THE TUH EEG CORPUS:
A Big Data Resource for Automated EEG Interpretation

A. Harati, S. López, I. Obeid and J. Picone
Neural Engineering Data Consortium
Temple University
Philadelphia, Pennsylvania, USA
{amir.harati, silvia.lopez, obeid, picone}@temple.edu

M. P. Jacobson, M.D. and S. Tobochnik
Department of Neurology
Lewis Katz School of Medicine, Temple University
Philadelphia, Pennsylvania, USA
jacobsm@tuhs.temple.edu

***Abstract*—The Neural Engineering Data Consortium (NEDC) is releasing its first major big data corpus – the Temple University Hospital EEG Corpus. This corpus consists of over 25,000 EEG studies, and includes a neurologist's interpretation of the test, a brief patient medical history and demographic information about the patient such as gender and age. For the first time, there is a sufficient amount of data to support the application of state of the art machine learning algorithms. In this paper, we present pilot results of experiments on the prediction of some basic attributes of an EEG from the raw EEG signal data using a 3,762 session subset of the corpus. Standard machine learning approaches are shown to be capable of predicting commonly occurring events from simple features with high accuracy on closed-loop testing, and can deliver error rates below 50% on a 6-way open set classification problem. This is very promising performance since commercial technology fails on this data.**

I. Introduction

The worldwide EEG market is growing substantially as EEGs are increasingly being used in preventive diagnostic procedures. The worldwide economic burden for brain-related illnesses reached more than $2T/yr. in 2014 ($1T/yr. in the U.S. alone). Diagnosis of a neurological disease like epilepsy is a life-altering event, impacting a person's ability to live a normal and fulfilling life. Hence, it is not surprising that the majority of research has focused on diagnosis of epilepsy and stroke. More recently, however, there has been significant interest in expanding the application of EEGs to conditions such as Middle Cerebral Artery (MCA) Infarct, Posterior Reversible Encephalopathy Syndrome (PRES), Alzheimer's disease and sleep disorders. Discovering the correlates of these conditions in an EEG signal is an open area of research.

An EEG may also be employed to determine the overall electrical activity of the brain, which is used to evaluate trauma, drug intoxication or blood flow during surgical procedures. Trauma represents an important emerging market for EEG technology since there is considerable interest today in diagnosing post-traumatic stress disorder (PTSD) in soldiers and chronic traumatic encephalopathy (CTE) in athletes. CTE in particular might require establishing a baseline EEG prior to one's involvement in contact sports, and monitoring changes over long periods of time. This will exceed the capacity of the healthcare system to manually interpret EEGs, and create additional need for a system capable of high performance automatic interpretation.

The primary output of an EEG test, as shown in Figure 1 is a physician's report. After the recording of a patient's EEG, it can take as long as several weeks for the interpretation of the signal by a certified neurologist. Reducing the lag time between recording and interpretation can positively impact the quality, efficiency (and profitability) of healthcare. As continuous monitoring has become more popular, the ability to analyze signals in real-time and generate alerts has become increasingly necessary. The goal of our research is not to simply generate a report automatically, but to identify the events in the signal that contributed to the diagnosis. Despite a large body of literature on the prediction of a diagnosis, only recently has research on low-level event detection appeared [1].

The ability to automatically predict life-threatening events from EEG signals has been actively researched for the past 40 years. Unfortunately, clinical use of such systems is limited due to poor classification performance. EEG events are defined as critical points in a signal, such as a spike or asymmetric wave shape, that correlate with the presence of a particular disease. Physicians have indicated that a classification error rate of 5% for these EEG events would be acceptable clinical performance. Current state of the art systems do not operate at this level of accuracy due to a lack of adequate machine learning resources and hence are not in widespread use. Therefore, NEDC has been developing a large database of clinical EEGs [2] collected at Temple University Hospital that we believe is crucial for the application of more powerful machine learning approaches.

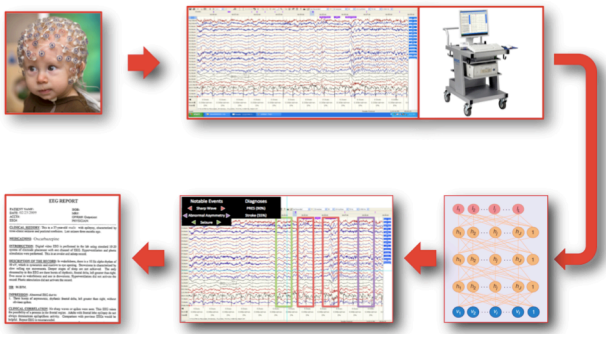


Figure 1. An overview of the EEG interpretation system. The primary outputs of the process are a labeled EEG signal and an EEG report.

