Big Data Resources for EEGs: Enabling Deep Learning Research[1]

L. Veloso, J. McHugh, E. von Weltin, S. Lopez, I. Obeid and J. Picone

The Neural Engineering Data Consortium, Temple University
{lillian.veloso, tuf57277, eva.vonweltin, tud22978, iobeid, picone}@temple.edu

The Temple University Hospital (TUH) electroencephalography (EEG) Corpus is the world's largest open source EEG corpus of its kind [1]. This corpus consists of over 25,000 EEG studies and over 14,000 patients, and includes a neurologist's interpretation of the test, a brief medical history of the patient, and demographic information about the patients such as gender and age. This database represents the efforts of the Department of Neurology and the Neural Engineering Data Consortium to support the use of EEG data in machine learning research. The data was collected in normal clinical settings and hence includes many non-epileptic features such as muscle and movement artifacts, and a variety of channel configurations that cannot be found in currently available, more sanitized datasets. This is the first dataset of its kind to contain a sufficient amount of EEG data to support the application of state of the art deep learning algorithms. The most recent release of this corpus is v1.0.0 which includes 13,550 patients, 23,218 EEG sessions with reports and 61,634 EEG files.

Several important subsets of the data that are designed to support research in specific subspecialties of EEG analysis. The first subset, created for the purpose of studying machine learning applications in automatic seizure detection, is the TUH EEG Seizure Corpus [2]. This subset has been manually annotated by a group of student researchers for seizure events. These events are classified by their type (intensive care unit (ICU), inpatient or outpatient), subtype (specific ICUs) and duration (routine EEG or Long Term Monitoring session). The training data, having been extended, contains 196 patients, 456 sessions and 1,505 files. The evaluation data contains 50 patients, 230 sessions and 984 files.

The second subset, meant to be used for the automatic detection of abnormal EEGs, is the TUH EEG Abnormal EEG Corpus [3] . It contains both normal and abnormal EEGs, with no patients overlapping the evaluation and training datasets. Each seizure event is classified by both a student researcher and a certified neurologist, with the positive agreement being 97% and higher, and the negative agreement being 1% or lower. The training data contains 2,132 patients and 2,740 files while the evaluation data contains 253 patients with 277 files.

The third subset is the TUH EEG Slowing Corpus [4]. This corpus was developed to aid the differentiation of seizure and slowing events. Its EEG files are term-based, meaning that events are annotated on every channel, to make it more useful for machine learning research. This subset contains 38 unique patients, 75 sessions, and 300 annotations in 112 aggregated files. The annotations include 100 samples of seizures events, independent slowing events and complex background events, all of which are 10 seconds in duration.

The fourth subset is the TUH EEG Epilepsy Corpus [5]. It was created to provide data for the purposes of automatic analysis of EEG. The patients were sorted by using a filter that categorized patients into two classes: epilepsy and not epilepsy. This was based on information in the session reports relating to their clinical history, medications at the time of recording, and EEG features associated with epilepsy. This subset contains European data format (EDF) files and corresponding neurologist reports for 1799 files in 570 sessions from 200 patients. From these, 1473 files in 436 sessions from 100 patients have epilepsy, whereas 326 files in 134 sessions from 100 patients do not have epilepsy.

1. 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.

These data are open source and freely available at https://www.isip.piconepress.com/projects/tuh_eeg/downloads/. There are more than 650 registered users, making it one of the most popular resources in the EEG research community. We have done preliminary experiments using deep learning algorithms with this dataset and look forward to the future research done in this field.

REFERENCES

[1] I. Obeid and J. Picone, "The Temple University Hospital EEG Data Corpus," *Front. Neurosci. Sect. Neural Technol.*, vol. 10, p. 196, 2016.

[2] M. Golmohammadi, V. Shah, S. Lopez, S. Ziyabari, S. Yang, J. Camaratta, I. Obeid, and J. Picone, "The TUH EEG Seizure Corpus," in *Proceedings of the American Clinical Neurophysiology Society Annual Meeting*, 2017, p. 1.

[3] S. Lopez, "Automated Identification of Abnormal EEGs," Temple University, 2017.

