The Stanford EEG Corpus:
A Large Open Dataset of Electroencephalograms from Children
And Adults to Support Machine Learning Technology

J. Kuo and C. Lee-Messer

Department of Neurology, Stanford University School of Medicine, Stanford, CA 94305
{jkuo3,cleemess}@stanford.edu

Introduction: The electroencephalogram (EEG) is a tool for diagnosing seizures and assessing brain electrical activity in physiological and pathological states. It forms the basis for brain-computer interfaces and studies of the basic science of brain function. Clinically, the current gold standard for analyzing EEG is visual inspection. Unfortunately, trained EEG readers are a limited resource and the process of reading EEGs is labor-intensive. This limits access to proper seizure monitoring and prevents advances in treatment.

Progress in automating EEG analysis has been slow relative to the advances in the last 10 years in visual object recognition and speech recognition using state of the art machine learning techniques. In these fields, a crucial step has been the appearance of high quality, large, open datasets, which allowed for comparison of algorithms from different groups against a common standard.

Our project aims to produce such a dataset based upon clinical recordings at Stanford University Hospital and Lucile Packard Children's Hospital in California.

Methods: We obtained Institutional Research Board approval for producing de-identified EEG files, including clinical annotations. These were compared with clinical EEG reports, obtained via chart review of the electronic medical record. Waveform de-identification was done via a scripted pipeline, translating the original proprietary Nihon-Kohden format into a hierarchical data format (HDF5) based form with standardized dates, while assigning individuals codes to replace names. EEG reports were analyzed via a naïve Bayesian classifier (built with NLTK and spaCy) in order to encode the report results.

Results: Over 136,000 de-identified EEGs resulted representing over 8 years of archived EEGs from over 12000 individuals and 16000 studies in 4.1 TB of compressed data (HDF5 supports built-in compression). The records include routine, ambulatory, long-term and intracranial EEGs, including 4600 studies of children and 375 neonates. 49.1% of the studies were from female patients. Our current text-processing pipeline yields automatic identification of reports with seizures at an 87.5% sensitivity and 95.1% specificity. The baseline rate of seizures in the studies is around 10% in our dataset.

Conclusions: We have produced a substantial new corpus of EEG for analysis by the scientific community. The objectives of our project closely mirror the goals of the Temple University Hospital EEG Corpus[1] and we view our effort as being complementary, with a different mix of patient ages and study types. We have begun sharing the dataset and look forward to collaborating with our colleagues in epilepsy, engineering and computer science to further enhance its value. Given the large size of the dataset, one of our next steps is looking at practical ways to distribute the data. Future plans include: (1) annotation of the data to better codify the EEG reports; (2) application of unsupervised and semi-supervised methods for improved waveform classification; and (3) annotation of the waveforms to identify seizures and other features in order to facilitate further supervised learning methods.

REFERENCES

[1] I. Obeid and J. Picone, "The Temple University Hospital EEG Data Corpus" *Frontiers in Neuroscience* 13 May 2016

The Stanford EEG Corpus

Jonathan Kuo and Chris Lee-Messer
Stanford University School of Medicine

Background

- EEG is a tool for diagnosing and localizing seizures. Analysis is via tedious visual inspection by human experts.
- Progress in automating EEG analysis has been slow relative to that in other fields such a visual object recognition.
- The lack of large, open datasets for machine learning has greatly slowed progress in the field.

Aim

- Produce a large open de-identified dataset of waveforms based upon clinical EEG Recordings in order to advance the use of modern machine learning techniques for EEG analysis

Methods

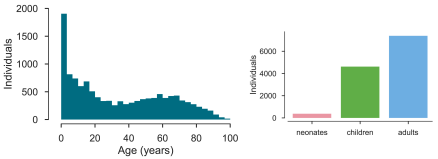
- An automated, scriptable pipeline was created to convert and de-identify Nihon-Kohden EEG files into open HDF5.
- Ascii-text EEG reports from the EMR were obtained and combined with NK database in order to classify and partially annotate the EEG

Results

- Our pipeline successfully produced 136k de-identified files
- A total of over 381+ million seconds of waveform data taking up 4.1 TB of compressed waveform data, sampled at 200-1000Hz with 19 to 128 channels per record.
- Approximately 19000 regions with seizures identified with high probability due to semi-annotated features

Population Properties

- Data from 12K + individuals seen at a tertiary/quaternary hospitals from 2008-2017
- 49% female, about 1/3 cases children
- 375+ neonates



EEG Characteristics

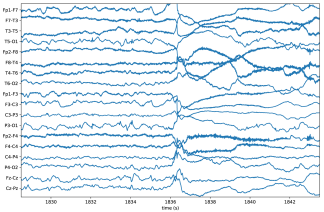
- The EEGs come from a mix of study types. Long-term monitoring (LTM) dominates. The dataset includes ~300 complete ambulatory studies (AMB). There are 604 explicitly marked intracranial studies (IE) and many more not yet specifically identified in the LTM category.
- The dataset also includes 10K total routine studies (Routine) which are usually 20-60 minutes.
- In-waveform annotations indicate a high probability of seizures in about 19K files (14%)
- A simple NLP naïve classifier, estimated that 55% of the studies were abnormal and about 8% contained seizures. This is a bit lower than the rate based upon in waveform annotations, possibly due to no longer counting multiple seizures within the same study

Files By Study Type

source	AMB	IE	LTM	Routine	Totals
lpch	204	604	29773	6060	36641
stanford	93		95623	4004	99720
Totals	297	604	125396	10064	136361

Waveform example

Visualization in the jupyter notebook of routine EEG with electrodecrement



Limitations

- The dataset is not fully representative of EEG activity present day to day at our center: it contains many clipped files which feature seizures.
- While sufficient for unsupervised and semi-supervised methods in machine learning, the EEGs lack the detailed annotations necessary for supervised ML methods.

Conclusions

- We have created a new, large EEG dataset in a open, easily accessible format suitable for machine learning techniques.

Future Directions

- We wish to provide open access to this dataset: Approval for a free license for use in research and education is currently being processed by Stanford's licensing and compliance departments.
- We plan to improve our text analysis tools to allow improved, cross-modal EEG annotation and machine learning.

Acknowledgments: This work was supported by the Stanford Department of Neurology and a grant from a patient-donor. IRB approval was obtained before work commenced.
Contact Information: Christopher Lee-Messer, Assist. Prof of Neurology, cleemess@stanford.edu.