, J., & Sundsten, J. W. (2015). *The human brain: an introduction to its functional anatomy* (7th ed.). St. Louis, Mo: Mosby.

Obeid, I., & Picone, J. (2015). *NSF ICORPS Team: AutoEEG*. Philadelphia, Pennsylvania, USA. (Retrieved from https://www.isip.piconepress.com/publications/reports/2016/nsf/icorps/report_v01.pdf.)

Obeid, I., & Picone, J. (2016). The Temple University Hospital EEG Data Corpus. *Frontiers in Neuroscience, Section Neural Technology*, *10*, 196.

Obeid, I., Picone, J., & Harabagiu, S. (2016). Automatic Discovery and Processing of EEG Cohorts from Clinical Records. In *Big Data to Knowledge All Hands Grantee Meeting* (p. 1). Bethesda, Maryland, USA: National Institutes of Health.

Picone, J. (1990). Continuous Speech Recognition Using Hidden Markov Models. *IEEE ASSP Magazine*, *7*(3), 26–41.

Picone, J. (1993). Signal modeling techniques in speech recognition. *Proceedings of the IEEE*, *81*(9), 1215–1247.

Picone, J., & Obeid, I. (2016). Fundamentals in Data Science: Data Wrangling, Normalization, Preprocessing of Physiological Signals. In M. Dunn (Ed.), *The BD2K Guide to the Fundamentals of Data Science* (p. 1). Bethesda, Maryland, USA: National Institutes of Health. (Retrieved from http://www.bigdatau.org/data-science-seminars.)

Picone, J., Obeid, I., & Harabagiu, S. (2015). Automatic Discovery and Processing of EEG Cohorts from Clinical Records. In *Big Data to Knowledge All Hands Grantee Meeting* (p. 1). Bethesda, Maryland, USA. (Retrieved from http://www.isip.piconepress.com/publications/conference_presentations/2015/nih_bd2k/cohort/.)

Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G., & Visweswariah, K. (2008). Boosted MMI for model and feature-space discriminative training. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Las Vegas, Nevada, USA.

Rabiner, L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Rosso, O. A., Martin, M. T., Figliola, A., Keller, K., & Plastino, A. (2006). EEG analysis using wavelet-based information tools. *Journal of Neuroscience Methods*, *153*(2), 163–182.

Rudrashetty, S. M., Pyakurel, A., Karumuri, B., Liu, R., Vlachos, I., & Iasemidis, L. (2015). Differential diagnosis of sleep disorders based on EEG analysis. *Journal of the Mississippi Academy of Sciences*. Mississippi Academy of Sciences.

Scheuer, M. L., Bagic, A., & Wilson, S. B. (2017). Spike detection: Inter-reader agreement and a statistical Turing test on a large data set. *Clinical Neurophysiology*, 128(1), 243–250.

Sheorajpanday, R. V. A., Nagels, G., Weeren, A. J. T. M., & De Deyn, P. P. (2011). Quantitative EEG in ischemic stroke: correlation with infarct volume and functional status in posterior circulation and lacunar syndromes. *Clinical Neurophysiology*: *Official Journal of the International Federation of Clinical Neurophysiology*, 122(5), 884–890.

Smith, S. (2005). EEG in the diagnosis, classification, and management of patients with epilepsy. *Journal of Neurology, Neurosurgery, and Psychiatry*, *76*(Suppl 2), ii2-ii7.

Society, A. E., (NINDS), N. I. of H., Pennsylvania, U. of, & Mayo Clinic. (2014). UPenn and Mayo Clinic's Seizure Detection Challenge. (Retrieved from *https://www.kaggle.com/c/seizure-detection*.)

Stroink, H., Schimsheimer, R.-J., de Weerd, A. W., Geerts, A. T., Arts, W. F., MC, E., … van Donselaar, C. A. (2006). Interobserver reliability of visual interpretation of electroencephalograms in children with newly diagnosed seizures. *Developmental Medicine & Child Neurology*, *48*(5), 374–377.

Sutter, R., Kaplan, P. W., Valença, M., & De Marchis, G. M. (2015). EEG for Diagnosis and Prognosis of Acute Nonhypoxic Encephalopathy: History and Current Evidence. *Journal of Clinical Neurophysiology*. United States: by the American Clinical Neurophysiology Society.

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.

Tatum, W., Husain, A., Benbadis, S., & Kaplan, P. (2007). *Handbook of EEG Interpretation*. (Kirsch, Ed.). New York City, New York, USA: Demos Medical Publishing.

Taylor, M. E. (2009). Transfer in reinforcement learning domains. In *Studies in Computational Intelligence* (Vol. 216, p. 218). Berlin, Germany: Springer.

Thiess, M., Krome, E., Golmohammadi, M., Obeid, I., & Picone, J. (2016). Enhanced visualizations for improved real-time EEG monitoring. In *2016 IEEE Signal Processing in Medicine and Biology Symposium* (SPMB) (p. 1).

van Donselaar, C., Schimsheimer, R., AT, G., & Declerck, A. (1992). Value of the electroencephalogram in adult patients with untreated idiopathic first seizures. *Archives of Neurology*, *49*(3), 231–237.

van Tricht, M. J., Ruhrmann, S., Arns, M., Müller, R., Bodatsch, M., Velthorst, E., … Nieman, D. H. (2014). Can quantitative EEG measures predict clinical outcome in subjects at Clinical High Risk for psychosis? A prospective multicenter study. *Schizophrenia Research*, 153(1–3), 42–47.

Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning* (pp. 1096–1103). New York, NY, USA.

Wang, C., Hu, X., Yao, L., Xiong, S., & Zhang, J. (2011). Spatio-temporal pattern analysis of single-trial EEG signals recorded during visual object recognition. *Science China Information Sciences*, 54(12), 2499–2507.

Waterstraat, G., Burghoff, M., Fedele, T., Nikulin, V., Scheer, H. J., & Curio, G. (2015). Non-invasive single-trial EEG detection of evoked human neocortical population spikes. *NeuroImage*, *105*, 13–20.

Wulsin, D. F., Gupta, J. R., Mani, R., Blanco, J. A., & Litt, B. (2011). Modeling electroencephalography waveforms with semi-supervised deep belief nets: fast classification and anomaly measurement. *Journal of Neural Engineering*, *8*(3), 36015.