[4] E. von Weltin, T. Ahsan, V. Shah, D. Jamshed, M. Golmohammadi, I. Obeid, and J. Picone, "Electroencephalographic Slowing: A Source of Error in Automatic Seizure Detection," in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, 2017, pp. 1–5.

[5] J. R. McHugh, J. Picone, and I. Obeid, "The TUH EEG Epilepsy Corpus," 2017. [Online]. Available: https://www.isip.piconepress.com/projects/tuh_eeg/downloads/tuh_eeg_epilepsy/. [Accessed: 14-Nov-2017].

NEURAL ENGINEERING DATA CONSORTIUM
www.nedcdata.org

Big Data Resources for EEGs: Enabling Deep Learning Research

L. Veloso, J. McHugh, E. von Weltin, S. Lopez, I. Obeid and J. Picone

The Neural Engineering Data Consortium, Temple University

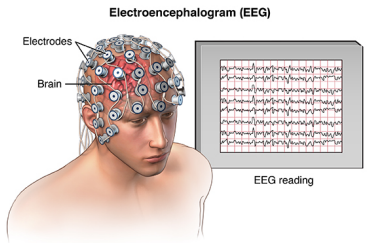
College of Engineering
Temple University

Abstract

- The Temple University Hospital Electroencephalography Corpus (TUHEEG) is the world's largest open source EEG corpus of its kind.
- Several important subsets of the data that are designed to support research in specific subspecialties of EEG analysis are:
  - TUH EEG Seizure Corpus: created for automatic seizure detection research; has been manually annotated for seizure events; events are classified by type (e.g., tonic) and subtype (e.g., ICU), and duration (e.g., routine or LTM).
  - TUH Abnormal EEG Corpus: supports research on classification of abnormal EEGs; includes patients ranging in age from 10-100 and many challenging benign conditions.
  - TUH EEG Slowing Corpus: developed to aid in the development of a tool that can differentiate between slowing at the end of a seizure and an independent non-seizure slowing event.
  - TUH EEG Epilepsy Corpus: contains 436 sessions from 100 patients with epilepsy and 134 sessions from 100 patients without epilepsy.
- These data are open source and freely available at https://www.isip.piconepress.com/projects/tuh_eeg/downloads/. There are more than 650 registered users of these resources, making them one of the most popular resources in the research community.

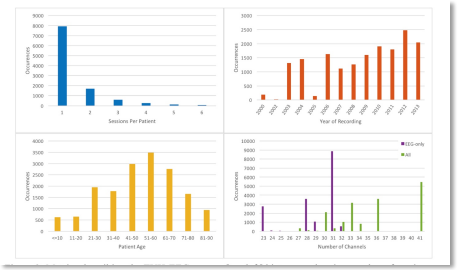
Introduction

- The data is harvested from clinical recordings collected at a large urban public hospital (TUH).
- All data and reports have been rigorously deidentified to ensure the data is HIPAA compliant.
- Each EEG session includes EEG signal data, a neurologist's report, and annotation information required to conduct machine learning research.
- All EEGs are collected in a clinical setting, meaning they have non-epileptic features such as muscle artifacts and patient movements.
- Each file is manually annotated by a team of neuroscience students; each session is typically reviewed by at least three annotators.
- Annotator accuracy has been validated against board-certified clinicians (Kappa ~ 0.8).



The TUH EEG Corpus (v1.0.0)

- The master corpus: contains all EEG sessions collected at TUH from 2002 to 2015. Data collection is ongoing (increasing at a rate of 3K sessions/yr).
- The corpus includes a rich and diverse set of patients and medical histories:

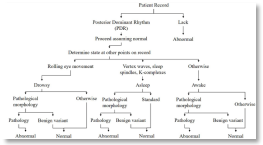


- The signal data totals 977 GB, while the reports are a total of 93 MB.
- Over 40 unique channel configurations; 90% of the database consists of Averaged Reference (AR) and Linked Ear (LE) EEGs; 95% of the data conforms to a standard 10/20 EEG configuration.

