**Software and Data Resources to
Advance Machine Learning Research in Electroencephalography**

*S. Rahman, M. Miranda, I. Obeid and J. Picone*

The Neural Engineering Data Consortium, Temple University, Philadelphia, Pennsylvania, USA

{tuh01696, matthewsm, iobeid, picone}@temple.edu

The Neural Engineering Data Consortium at Temple University has been providing key data resources to support the development of deep learning technology for electroencephalography (EEG) applications [1-4] since 2012. We currently have over 1,700 subscribers to our resources and have been providing data, software and documentation from our web site [5] since 2012. In this poster, we introduce additions to our resources that have been developed within the past year to facilitate software development and big data machine learning research.

Major resources released in 2019 include:

* **Data:** The most current release of our open source EEG data is v1.2.0 of TUH EEG and includes the addition of 3,874 sessions and 1,960 patients from mid-2015 through 2016.
* **Software:** We have recently released a package, PyStream, that demonstrates how to correctly read an EDF file and access samples of the signal. This software demonstrates how to properly decode channels based on their labels and how to implement montages. Most existing open source packages to read EDF files do not directly address the problem of channel labels [6].
* **Documentation:** We have released two documents that describe our file formats and data representations: (1) *electrodes and channels [6]:* describes how to map channel labels to physical locations of the electrodes, and includes a description of every channel label appearing in the corpus; (2) *annotation standards [7]:* describes our annotation file format and how to decode the data structures used to represent the annotations.

Additional significant updates to our resources include:

* *NEDC TUH EEG Seizure (v1.6.0):* This release includes the expansion of the training dataset from 4,597 files to 4,702. Calibration sequences have been manually annotated and added to our existing documentation. Numerous corrections were made to existing annotations based on user feedback.
* *IBM TUSZ Pre-Processed Data (v1.0.0):* A preprocessed version of the TUH Seizure Detection Corpus using two methods [8], both of which use an FFT sliding window approach (STFT). In the first method, FFT log magnitudes are used. In the second method, the FFT values are normalized across frequency buckets and correlation coefficients are calculated. The eigenvalues are calculated from this correlation matrix. The eigenvalues and correlation matrix's upper triangle are used to generate feature.
* *NEDC TUH EEG Artifact Corpus (v1.0.0):* This corpus was developed to support modeling of non-seizure signals for problems such as seizure detection. We have been using the data to build better background models. Five artifact events have been labeled: (1) eye movements (EYEM), (2) chewing (CHEW), (3) shivering (SHIV), (4) electrode pop, electrostatic artifacts, and lead artifacts (ELPP), and (5) muscle artifacts (MUSC). The data is cross-referenced to TUH EEG v1.1.0 so you can match patient numbers, sessions, etc.
* *NEDC Eval EEG (v1.3.0):* In this release of our standardized scoring software, the False Positive Rate (FPR) definition of the Time-Aligned Event Scoring (TAES) metric has been updated [9]. The standard definition is the number of false positives divided by the number of false positives plus the number of true negatives: #FP / (#FP + #TN).

We also recently introduced the ability to download our data from an anonymous rsync server. The rsync command [10] effectively synchronizes both a remote directory and a local directory and copies the selected folder from the server to the desktop. It is available as part of most, if not all, Linux and Mac distributions (unfortunately, there is not an acceptable port of this command for Windows). To use the rsync command to download the content from our website, both a username and password are needed. An automated registration process on our website grants both. An example of a typical rsync command to access our data on our website is:

*rsync -auxv nedc\_tuh\_eeg@www.isip.piconepress.com:~/data/tuh\_eeg/*

Rsync is a more robust option for downloading data. We have also experimented with Google Drive and Dropbox, but these types of technology are not suitable for such large amounts of data.

All of the resources described in this poster are open source and freely available at *https://www.isip.piconepress.com/projects/tuh\_eeg/downloads/*. We will demonstrate how to access and utilize these resources during the poster presentation and collect community feedback on the most needed additions to enable significant advances in machine learning performance.

Acknowledgments

Research reported in this publication was most recently supported by the National Science Foundation Partnership for Innovation award number IIP-1827565 and the Pennsylvania Commonwealth Universal Research Enhancement Program (PA CURE).

Several grants over the years have supported this database development project. Significant contributors include National Human Genome Research Institute of the National Institutes of Health award number U01HG008468, DARPA Microsystems Technology Office award number D13AP00065, National Science Foundation Division of Computer and Network Systems award number CNS-1305190, the Temple University Office of the Vice-Provost for Research and the Temple University College of Engineering.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the official views of any of these organizations.

References

1. S. Ferrell, E. von Weltin, I. Obeid, and J. Picone, “Open Source Resources to Advance EEG Research,” *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, 2018, pp. 1–3.
2. L. Veloso, J. R. McHugh, E. von Weltin, I. Obeid, and J. Picone, “Big Data Resources for EEGs: Enabling Deep Learning Research,” *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, 2017, p. 1.
3. I. Obeid and J. Picone, “The Temple University Hospital EEG Data Corpus,” *Augmentation of Brain Function: Facts, Fiction and Controversy. Volume I: Brain-Machine Interfaces*, 1st ed., vol. 10, M. A. Lebedev, Ed. Lausanne, Switzerland: Frontiers Media S.A., 2016, pp. 394–398.
4. N. Shawki et al., “The Temple University Digital Pathology Corpus,” in *Machine Learning Applications in Medicine and Biology* (Tentative), 1st ed., I. Obeid and J. Picone, Eds. New York City, New York, USA: Springer-Verlag, 2019, p. 45.
5. S. I. Choi, S. Lopez, I. Obeid, M. Jacobson, and J. Picone, “The Temple University Hospital EEG Corpus,” The Neural Engineering Data Consortium, College of Engineering, Temple University, 2017. [Online]. Available: *http://www.isip.piconepress.com/projects/tuh\_eeg*.
6. S. Ferrell, V. Mathew, T. Ahsan, and J. Picone, “The Temple University Hospital EEG Corpus: Electrode Location and Channel Labels,” Philadelphia, Pennsylvania, USA, 2019.
7. S. Ferrell, L. Jakielaszek, T. Elseify, and J. Picone, “The Temple University Hospital EEG Corpus: Annotation File Formats,” Philadelphia, Pennsylvania, USA, 2019.
8. S. Roy, U. Asif, J. Tang, and S. Harrer, “Machine Learning for Seizure Type Classification: Setting the benchmark,” *arXiv*, pp. 1–5, 2019.
9. V. Shah, M. Golmohammadi, I. Obeid, and J. Picone, “Objective evaluation metrics for automatic classification of EEG events,” *J. Neural Eng.*, pp. 1–21, 2019.
10. A. Tridgell and P. Mackerras, “First release of rsync - rcp replacement,” Newsgroup: comp.os.linux.announce, 1996. [Online]. Available: https://groups.google.com/forum/#!msg/comp.os.linux.announce/tZE1qtTcQaU/IF8GhGQ\_uTsJ. [Accessed: 31-Oct-2019].