Wulsin, D., Blanco, J., Mani, R., & Litt, B. (2010). Semi-Supervised Anomaly Detection for EEG Waveforms Using Deep Belief Nets. In *International Conference on Machine Learning and Applications (ICMLA)* (pp. 436–441). Washington, D.C., USA.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., … Zweig, G. (2016). Achieving Human Parity in Conversational Speech Recognition. Ithaca, New York, USA: Cornell University Library (arvix.org). (Retrieved from https://arxiv.org/abs/1610.05256.)

Yamada, T., & Meng, E. (2009). *Practical Guide for Clinical Neurophysiologic Testing: EEG*. Philadelphia, Pennsylvania, USA: Lippincott Williams & Wilkins.

Yang, S., López, S., Golmohammadi, M., Obeid, I., & Picone, J. (2016). Semi-automated Annotation of Signal Events in Clinical EEG Data. In *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium* (pp. 1–5). Philadelphia, Pennsylvania, USA.

Yu, D., Li, J., & Deng, L. (2011). Calibration of Confidence Measures in Speech Recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, *19*(8), 2461–2473.

Ziyabari, S., Golmohammadi, M., Obeid, I., & Picone, J. (2017). An Analysis of Objective Performance Metrics for Automatic Interpretation of EEG Signal Events. Under Review. (Available at: *http://www.isip.piconepress.com/publications/unpublished/journals/2017/neural_engineering/metrics*.)

LIST OF FIGURES

Figure 13. An overview of our iterative HMM training procedure is shown. An active learning approach is used to bootstrap the system to handle large amounts of data.

Figure 14. The general process for identifying an abnormal EEG depends heavily on the observation of the Posterior Dominant Rhythm (PDR).

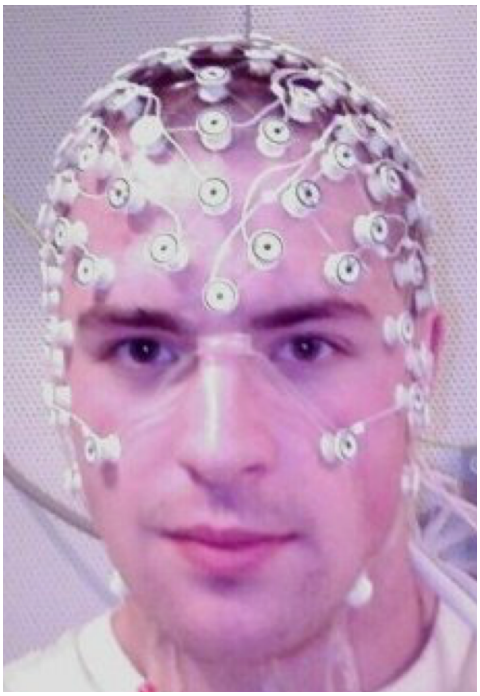


Figure 1. An EEG measures electrical activity (ionic currents) along the scalp using gold-plated or silver/silver chloride electrodes. The signals, typically in the microvolt range, are very noisy and must be viewed by examining differential voltages between and electrode and a ground point such as an electrode connected to the left ear. Photo adapted from Electroencephalography (2017, April 19).

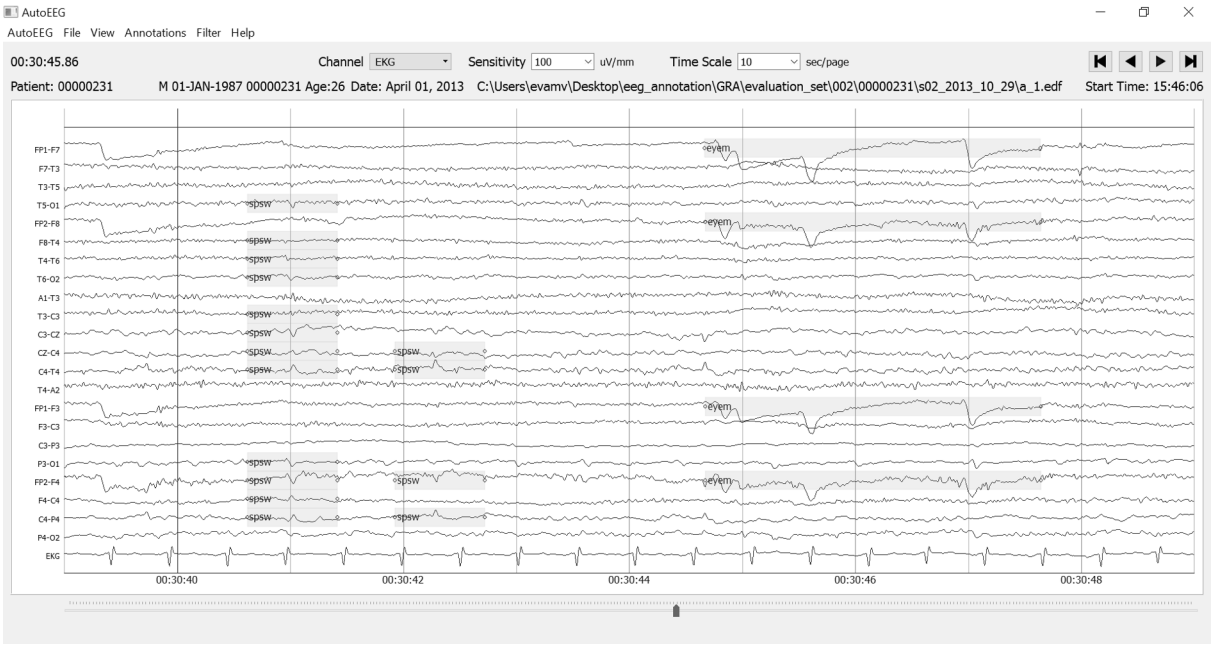
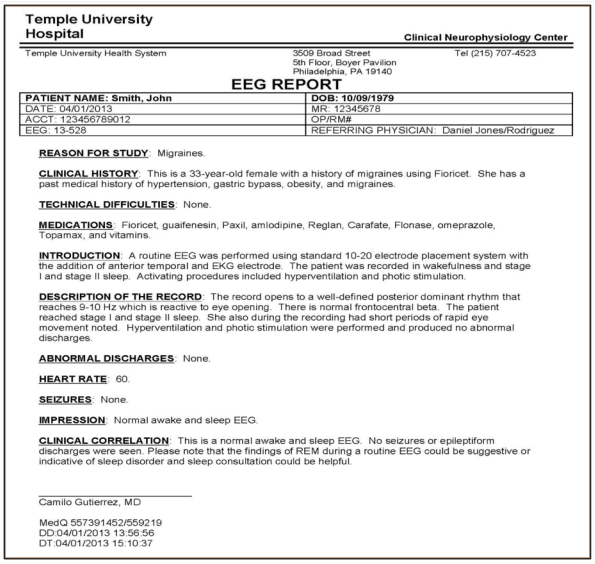


Figure 2. An EEG signal is a multi-channel signal typically consisting of 22 channels. Neurologists often visualize the data using a montage – a set of differential voltages designed to accentuate anomalous behavior like spikes and sharp waves. Event-based annotations, which are created on a per-channel basis, are shown.