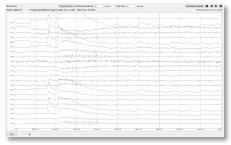
No. of Patients	No. of Sessions	Total Duration
13,551	23,257	15,968 hours

The TUH Abnormal EEG Corpus

- All EDF files are 15 minutes or longer
- Each seizure event is classified by both a student annotator and a certified neurologist
- Positive agreement between the two groups was 97% and negative agreement was 1% or lower.



- Abnormal: epileptiform features, such as spike and wave discharges, are present at the vertex of the scalp.
- Normal: eye blink artifacts and posterior dominant rhythm (PDR) are both normal features.



	No. of Patients	No. of Sessions	Total Duration
Training	2,132	2,740	1,045 hours
Evaluation	253	277	103 hours

The TUH EEG Epilepsy Corpus

- Patients were filtered based on criteria in the reports that indicated signs of epilepsy as determined by neurologists from NIH.
- Reports were searched for keywords and medications that are indicative of epilepsy:
  - Medications that normally indicate a history of epilepsy (e.g., Keppra, Levetircatam, Vimpat).
  - Keywords that indicate seizure behavior:

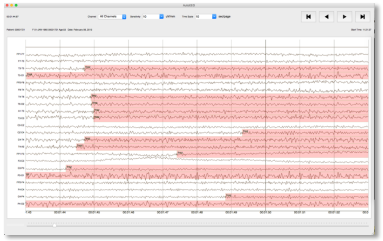
Typical Inclusion Criteria	Typical Exclusion Criteria
Spike and wave	No sharp wave or spike
Sharp wave	Single sharp wave or spike
Sharp waves	No focal or epileptiform
Spike	Left anterior temporal sharp wave

- Search results were manually reviewed to make sure they conformed to the requirements.
- Data is being used to correlate seizures in EEG signals with patterns in interictal EKGs.
- This corpus represents one of the first efforts to subdivide TUH EEG for use in machine learning.

Epilepsy Diagnosis	No. of Patients	No. of Sessions	Total Duration
Yes	100	436	351 hours
No	100	134	72 hours

The TUH EEG Seizure Corpus (v1.2.0)

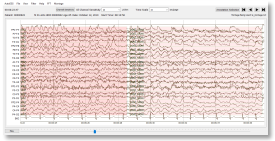
- A subset of TUHEEG created to support deep learning research into automatic seizure detection.
- Each file has been reviewed and manually annotated by a team of students for seizure events (start time, stop time, type of seizure).
- Each session is classified by it's type: Inpatient, Outpatient, EMU, and ICU.
- Each session has a subtype: ER, OR, General, Outpatient, EMU, and the different ICUs: BURN, CICU, ICU, NICU, NSICU, PICU, RICU, and SICU.
- Both routine (20 min.) and long-term (24 hr.) EEGs.

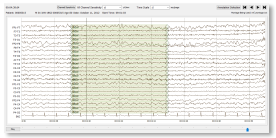


	No. of Patients	No. of Sessions	No. of Seizures
Training	196	456	1,303
Evaluation	50	230	649

The TUH EEG Slowing Corpus

- Created to aid in the differentiation of seizure events and slowing events in machine learning, which is the most common single error modality.
- Contains 100 samples of seizure, independent slowing, and complex background events.
- The samples of slowing and complex background were collected manually, while the seizure events were taken from the TUH EEG Seizure Corpus.
- Each sample is 10 seconds long to facilitate simple machine learning experiments using neural networks, which prefer fixed-length patterns.
- Post-ictal slowing which is observed at the termination of many seizure events.





- Intermittent electrographic slowing that can cause false alarms.

No. of Unique Patients	No. of Sessions	Total No. of Events
38	75	100

Summary

- The TUH EEG Corpus is enabling the application of state of the art machine learning algorithms to problems such as seizure detection.
- The open source nature of the data (e.g., no IRB or data-sharing agreements are required) makes it accessible to a large community of researchers.
- There are currently over 650 registered users of these resources. Plans for 2018 include an open Kaggle-style competition on seizure detection.
- Future plans include:
  - Data collection at TUH will continue for at least the next three years, growing the corpus at a rate of at least 3,000 sessions per year.
  - Parsed medical reports that contain medical concepts, their attributes, and knowledge representations that describe how these concepts relate to one another.
- A digital pathology corpus of 1M images!

Acknowledgements

- 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.