The Temple University Hospital EEG Corpus

A. Harati, S. Choi, M. Tabrizi, I. Obeid and J. Picone
Neural Engineering Data Consortium
Temple University
Philadelphia, Pennsylvania, USA
{amir.harati, sung.choi, masih, obeid, picone}@temple.edu

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

***Abstract*—The recently established Neural Engineering Data Consortium (NEDC) is in the process of developing its first large-scale corpus. This corpus, known as the Temple University Hospital EEG Corpus, upon completion, will total over 20,000 EEG studies, and include patient information, medical histories and physician assessments, making it the largest and most comprehensive publicly released EEG corpus. For the first time, there will be sufficient data to support the application of state of the art machine learning algorithms. In this paper, we present pilot results of experiments in which we attempted to predict some basic attributes of an EEG from the raw EEG data using a pilot database of 100 EEGs. 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 slightly below 50% on a 12-way open set classification problem.**

I. INTRODUCTION

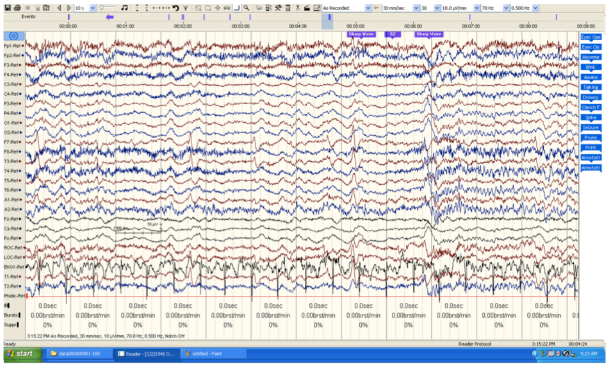
Data-driven approaches have made enormous advances in recent years [1]-[3] in terms of their ability to predict events through supervised training on big data resources. Equally important, however, is the fact that many of these techniques have the ability to discover underlying structure of the data using latent variables and unsupervised training techniques. These types of algorithms can provide enormous insight into the data. The only impediment to applying these techniques has been the lack of a suitable amount of data to support comprehensive experimentation. The Neural Engineering Data Consortium (NEDC) is being established at Temple University to address this issue.

The past two decades have seen an explosion in Brain Computer Interface (BCI) research. However, despite significant progress, overall progress in the field does not appear to have been commensurate with the scope of investment (over $200M in the last decade from NIH and NSF alone). In particular, efforts to commercialize research findings have been tepid, hampered by a general lack of robustness when translating technologies to uncontrolled environments beyond the research laboratory. NEDC [4] is being launched to focus the attention of the research community on a progression of neural engineering research questions and to generate and curate massive data sets to be used in addressing those questions. A community-wide assessment, funded by a planning grant from the National Science Foundation, is being conducted to better define and prioritize the required resources needed by researchers to fuel innovation. These activities will be discussed extensively at GlobalSIP 2013.

The existence of massive corpora has proven to substantially accelerate research progress by eliminating unsubstantiated research claims [5]. NEDC will broaden participation by making data available to research groups who have significant signal processing expertise but who lack capacity for data generation. This effort is modeled in part after similar successful endeavors, particularly in the human language technology field where a data consortium has led to systematic research and technology advances over a 20-year span [6].

In this paper, we present some preliminary results on NEDC's first corpus – clinical electroencephalogram (EEG) recordings, as shown in Figure 1, conducted at Temple University Hospital (TUH) from 2002 to 2013 (and beyond). This corpus will support the development of technology to automatically interpret EEGs in addition to advancing the basic science of what aspects of a patient's medical record correlate with various pathologies that can be diagnosed from EEG studies.

Automatic interpretation of EEG data using machine learning approaches has evolved in recent years. Several specific applications have been studied extensively, including seizure detection [7], movement [8] and brain activity [9]. What most of these studies have in common, however, is that the data sets are small, typically involving 100 or less EEG studies. Such small studies simply do not produce statistically significant outcomes, and do not represent enough data to support complex statistical models such as k-nearest neighbors (kNN) [10], neural networks (NN) [2] or random forests (RF) [3]. Generalization of the findings presented in



3:15:22 PM As Recorded, 30 mm/sec, 10 µV/mm, 70.0 Hz, 0.500 Hz, Notch Off

Figure 1. The source data, which consists of 24-channel recordings plus annotations, is displayed using Natus Medical Incorporated's Nicolet^{TM} NicVue v5.71.4.2530).

such small studies is not possible. Further, when correlates such as drug treatments, patient medical histories, or patient gender or age are factored in, studies consisting of 100 subjects are not sufficient to draw conclusions about best practices. It is the goal of this project to fundamentally change this.