II. The TUH EEG Corpus

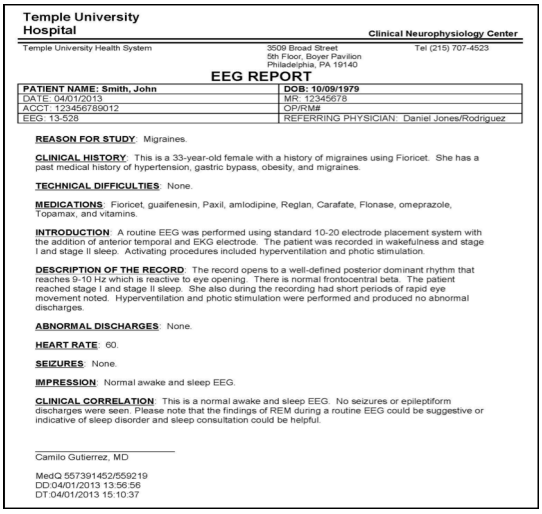
The TUH EEG Corpus is the world's largest publicly available database of clinical EEG data, comprising more than 25,000 EEG records and over 14,000 patients. It represents the collective output from Temple University Hospital's Department of Neurology since 2002. EEG signals were recorded using several generations of Natus Medical Incorporated's NicoletTM EEG recording technology. The raw signals obtained from the studies consist of recordings that vary between 20 and 128 channels sampled at 250 Hz minimum using a 16-bit A/D converter. The data is stored in a proprietary format that has been exported to EDF with the use of NicVue v5.71.4.2530. These EDF files contain a 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 project website [3].

The medical record numbers, names, exact dates of birth and study numbers were redacted in the headers order to ensure the patients' anonymity. However, 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.

For every EEG, there is also a report, shown in Figure 2, 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. 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). The reports have been manually de-identified so that a patient's identity remains anonymous. The reports are provided as flat unstructured text files. More research is needed on the organization and representation of these reports. However, we have been able to automatically pair EEG event classes with sessions using these reports.

Table 1. Selected Fields From an EDF Header.

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



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	CP/RM#
EEG: 13-526	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 2. An example of a physician's EEG Report.

A distribution of the number of records per year is presented in Figure 3. The number of EEGs recorded at TUH has been steadily increasing in recent years, and we hope to continue augmenting the data as it is collected. Approximately 75% of the sessions are standard EEGs less than one hour in duration, while the remaining 25% are from long-term monitoring sessions. There are about six different formats for the reports depending on the year and the type of EEG, though the body of the report contains similar information.

To put the size of this corpus in perspective, the EEG signal data requires about 1.8T of storage with a median file size of 20 Mbytes. The EEG signal data is "pruned" which simply means the EEG technician identified sections of the recording that were of clinical value and discarded the rest. Even so, the amount of data is staggering. For example, if we treat each channel of data as an independent signal, there is over 1B seconds of data. Though this might seem huge at first, the events we are interested in are relatively rare, often occupying less than 1% 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.

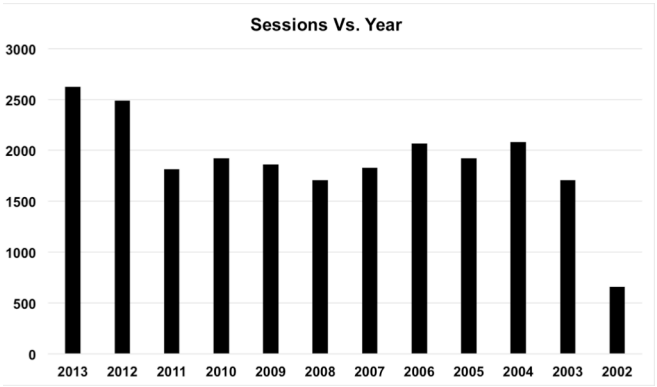


Figure 3. The number of EEG sessions per year.

III. Preliminary Experiments

An important part of our research on this data has been iterating on the definition of EEG events with neurologists. This requires a deeper understanding of how EEGs are manually interpreted and the translation of this process into an algorithm description. After several iterations, we are focusing on a 6-way classification:

(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 Blinks (EYEBL)*: 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 [4] to perform stroke and epilepsy detection. In fact, we have replicated state of the art results on these tasks using the technology described in this paper. 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.

The last three classes are used to improve our background model. 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). 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 follows a very successful approach that we have used in speech recognition [5].

Artifacts and eye blinks occur frequently enough that they merit separate classes. 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. 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.

Feature extraction was performed on the data using a standard filter bank/cepstral coefficient approach [6]. The overall system at present is not extremely sensitive to the core feature set as long as they adequately model the spectral range from approximately 0.5 Hz to 25 Hz. We use an 8-band filter bank analysis that is transformed into a 9-element feature vector that includes 8 cepstral coefficients and energy. The latter is calculated using a frequency domain approach.