Temple University Hospital

Clinical Neurophysiology Center

Temple University Health System
3509 Broad Street
5th Floor, Boyer Pavilion
Philadelphia, PA 19140
Tel (215) 707-4523

EEG REPORT

PATIENT NAME: Smith, John	**DOB: 10/09/1979**
DATE: 04/01/2013	MR: 12345678
ACCT: 123456789012	OP/RM#
EEG: 13-528	REFERRING PHYSICIAN: Daniel Jones/Rodriguez

REASON FOR STUDY: Migraines.

CLINICAL HISTORY: This is a 33-year-old female with a history of migraines using Fioricet. She has a past medical history of hypertension, gastric bypass, obesity, and migraines.

TECHNICAL DIFFICULTIES: None.

MEDICATIONS: Fioricet, guaifenesin, Paxil, amlodipine, Reglan, Carafate, Flonase, omeprazole, Topamax, and vitamins.

INTRODUCTION: A routine EEG was performed using standard 10-20 electrode placement system with the addition of anterior temporal and EKG electrode. The patient was recorded in wakefulness and stage I and stage II sleep. Activating procedures included hyperventilation and photic stimulation.

DESCRIPTION OF THE RECORD: The record opens to a well-defined posterior dominant rhythm that reaches 9-10 Hz which is reactive to eye opening. There is normal frontocentral beta. The patient reached stage I and stage II sleep. She also during the recording had short periods of rapid eye movement noted. Hyperventilation and photic stimulation were performed and produced no abnormal discharges.

ABNORMAL DISCHARGES: None.

HEART RATE: 60.

SEIZURES: None.

IMPRESSION: Normal awake and sleep EEG.

CLINICAL CORRELATION: This is a normal awake and sleep EEG. No seizures or epileptiform discharges were seen. Please note that the findings of REM during a routine EEG could be suggestive or indicative of sleep disorder and sleep consultation could be helpful.

_____
Camilo Gutierrez, MD

MedQ 557391452/559219
DD:04/01/2013 13:56:56
DT:04/01/2013 15:10:37

Figure 3. An example of a typical EEG Report. The format of the report is fairly standard across institutions and contains a brief medical history, medication history, a description of the neurologist's interpretation of the EEG signal, and the implications of these findings (clinical correlation).

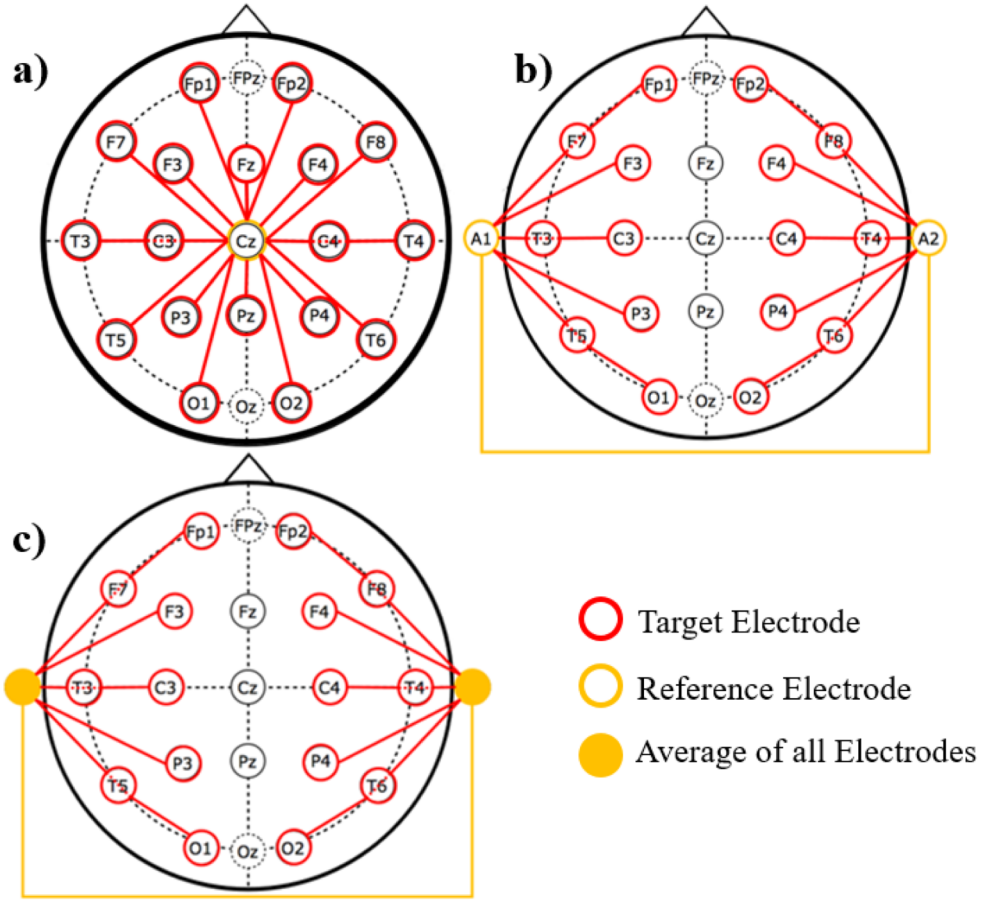
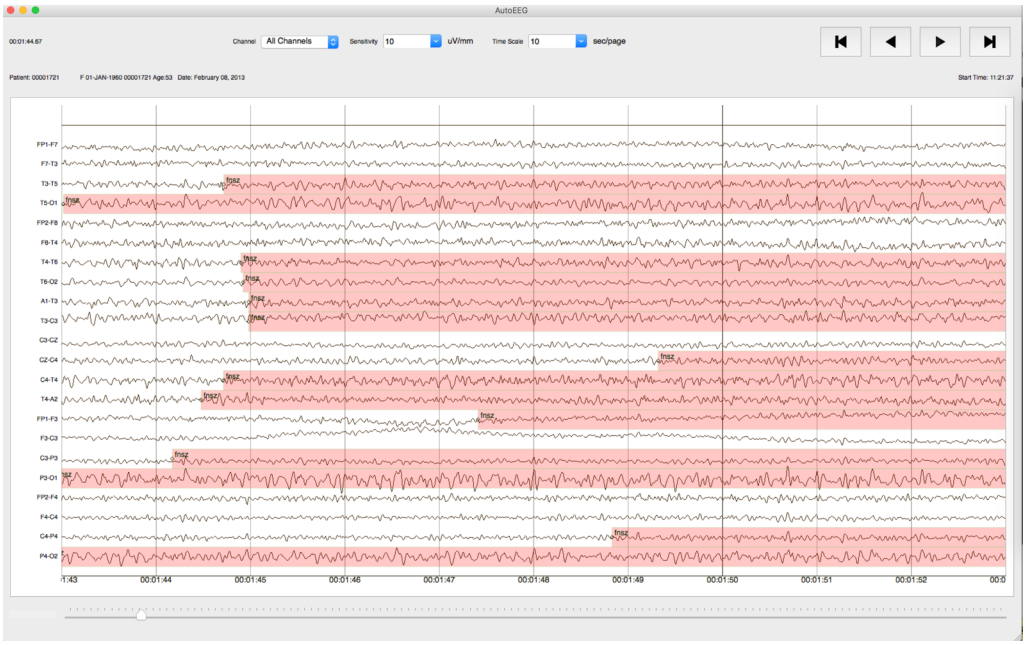