II. The TUH EEG Corpus

The Temple University Hospital EEG Corpus (TUH-EEG) will be the world's largest publicly available database of clinical EEG data. The database, upon completion, will comprise over 20,000 clinical EEG records made at TUH dating back to 2002. We expect this to be an on-going project with annual updates to the corpus. Although information disclosing a patient's identity, such as name and corresponding video are being carefully redacted, other information such as gender, age, relevant medical history, and medications will be retained. In this manner, it will, for example, be possible to mine the data set for statistically significant changes in EEG activity in response to various medications. The complete corpus is expected to be available by the end of 2013 (to coincide with GlobalSIP 2013).

The raw signal data consists of recordings containing between 24 and 36 channels sampled at a minimum of 250 Hz using a 16-bit A/D converter. More information on the fundamentals of EEG recordings can be found here [11]. TUH has been using a Natus Medical Incorporated's Nicolet^{TM} EEG recording facility for the majority of the data collected. The raw data files are stored in a proprietary format. The data files are being exported using NicVue v5.71.4.2530 from their proprietary format to an EDF+ format [12]. This file format consists of a machine-readable header containing metadata about the study, and the binary signal data. The EDF+ header contains 24 unique fields in addition to the actual signal data. Selected fields from this header that contain important metadata are shown below in Table 1. The redacting process involves modifying the patient ID (Field 2), date of birth (F6), patient name (F8) and the study number (F13) so that the patient's identity remains anonymous. There are additional fields that describe signal conditions, such as the maximum amplitudes of the signals, which are stored for every channel. A complete description of the header and its contents can be found at the project web site [13].

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[1] Prefiltering	HP:1.000 Hz LP:70.0 Hz N:60.0
28	Signal[1] No. Samples/Rec.	250

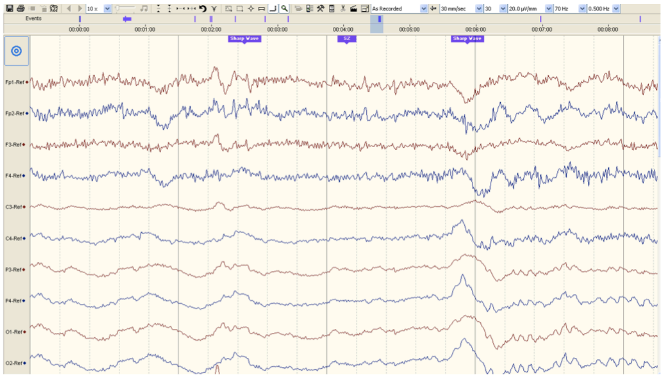
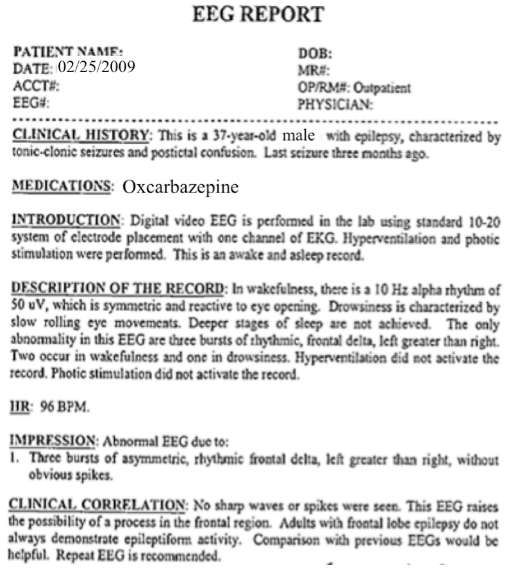


Figure 3. The EEG data contains detailed annotations including markers, shown in blue above, synchronized with the waveforms that identify critical events.

