Open Source Resources to Advance EEG Research[1,2]

S. Ferrell, E. von Weltin, I. Obeid and J. Picone

Neural Engineering Data Consortium, Temple University, Philadelphia, Pennsylvania, USA
{sean.ferrell, eva.vonweltin, iobeid, picone}@temple.edu

Since 2012, the Neural Engineering Data Consortium (NEDC) at Temple University has been providing many key data resources to support machine learning research in bioengineering [1]. In this poster, we present an update on several significant resources available from NEDC that were established to support a new generation of electroencephalogram (EEG) technology development.

Our flagship product, the Temple University Hospital Electroencephalography Corpus (TUH EEG), is the world's largest open source EEG corpus [2]. Over 1,200 sites worldwide are using this corpus to support their machine learning research. For the past 6 years we have collected every EEG session conducted at Temple University Hospital. Our historical archive goes back to 2002 and has resulted in a database of over 30,000 EEG studies from more than 16,000 patients. Each study includes a report that consists of detailed information about the patient, the patient's medical history, and a neurologist's review of the study.

Our most current release is v1.1.0 and includes 13,539 patients, 23,002 EEG sessions and reports, and 53,506 EEG files. A new feature of this release is the inclusion of three types of automatically generated annotations: (1) seizure events that include the start time, stop time, channel label and type of seizure, (2) normal/abnormal classification of a session, and (3) a six-way classification of each one-second epoch (spike/sharp waves, periodic lateralized epileptiform discharges, generalized periodic epileptiform discharges, artifacts, eye movement, and background). These annotations were generated automatically using a state-of-the-art deep learning system [3].

There are a number of subsets of this corpus that were created to support specific types of EEG analysis:

- *The TUH EEG Seizure Corpus (v1.3.0):* supports the development of automatic seizure detection technology. The data has been manually annotated for seizure events [4]. The training dataset has been extended and contains 264 patients, 580 sessions, and 1,987 files. The development test dataset consists of 50 patients, 238 sessions, and 1,013 EEG files. Also included is a held-out blind evaluation dataset which will be used in an upcoming Kaggle-style challenge hosted by IBM. This blind evaluation set consists of 50 patients, 152 sessions, and 1,023 EEG files. More details about this challenge will follow within the next few months.
- *The TUH Abnormal EEG Corpus (v2.0.0):* supports the development of automatic detection of abnormal EEGs. The training dataset consists of 2,310 patients, 2,717 sessions, and 2,717 EEG files. The evaluation dataset includes 253 patients, 276 sessions, and 276 EEG files. Each session is labeled as either normal or abnormal using a decision-making process described by Lopez et al. [5]. Approximately 50% of the data constitute abnormal EEG sessions.
- *The TUH EEG Slowing Corpus (v1.0.1):* aids in the development of technology that can differentiate between post-ictal and transient slowing. Slowing can be a focal or generalized decrease in frequency and is either a part of a seizure or an independent event. This corpus consists of 38 patients, 75

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.

2. Portions of this work were also sponsored by Temple University's Office of the Senior Vice-Provost for Research through a Pennsylvania Commonwealth Universal Research Enhancement (PA-CURE) grant.

sessions, and 300 annotations in 112 aggregated files. It has been used to study common error modalities in automated seizure detection [6][7].

- *The TUH EEG Epilepsy Corpus (v1.0.0):* supports the study of the characteristics of patients diagnosed as epileptic. These patients were sorted into two classes (epileptic/non-epileptic) based on medications listed during the recording, the clinical history of the patient, and EEG signal features associated with epilepsy. There are 237 patients, 561 sessions, and 1,648 files in this dataset.
- *The TUH EEG Events Corpus (v1.0.1):* used to develop the six-way classification system previously described. This corpus consists of two datasets, training (359 files) and evaluation (159 files).

An important feature of all these corpora is that patient numbering is now consistent. Therefore, users can study specific patients or diseases across a broad range of conditions. Previous releases of these corpora had not been reconciled because they were preliminary releases made at different points in time and were anonymized individually.

We also have expanded the documentation about the corpus to explain electrode locations, channel labels, sample frequencies and annotation formats. There is a wealth of metadata now available for each patient and session.

In addition to data resources, NEDC also provides some supporting tools to facilitate research. These include:

- *The NEDC Demo System (v0.4.1)*: a visualization tool developed to rapidly annotate EEG signals [8]. Our demo system allows users to annotate EEG signals on a per-channel basis. This tool is written in Python, using the PyQt toolkit, and is easily customized to support specific annotation tasks. We have used it on most popular operating systems including Windows, Linux, and Mac systems. A cohort retrieval system has been integrated into this viewer that allows users to query the TUH EEG Corpus for sessions that match particular search criteria. Both signal and report events can be searched.
- *NEDC Eval EEG (v1.2.0):* a standardized scoring package that is an important piece of any common evaluation framework [9]. This software implements a variety of popular scoring metrics based on common measures such as sensitivity, specificity, and Cohen's kappa statistic. It also includes new metrics based on time-aligned and epoch scoring, providing a more balanced view of performance.

In this poster, we will discuss these resources and describe how they fit together to promote the development of deep learning technology for automatic interpretation of EEGs. All of the data and resources presented here are freely available at *https://www.isip.piconepress.com/projects/tuh_eeg/downloads/*. No licensing or data sharing agreements are needed. An automated registration process provides users with a username and password to access the data. The unencumbered nature of these resources is a very important differentiating feature of these NEDC resources.