Figure 4. The electrode locations are shown for three common referential montages for a standard 10/20 configuration: a) the Common Vertex Reference (C_z), b) the Linked Ears Reference (LE) and c) the Average Reference (AR).

a) a typical seizure event



b) the corresponding spectrogram for a subset of the channels

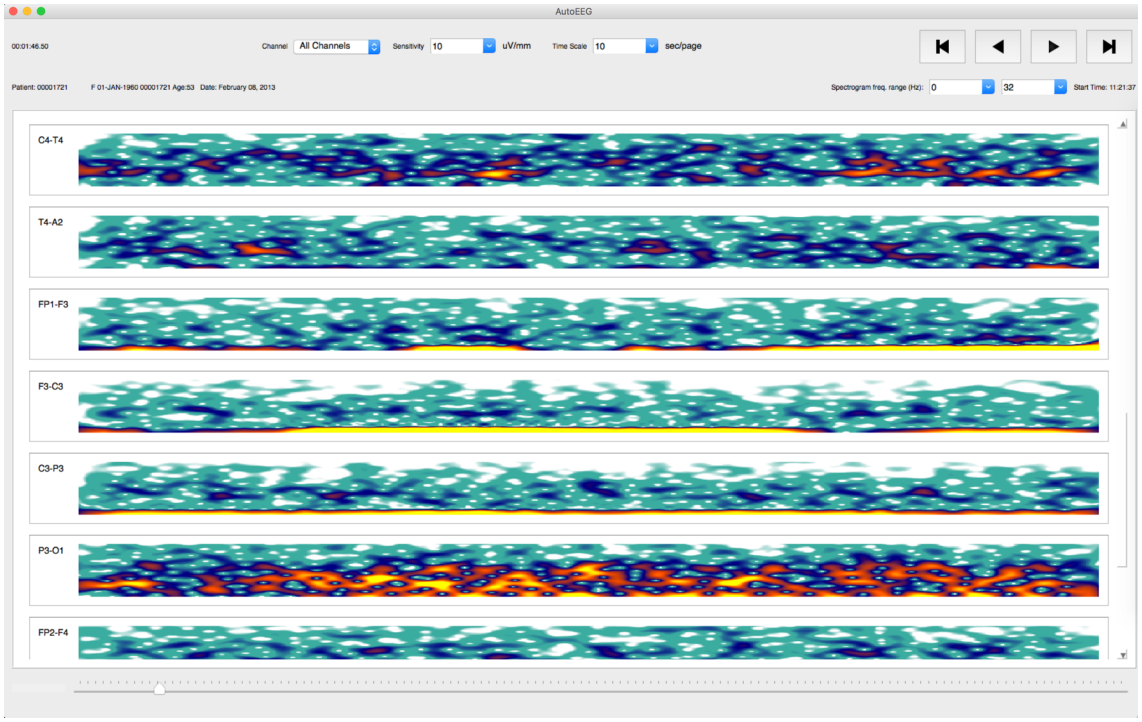


Figure 5. A typical seizure event is shown in a) a waveform plot and b) a spectrogram plot for a subset of the channels. Note that the seizure begins on a few channels (e.g., T5-O1) and then spreads to adjacent channels.