The EDF+ file also contains time-aligned transcriptions of the data. An example of this is shown in Figure 3. The blue markers shown along the top of the waveform represent event annotations provided by technicians and physicians. This type of information will presumably be useful in localizing points of interest in the waveforms, and in developing hierarchical classification models that use the symbolic representation of these markers as internal states.

In addition to the signal data, for each EEG, a physician's EEG Report is available. An example of such a report is shown in Figure 2. This report contains a summary of the patient's clinical history and medications. It also includes two fields, Impression and Clinical Correlation, which contain the physician's findings. This report information is available in an Excel spreadsheet in a name/value pair format. The more recently collected EEGs (since 2011) are also coded in International Statistical Classification of Diseases codes (ICD-9). These codes can form the basis for the classification labels used in machine learning experiments.



EEG REPORT

PATIENT NAME:	DOB:
DATE: 02/25/2009	MR#:
ACCT#:	OP/RM#: Outpatient
EEG#:	PHYSICIAN:

CLINICAL HISTORY: This is a 37-year-old male with epilepsy, characterized by tonic-clonic seizures and postictal confusion. Last seizure three months ago.

MEDICATIONS: Oxcarbazepine

INTRODUCTION: Digital video EEG is performed in the lab using standard 10-20 system of electrode placement with one channel of EKG. Hyperventilation and photic stimulation were performed. This is an awake and asleep record.

DESCRIPTION OF THE RECORD: In wakefulness, there is a 10 Hz alpha rhythm of 50 uV, which is symmetric and reactive to eye opening. Drowsiness is characterized by slow rolling eye movements. Deeper stages of sleep are not achieved. The only abnormality in this EEG are three bursts of rhythmic, frontal delta, left greater than right. Two occur in wakefulness and one in drowsiness. Hyperventilation did not activate the record. Photic stimulation did not activate the record.

HR: 96 BPM.

IMPRESSION: Abnormal EEG due to:

1. Three bursts of asymmetric, rhythmic frontal delta, left greater than right, without obvious spikes.

CLINICAL CORRELATION: No sharp waves or spikes were seen. This EEG raises the possibility of a process in the frontal region. Adults with frontal lobe epilepsy do not always demonstrate epileptiform activity. Comparison with previous EEGs would be helpful. Repeat EEG is recommended.

Figure 2. An example of a physician's EEG Report.

III. Pilot Experiments

To better understand the nature of our data and the challenges processing it, we selected 140 studies and formed a pilot corpus. We first analyzed the metadata to understand something about the nature of the patients. A summary of key metadata fields is shown in Table 2. The corpus is fairly evenly divided between males and females (46% male / 54% female) and contains a good distribution of ages ranging from 20 years old to 95 years old. There is a total of over 42 hours of signal data, or an average of 17 minutes per study. The number of channels, including an annotation channel, varies from 28 to 37 with the single largest group having been recorded using 31 channels (48 studies). Prefiltering was turned off. The sample rate was 250 Hz for a majority of the studies.

Table 2. An Analysis of Metadata in the Pilot Corpus.

Field	Description	Example
3	Gender	M (46%) F (54%)
4	Age (Derived from DOB)	Min (20) Max (94) Avg (53) Stdev (19)
13,14	Duration	42 hours (17 mins./study)
15	Number of Channels	28 (26) 29 (12) 31 (48) 32 (19) 35 (5) 37 (33)
23	Prefiltering	HP:0.000 Hz LP:0.0 Hz N:0.0
24	Sample Frequency	250 Hz (100) 256 Hz (43)

A technician annotated each EEG at the time of recording. A listing of some of the most common markers is shown in Table 4. Though these appear as time-aligned markers in the data, technicians who enter these markers do not pay close attention to the time alignment. Hence, their location in time is only approximate but still informative.

Table 4. The Most Common Annotation Markers

Marker	Frequency
Eyes Open	515
Eyes Closed	355
Movement	240
Swallow	98
Awake	61
Drowsy / Sleeping	49
Hyperventilation	40
Talking	21

We collapsed these into 12 categories to faciliate some simple machine learning pilot experiments. The categories are shown in Table 3. We selected these categories by clustering the specific markers into more general representations so that there would be an adequate amount of examples for training that corresponded to each marker. We removed some anomalous files and ended up with 103 files that had at least one instance of one of these 12 markers. Each study, or file, can contain one or more of these markers. Note that only 2 of the 103 files contained markers denoting Seizure during a session. Events such as seizures during a session are exceedingly rare.

