**B****ig Data Resources for EEGs: Enabling Deep Learning Research1**

*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.

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.

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.

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].