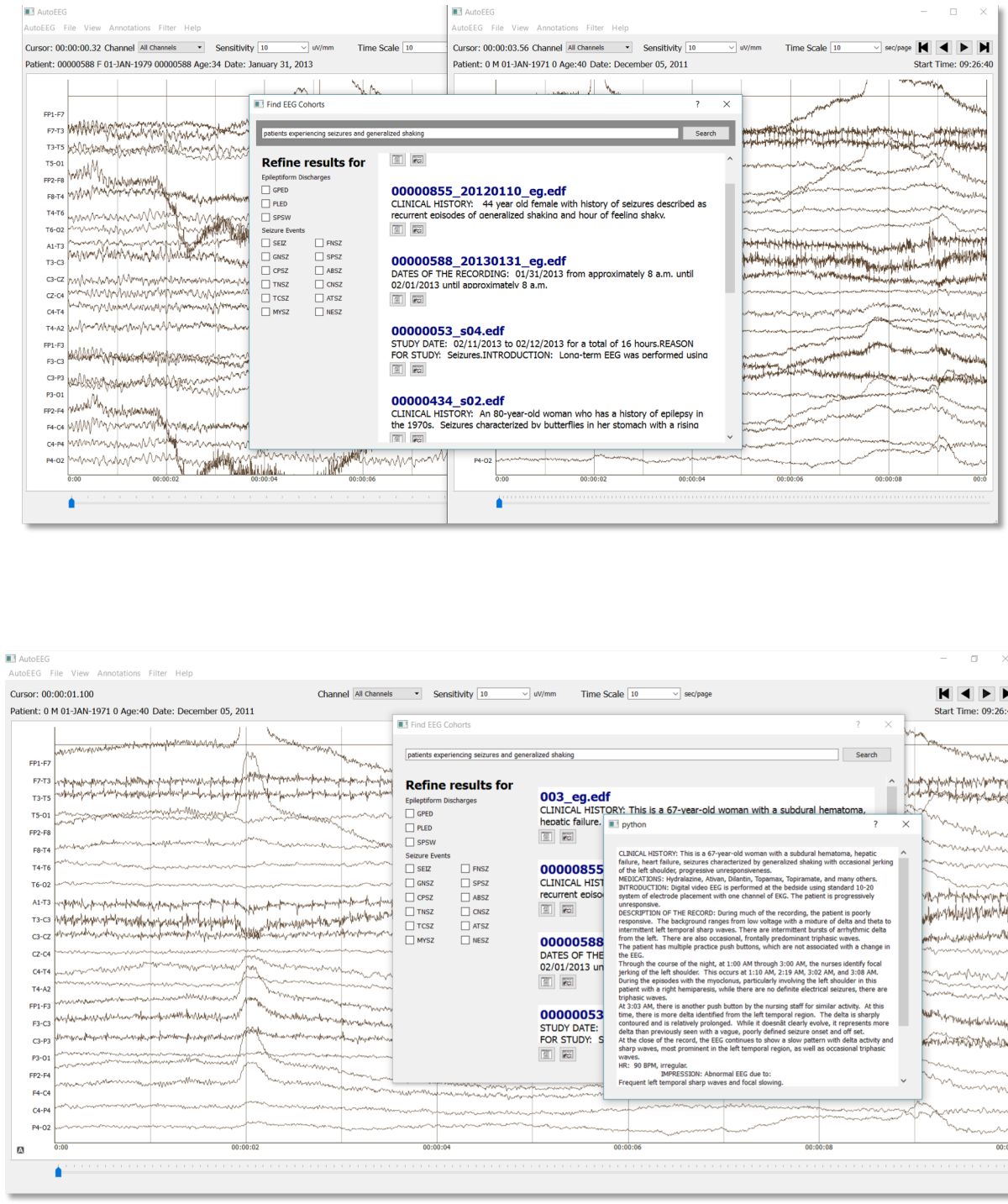


Figure 6. A cohort retrieval system can provide relevant decision support that improves a neurologist's ability to manually interpret an EEG.

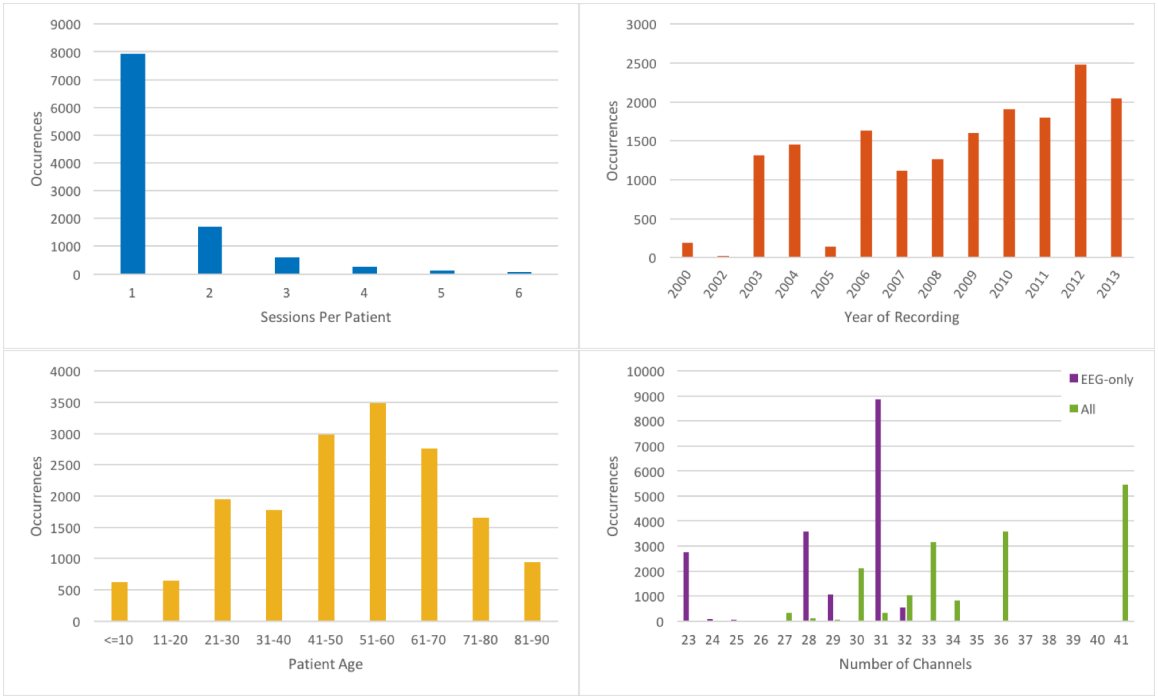


Figure 7. Metrics describing the TUH EEG Corpus: [top left] histogram showing number of sessions per patient; [top right] histogram showing number of sessions recorded per calendar year; [bottom left] histogram of patient ages; [bottom right] histogram showing number of EEG-only channels (purple) and total channels (green).