Table 3. Clustered Marker Categories

Numeric Label	Name
1	Hyperventilation
2	Movement
3	Sleeping
4	Cough
5	Drowsy
6	Talking
7	Chew
8	Seizure
9	Swallow
10	Spike
11	Dizzy
12	Twitch

We constructed a simple machine learning experiment based on these markers. We first converted each *N*-channel EEG signal to a single feature vector by computing some popular aggregate features: the signal mean, variance and peak value. We computed these features on the first 16 channels because these channels were common to all EEGs in the corpus. We concatenated each vector into one supervector for each file that had a dimension of 48 (3 features per channel x 16 channels). Though more sophisticated feature extraction algorithms will be used in future studies, for these pilot experiments we were mainly interested in establishing the consistency of the data – understanding to what extent it would support machine learning experiments.

These vectors, along with the associated numeric class labels, were applied to three standard machine learning algorithms available in MATLAB. We selected these algorithms based on our familiary with their performance on a wide range of tasks [14]. The three algorithms employed were: (1) a *K*-nearest neighbor (kNN) based on *knnsearch*; (2) a neural network (NN) algorithm based on the "newff" function that uses a single hidden layer; and (3) a random forest (RF) algorithm based on *TreeBagger*. The kNN algorithm required specifying the value of the number of nearest neighbors (*K*). The NN algorithm used tangent sigmoid transfer functions between each layer and resilient back propagation to train the network. The number of neurons (*N*) must be specified. The RF algorithm simply required providing the number of trees (*T*) used in the ensemble.

We used a "leave-one-out" cross-validation approach to conducting the evalutation. Given *N* tokens, *N-1* of these tokens are used for training, and then the classifier is used to predict the class for the N^{th} token. The data set is cycled so that all *N* tokens appear exactly once as the evaluation token. The overall error rate is then computed as the average of the *N* experiments with a single token. This is a very effective technique when faced with a limited amount of data.

A summary of the results is shown in Table 5. The first column lists the algorithm; the second column shows the setting for the relevant design parameter; and the remaining columns show the error rate as a percentage. A forced choice scheme was used (one of 12 labels must be output). Results are given for both closed-set testing, where we train and evaluate on the same tokens, and open-set testing, where we use the leave-one-out method previously described. Results were also given for the raw features, and a post-processed version of these feaures in which each vector is normalized to have a norm of *1*. Performance on closed-set testing is informative because the behavior is generally stable and consistent. Performance on open-set testing evaluates the generalization capability of the classifier. When performance diverges

TABLE 5. ERROR RATES AS A FUNCTION OF THE CLASSIFIER

Alg.	Setting	Closed		Open	
		Raw	Norm	Raw	Norm
kNN	K=1	0.0%	61.5%	72.1%	62.5%
kNN	3	27.9%	61.5%	63.5%	49.0%
kNN	5	39.4%	61.5%	64.4%	69.2%
NN	N=5	49.0%	70.2%	51.9%	75.0%
NN	10	50.0%	71.2%	51.9%	77.9%
NN	15	49.0%	78.9%	50.0%	76.0%
NN	20	51.9%	76.9%	55.8%	78.9%
RF	T=1	19.2%	54.8%	62.5%	60.6%
RF	20	0.0%	49.0%	62.5%	57.7%
RF	50	0.0%	56.7%	61.5%	55.8%
RF	100	0.0%	50.0%	65.4%	54.8%

between closed and open set testing, it is a sign that the data set might have problems or be too small to support these types of experiments.

The overall best performance on open-set testing is achieved by the kNN algoithm for *K=3*. However, closed-set performance is a little unstable since we would expect performance to improve as a function of *K*. Not surprisingly, RF achieves an error rate of 0% on closed-set testing. The RF algorithm is very good at learning the nuances of a specific data set. However, its ability to generalize on the open-set test is not optimal, since this error rate increases to 60%. The performance for NN is very stable across a range of conditions. However, like RF, there is a tendency to overtrain and hence peformance on open-set testing is not optimal.