After extracting features, we have trained a standard hidden Markov Model (HMM) for each class [7]. HMMs are a class of doubly stochastic processes in which discrete state sequences are modeled as a Markov chain and have been used extensively to model time series data. Expectation-Maximization (EM) algorithm is used to train the models. Figure 4 provides an overview of the training procedure. An active learning approach is used to bootstrap the system from small amounts of data to larger subsets.

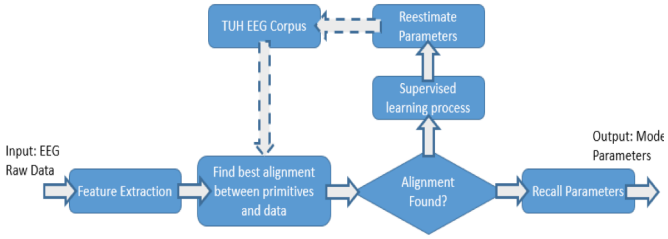


Figure 4. An overview of our iterative HMM training procedure is shown. An active learning approach is used to bootstrap the system to handle large amounts of data.

It should also be noted that data preparation is a large part of the challenge in processing this clinical data. This involves clustering files into the appropriate classes based on information automatically extracted from a physician's report. We initially trained our system in a completely unsupervised manner using an active learning approach. We then had a small amount of data manually labeled by an expert. We carefully selected 100 10-second epochs that contained ample examples of the SPSW class along with a few GPED and PLED examples. We used this data to guide the training process. We are in the process of annotating an additional 70 10-second epochs to serve as a held-out evaluation set.

Toward this goal we have also developed a software tool to facilitate labeling of the data and review of the machine learning output. We refer to this tool as the EEG Annotator. It is written in Python based on the PyQt toolkit, and accepts EDF files as input. Annotations are stored in a separate text file to facilitate interfacing to other software tools. A screenshot of the program can be seen in Figure 5.

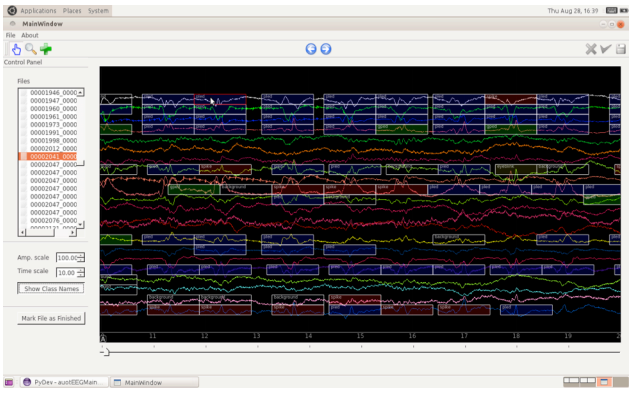


Figure 5. A screeenshot of EDF Annotator is shown. The colors depict different classes (e.g., red is SPSW).

Note that in these preliminary experiments, each channel is processed independently. Some events are easily identified by looking at behaviors across channels. However, in this study, we have restricted ourselves to analyzing each channel independently. We are in the process of developing a second machine learning module that will post process hypotheses from all channels and classify sessions accordingly. This module will substantially reduce the false alarm rate [1].

In TABLE 2 we present a confusion matrix for the HMM-based system on the evaluation data. We observe the correct recognitions for the three primary event classes (SPSW, PLED, and GPED) are above 40% though misrecognitions are also about 40%. Fortunately, to be relevant for clinical use it is not necessary to detect every spike correctly. It is adequate to detect enough of these spikes that a neurologist can quickly develop and impression. Of greater concern is a high false alarm rate which results in a need to review too much erroneous data.

To put these results in perspective, we have also compared our system to Wulsin's approach using Deep Belief Networks [4]. Wulsin et al. used a small proprietary database of clinical EEGs that was manually transcribed by two neurologists. Both studies used very similar classes though the methodologies for modeling non-spike portions to the signals were somewhat different. In their study, the performance metric used was the F_{score}, the harmonic mean of sensitivity and specificity:

$$F_{score} = 2\frac{sensitivity \bullet specificity}{sensitivity + specificity} \quad (1)$$

The highest F_{score} reported in [4] is 0.476 while our system produced an F_{score} of 0.702 for the evaluation data and 0.772 for the training data on the TUH EEG Corpus.

Figure 6 shows a tradeoff between false alarms and detections (correct recognitions). We can change the operating point along this curve by simply modifying a threshold on the likelihood values of the most likely class. Clinicians prefer a low false alarm rate since that reduces the amount of time spent reviewing the data and can increase their productivity. Based on discussions with Temple University Hospital neurologists, our target for the detection rate on the three primary event classes is 95% and our target for the false alarm rate is 5%.