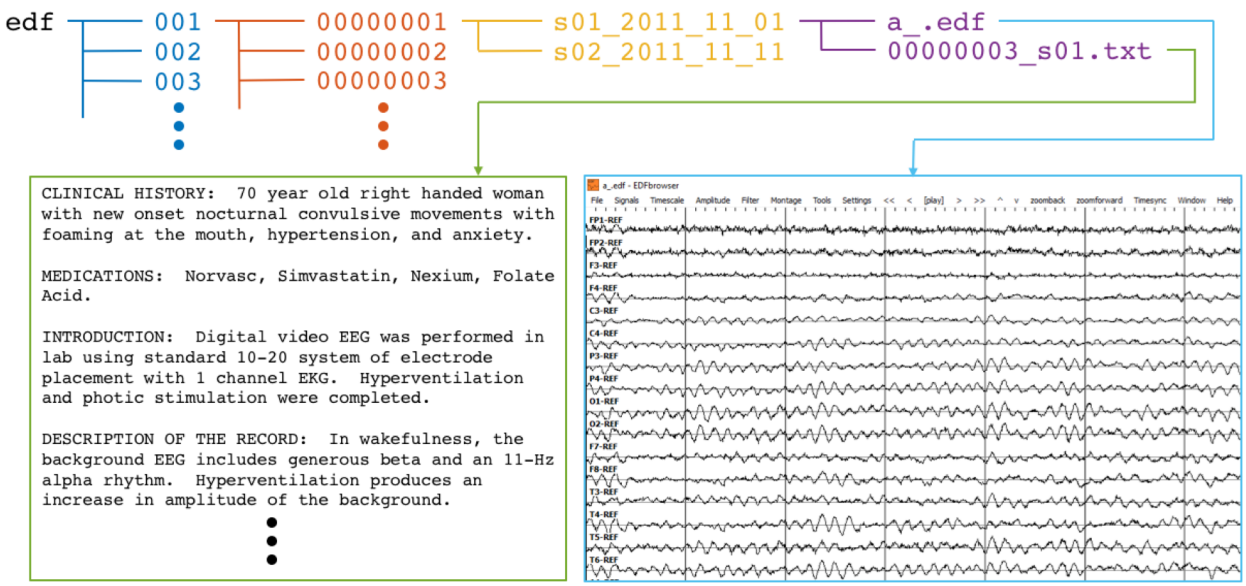


Figure 8. The directory and file structure of TUH-EEG is shown. Data is organized by patient (orange) and then by session (yellow). Each session contains one or more signal files (edf) and a physician's report (txt). To accommodate filesystem management issues, patients are grouped into sets of about 100 (blue).

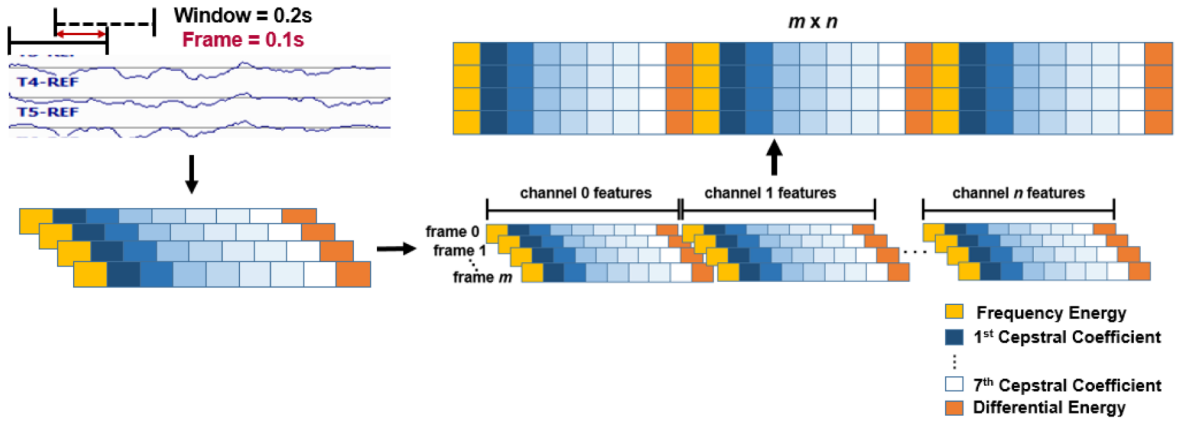


Figure 9. An illustration of the frame-based analysis that is used to extract features, and how features are stacked so that both temporal and spatial context can be exploited.

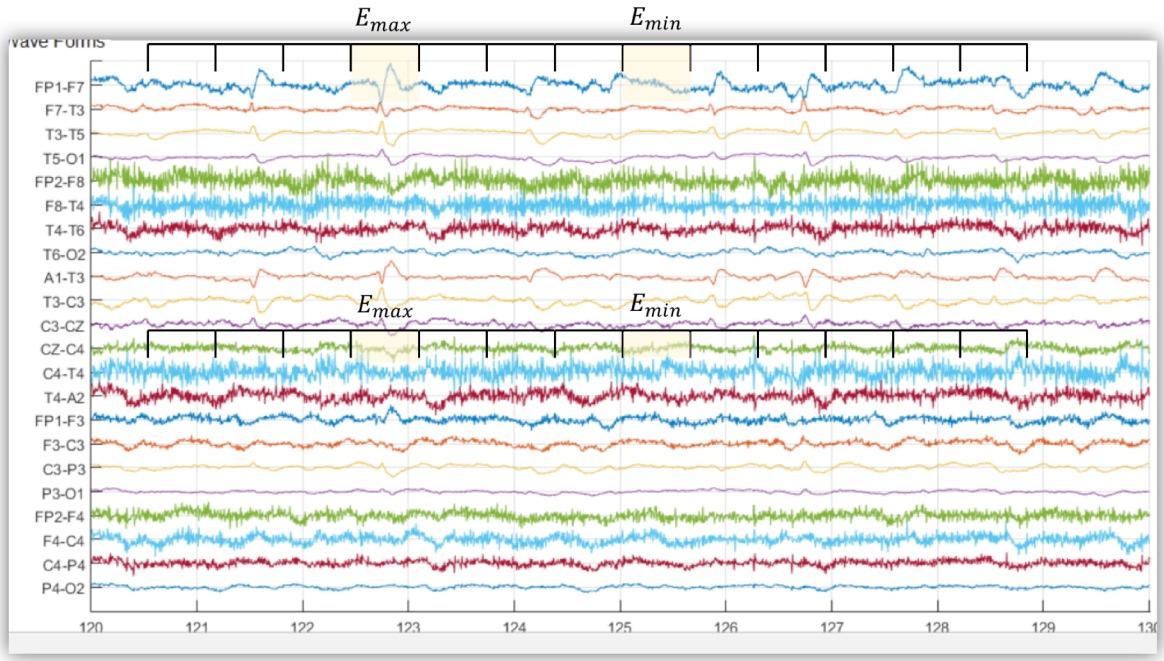


Figure 10. An illustration of how the differential energy term accentuates the differences between spike-like behavior and noise-like behavior. Detection of SPSW events is critical to the success of the overall system.