IV. SUMMARY

In this paper, we have introduced the TUH-EEG Corpus, which upon completion will consist of over 20,000 EEGs, making it the largest publicly available EEG corpus. We expect it will have a major impact on the development of clinical tools to automatically interpret EEGs. 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.

Preliminary results presented on a pilot corpus of over 100 EEGs demonstrated that it is possible to predict some annotations directly from the data. We showed that a kNN-based predictor trained using a leave-one-out cross-validation approach achieved a closed-loop error of 0% and an open-loop error of 49%. An analysis of the actual errors was most informative. Not suprisingly, all algorithms tend to hypothesize the most frequently occurring classes – in this case the first two classes in TABLE 3. This is perhaps the greatest challenge in these types of bioengineering corpora. For example, consider the problem of predicting that a patient has a seizure disorder. In these 103 studies, the seizure marker only appeared twice. Needless to say it was incorrectly classified because in a Bayesian sense, the prior probabilities are so low that it makes sense to simply always choose a non-seizure classification. This is a good example of why it is so important for the TUH EEG Corpus to be large.

It is also clear from our pilot experiments that features are very important. The coarse aggregate feaures used here are not designed to detect events that manifest themselves by short temporal bursts. Sequential decoding of the EEG using contemporary technology such as hidden Markov models will be crucial to identification and classification of such events. This will be the topic of future studies, and we expect to present prelminary results at the symposium. Self-organizing systems that can learn internal structure will be important.

ACKNOWLEDGEMENTS

This research was supported in part by the Defense Advanced Research Projects Agency (Grant No. 12345678), 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.

We would also like to thank the staff of the Department of Neurology, particularly Adam Bolden and Steven D. Tobochnik, for their generous assistance in this project.

REFERENCES

[1] I. Arel, D. C. Rose, and T. P. Karnowski, "Deep Machine Learning - A New Frontier in Artificial Intelligence Research [Research Frontier]," *IEEE Computational Intelligence Magazine*, vol. 5, no. 4, pp. 13–18, 2010.

[2] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohammed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 83–97, Nov. 2012.

[3] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] I. Obeid and J. Picone, "The Neural Engineering Data Consortium (NEDC)," *College of Engineering, Temple University*, 2012. [Online]. Available: http://www.temple.edu/engineering/nedc. [Accessed: 06-Jan-2013].

[5] "The History of Automatic Speech Recognition Evaluations at NIST," *NIST*, 2009. [Online]. Available: *http://www.itl.nist.gov/iad/mig/publications/ASRhistory/index.html*. [Accessed: 03-Feb-2013].

[6] C. Cieri, "20 Years of Progress in Human Language Technology," in *Proceedings of the IEEE Global Conference on Signal and Information Processing*, Austin. Texas, USA, 2013.

[7] W. Zhou, Y. Liu, Q. Yuan, and X. Li, "Epileptic Seizure Detection Using Lacunarity and Bayesian Linear Discriminant Analysis in Intracranial EEG," *IEEE Transactions on Biomedical Engineering,* vol. PP, no. 99. pp. 1–7, 2013.

[8] P. Ahmadian, S. Cagnoni, and L. Ascari, "How capable is non-invasive EEG data of predicting the next movement? A mini review.," *Frontiers in human neuroscience*, vol. 7, p. 124, Jan. 2013.

[9] P. L. Purdon, et al., "Electroencephalogram signatures of loss and recovery of consciousness from propofol," *Proceedings of the National Academy of Sciences*, vol. 110, no. 12, pp. E1142–E1151, Mar. 2013.

[10] Mathworks, "Knnsearch," *Mathworks Documentation Center*, 2013. [Online]. Available: http://www.mathworks.com/help/stats/knnsearch.html?searchHighlight=knnsearch.

[11] E. Niedermeyer and F. H. L. da Silva, *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Philadelphia, Pennsylvania, USA: Lippincott Williams & Wilkins, 2005, p. 1309.

[12] R. Kemp, "European Data Format," *Department of Neurology, Leiden University Medical Centre, The Netherlands*, 2013. [Online]. Available: http://www.edfplus.info. [Accessed: 06-Jan-2013].

[13] 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].

[14] J. Steinberg, "A Comparative Analysis of Bayesian Nonparametric Inference Algorithms for Acoustic Modeling in Speech Recognition," Temple University, 2013.