Note that these results were generated using closed-set testing since we have a very limited amount of transcribed data. Informal analyses of open-set results at this level of performance indicate they are comparable. Once the held-out data is transcribed we will be in a better position to do open-set testing. It is encouraging that confusions between the three primary event classes and the three background event classes are relatively small.

TABLE 2. EXPERIMENTAL RESULTS

	SPSW	PLED	GPED	ARTF	EYBL	BCKG
SPSW	38%	8%	27%	9%	11%	7%
PLED	19%	54%	19%	1%	4%	2%
GPED	12%	20%	42%	14%	2%	9%
ARTF	6%	3%	4%	39%	2%	47%
EYBL	3%	9%	1%	1%	84%	1%
BCKG	9%	2%	8%	6%	3%	72%

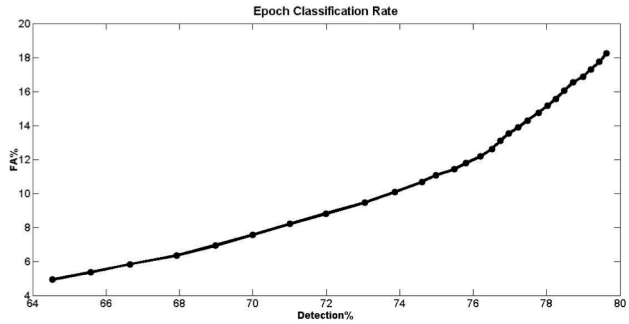


Figure 6. The tradeoff between false alarms and correct detections is depicted for a baseline system on the TUH EEEG Corpus. A system with a low false alarm rate is much preferred over a high detection rate.

IV. SUMMARY

In this paper, we have introduced the TUH EEG Corpus, which consists of over 25,000 EEG sessions from over 14,000 patients. The data spans over a decade of clinical studies, and includes a rich library of metadata, patient histories and physician's interpretations. It is ideal for large-scale machine learning experiments. We expect it will have a major impact on the development of clinical tools to automatically interpret EEGs. We expect the complete corpus to be released by Spring 2015, and will be available at *www.nedcdata.org*. Preliminary releases have been available from the project web site since February 2014. Feedback on the data is encouraged.

Preliminary results presented on a pilot corpus of 3,762 sessions demonstrated that it is possible to predict some annotations directly from the data using unsupervised and partially-supervised learning techniques. Our HMM-based system was able to successfully detect SPSW, GPED and PLED events. A detection rate of 76% was achieved with a false alarm rate of 12%. The background and artifact modeling performed well. Event detection overall was promising given that only single channel processing was performed.

We expect to continue improving performance by incorporating a postprocessing scheme that pools outputs across channels and integrates features that incorporate both short-term and long-term spectral information. Increasing the temporal context and correlating channel identities with events will also greatly improve our ability to differentiate SPSW events from GPED and PLEDs. Additional levels of postprocessing will be used to determine epoch labels.

ACKNOWLEDGEMENTS

Portions of this work were sponsored by the Defense Advanced Research Projects Agency (DARPA) MTO under the auspices of Dr. Doug Weber through the Contract No. D13AP00065, Temple University's College of Engineering and Office of the Senior Vice-Provost for Research, and the National Science Foundation through Major Research Instrumentation Grant No. CNS-09-58854 and Grant No. CNS-1305190. 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.

REFERENCES

[1] D. Wulsin, *Bayesian Nonparametric Modeling of Epileptic Events*, University of Pennsylvania, 2013.

[2] A. Harati, S. I. Choi, M. Tabrizi, I. Obeid, M. Jacobson, and J. Picone, "The Temple University Hospital EEG Corpus," in *Proc. of the IEEE Global Conf. on Signal and Information Processing*, 2013, pp. 29–32.

[3] S. I. Choi, I. Obeid, M. Jacobson, and J. Picone, "The Temple University Hospital EEG Corpus," The Neural Engineering Data Consortium, College of Eng., Temple Univ., 2013. [Online]. Available: *http://www.isip.piconepress.com/projects/tuh_eeg*. [Accessed: 06-Jan-2013].

[4] D. Wulsin, J. Blanco, R. Mani, and B. Litt, "Semi-Supervised Anomaly Detection for EEG Waveforms Using Deep Belief Nets," in *International Conference on Machine Learning and Applications (ICMLA)*, 2010, pp. 436–441.

[5] J. Picone, "Continuous speech recognition using hidden Markov models," *IEEE ASSP Mag.*, vol. 7, no. 3, pp. 26–41, Jul. 1990.

[6] M. Brookes, "Voicebox: Speech processing toolbox for matlab," *Dept. of Electrical & Electronic Engineering, Imperial College*, 1997. [Online]. Available: Software, available [Mar. 2011] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html. [Accessed: 28-Aug-2014].

[7] L. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286, 1989.