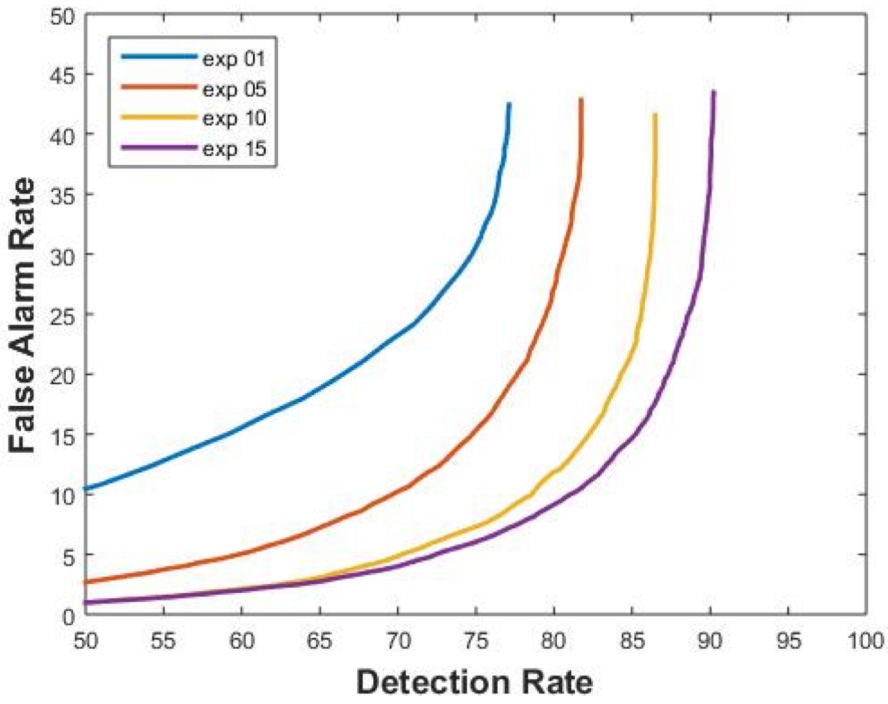


Figure 11. A DET curve analysis of feature extraction systems that compares absolute and differential features. The addition of first derivatives provides a measurable improvment in performance while second derivatives are less beneficial

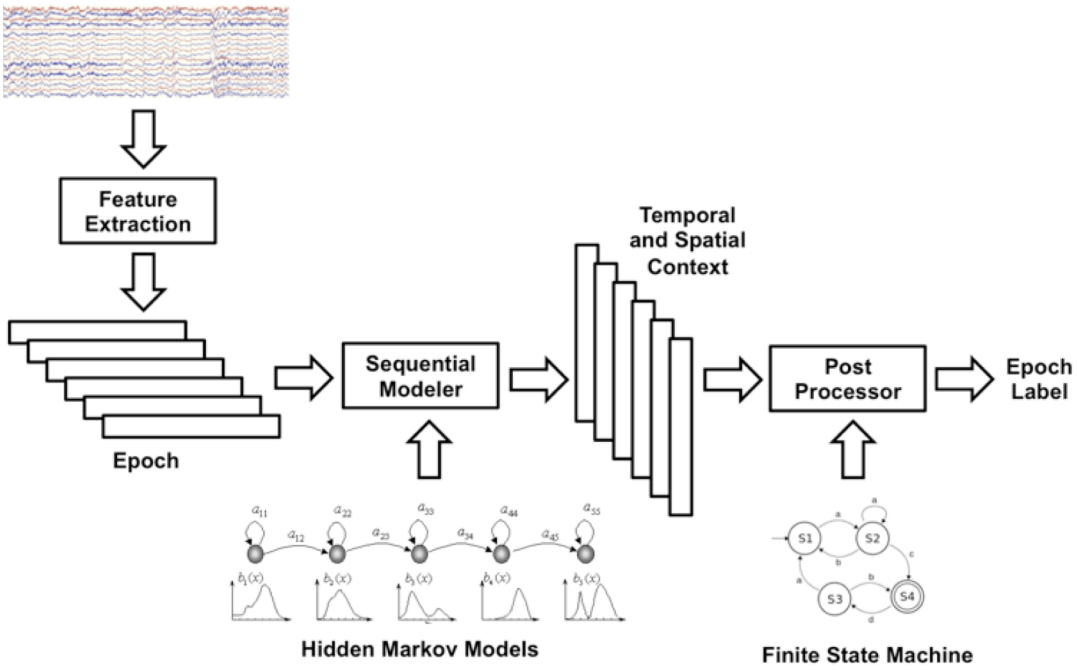


Figure 12. A two-level architecture for automatic interpretation of EEGs that integrates hidden Markov models for sequential decoding of EEG events with deep learning for decision-making based on temporal and spatial context.

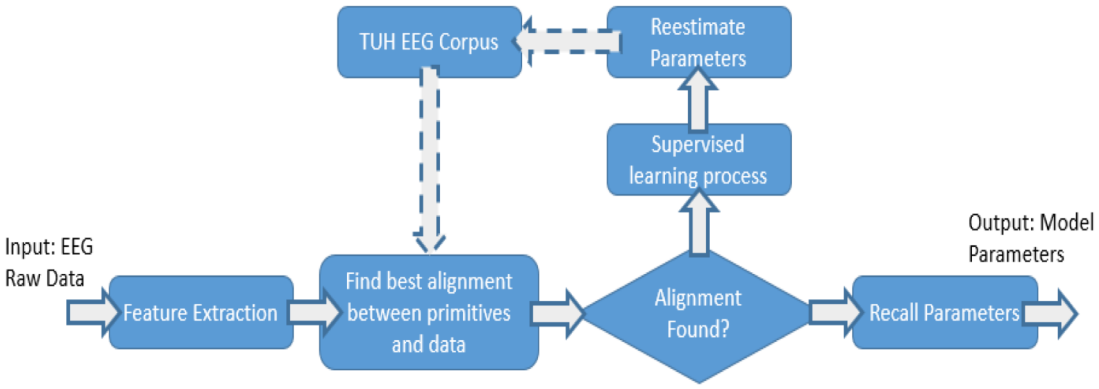


Figure 13. An overview of our iterative HMM training procedure is shown. An active learning approach is used to bootstrap the system to handle large amounts of data.

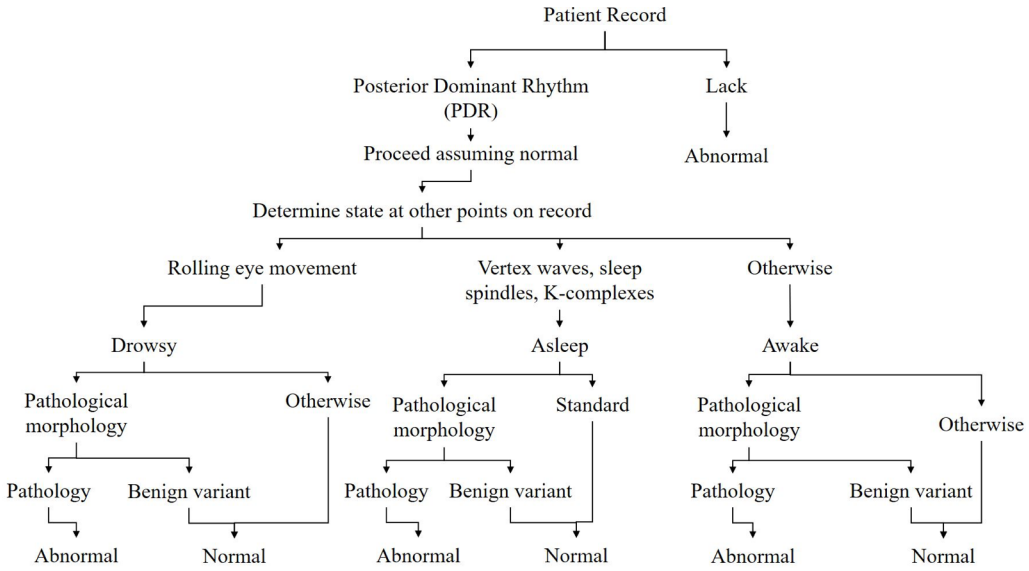


Figure 14. The general process for identifying an abnormal EEG depends heavily on the observation of the Posterior Dominant Rhythm (PDR).

LIST OF TABLES

Table 4. The impact of differential features on performance is shown. For the overall best systems (nos. 10 and 15), second derivatives do not help significantly. Differential energy and derivatives appear to capture similar information.

Field	Description	Example
1	Version Number	0
2	Patient ID	TUH123456789
4	Gender	M
6	Date of Birth	57
8	Firstname_Lastname	TUH123456789
11	Startdate	01-MAY-2010
13	Study Number/ Tech. ID	TUH123456789/TAS X
14	Start Date	01.05.10
15	Start time	11.39.35
16	Number of Bytes in Header	6400
17	Type of Signal	EDF+C
19	Number of Data Records	207
20	Dur. of a Data Record (Secs)	1
21	No. of Signals in a Record	24
27	Signal Prefiltering	HP:1.000 Hz LP:70.0 Hz N:60.0
28	No. Signal Samples/Rec.	250

Table 1. Selected fields from an EDF header that contain de-identified patient information and key signal processing paramters are shown.

Event	Train		Eval	
	No.	% (CDF)	No.	% (CDF)
SPSW	645	0.8% (1%)	567	1.9% (2%)
GPED	6184	7.4% (8%)	1,998	6.8% (9%)
PLED	11,254	13.4% (22%)	4,677	15.9% (25%)
EYEM	1,170	1.4% (23%)	329	1.1% (26%)
ARTF	11,053	13.2% (36%)	2,204	7.5% (33%)
BCKG	53,726	63.9% (100%)	19,646	66.8% (100%)
Total:	84,032	100.0% (100%)	29,421	100.0% (100%)

Table 2. An overview of the distribution of events in the subset of the TUH EEG Corpus used to develop our baseline technology for the six-event classification task.

No.	System Description	Dims.	6-Way	4-Way	2-Way
1	Cepstral	7	59.3%	33.6%	24.6%
2	Cepstral + E_f	8	45.9%	33.0%	24.0%
3	Cepstral + E_t	8	44.9%	33.7%	24.8%
4	Cepstral + E_d	8	55.2%	32.8%	24.3%
5	Cepstral + E_f +E_d	9	39.2%	30.0%	20.4%

Table 3. Performance on the TUH EEG Short Set of the base cepstral features augmented with an energy feature. System no. 5 uses both frequency domain and differential energy features. Note that the results are consistent across all classification schemes.

No.	System Description	Dims.	6-Way	4-Way	2-Way
6	Cepstral + Δ	14	56.6%	32.6%	23.8%
7	Cepstral + E_f + Δ	16	43.7%	30.1%	21.2%
8	Cepstral + E_t + Δ	16	42.8%	31.6%	22.4%
9	Cepstral + E_d + Δ	16	51.6%	30.4%	22.0%
10	Cepstral + E_f +E_d + Δ	18	35.4%	25.8%	16.8%
11	Cepstral + Δ + ΔΔ	21	53.1%	30.4%	21.8%
12	Cepstral + E_f + Δ + ΔΔ	24	39.6%	27.4%	19.2%
13	Cepstral + E_t + Δ + ΔΔ	24	39.8%	29.6%	21.1%
14	Cepstral + E_d + Δ + ΔΔ	24	52.5%	30.1%	22.6%
15	Cepstral + E_f +E_d + Δ + ΔΔ	27	35.5%	25.9%	17.2%
16	(15) but no ΔΔ for E_d	26	35.0%	25.0%	16.6%

Table 4. The impact of differential features on performance is shown. For the overall best systems (nos. 10 and 15), second derivatives do not help significantly. Differential energy and derivatives appear to capture similar information.

System	DET	FA	Error
1: Simple Heuristics	**99%**	**64%**	**74%**
2: Random Forests	**85%**	**6%**	**37%**
3: Autoencoder	**84%**	**4%**	**37%**

Table 5. A comparison of performance for three postprocessing algorithms for the detection rate (DET), false alarm rate (FA) and classification error rate (Error). The FA rate is the most critical to this application.

No.	System Description	Error
1	Random Guessing	49.8%
2	kNN (k = 20)	41.8%
3	RF (N_t = 50)	31.7%
4	PCA-HMM (#GM = 3 #HMM States = 3)	25.6%
5	GMM-HMM (#GM = 3 #HMM States = 3)	17.0%

Table 6. A comparison of performance for abnormal classification of an EEG for several standard classification approaches including k-Nearest Neighbor (kNN), Random Forests (RF) and two variants of a hidden Markov model (HMM) system.

Channel	Error (%)
FP1-F7	19.8%
T5-O1	17.0%
F7-T3	19.8%
C3-Cz	20.7%
P3-O1	23.6%

Table 7. A comparison of performance on an abnormal EEG classification task as a function of the channel selected for analysis. Neuroscience considerations support the use of T5-O1, which delivers the lowest error rate.