REFERENCES

[1] L. Veloso, J. R. McHugh, E. von Weltin, I. Obeid, and J. Picone, "Big Data Resources for EEGs: Enabling Deep Learning Research," presented at the *IEEE Signal Processing in Medicine and Biology Symposium*, 2017, p. 1.

[2] I. Obeid and J. Picone, "The Temple University Hospital EEG Data Corpus," *Front. Neurosci. Sect. Neural Technol.*, vol. 10, p. 00196, 2016.

[3] M. Golmohammadi, A. Harati Nejad Torbati, S. Lopez, I. Obeid, and J. Picone, "Automatic Analysis of EEGs Using Big Data and Hybrid Deep Learning Architectures," *Front. Hum. Neurosci.*, pp. 1–30, 2018 (in review). *https://www.isip.piconepress.com/publications/unpublished/journals/2018/frontiers_neuroscience/hybrid/*.

[4] V. Shah, E. von Weltin, S. Lopez, J. R. McHugh, L. Veloso, M. Golmohammadi, I. Obeid, and J. Picone, "The TUH EEG Seizure Corpus," *Front. Neurosci.*, pp. 1–9, 2018 (in review). *https://www.isip.piconepress.com/publications/unpublished/journals/2018/frontiers_neuroinformatics/tuh_eeg_seizure/*.

[5] S. Lopez, I. Obeid, and J. Picone, "Automated Interpretation of Abnormal Adult Electroencephalograms," presented at the 26th Conference on Intelligent Systems for Molecular Biology, 2018, p. 1.

[6] M. Golmohammadi, I. Obeid, and J. Picone, "Deep Residual Learning for Automatic Seizure Detection," presented at the 26th Conference on Intelligent Systems for Molecular Biology, 2018, p. 1.

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

[8] N. Capp, E. Krome, I. Obeid, and J. Picone, "Rapid Annotation of Seizure Events Using an Extensible Visualization Tool," in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, 2017, p. 1.

[9] V. Shah, S. Ziyabari, M. Golmohammadi, I. Obeid, and J. Picone, "Objective Evaluation Metrics for Automatic Classification of EEG Events," *J. Neural Eng.*, pp. 1–19, 2018 (in review). *https://www.isip.piconepress.com/publications/unpublished/journals/2018/iop_jne/metrics/*.

Open Source Resources to Advance EEG Research

S. Ferrell, E. von Weltin, I. Obeid and J. Picone

The Neural Engineering Data Consortium, Temple University

NEURAL ENGINEERING DATA CONSORTIUM
www.nedcdata.org

College of Engineering
Temple University

Abstract

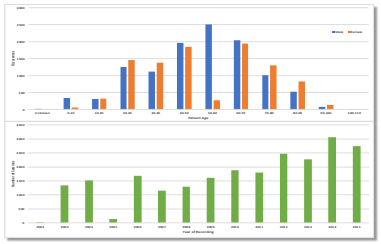
- Created in 2012, the Temple University Hospital Electroencephalography Corpus (TUH EEG) is the world's largest open source EEG corpus.
- There are a number of important subsets of this corpus that were developed to support specific types of EEG analysis:
  - TUH EEG Seizure Corpus: supports the development of automatic seizure detection technology.
  - TUH Abnormal EEG Corpus: supports the development of automatic detection of abnormal EEGs.
  - TUH EEG Slowing Corpus: aids in the development of technology that can differentiate between post-ictal and transient slowing.
  - TUH EEG Epilepsy Corpus: supports the study of patients diagnosed as epileptic.
  - TUH EEG Events Corpus: used to develop the six-way classification system.
- In addition to data resources, supporting tools to facilitate research are also provided:
  - NEDC Demo System: a visualization tool developed to rapidly annotate EEG signals.
  - NEDC Eval EEG: a standardized scoring package that is an important piece of any common evaluation framework.

Introduction

- NEDC's historical archive of EEG data dates as far back as 2002 and includes every EEG collected at Temple University Hospital (TUH).
- This corpus now includes over 40,000 EEG studies from more than 18,000 patients.
- This dataset is the first database large enough to support the application of state of the art machine learning algorithms.
- Each study includes a report that contains information about the patient, the patient's medical history, and a neurologist's findings.
- The patient numbering is consistent throughout these corpora; users can study specific patients or diseases across a broad range of conditions.
- Documentation about the corpus has been expanded to explain electrode locations, channel labels, sample frequencies, and annotation formats.
- Automatically generated annotations of the data are also available (e.g., normal/abnormal).

The TUH EEG Corpus (v1.1.0)

- The main corpus: contains every EEG session collected at TUH from 2002 to 2015. Data collection is ongoing (increasing at a rate of 3K sessions/yr).
- There is a variety of patients and medical histories included in the corpus:



- The signal data totals 959 GB, while the reports are a total of 94 MB.
- The latest release includes three types of automatically generated annotations that include: (1) seizure events, (2) abnormal events, and (3) six-way classification of each one-second epoch.

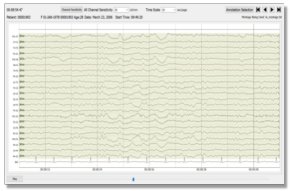
No. of Patients	No. of Sessions	Total Duration
13,539	23,002	15,757 hours

The TUH Abnormal EEG Corpus (v2.0.0)

- Supports the development of automatic detection of abnormal EEGs.
- Approximately 50% of the data constitute abnormal EEG sessions.
- Each session is labeled as either normal or abnormal using a specific decision-making process.

The TUH EEG Slowing Corpus (v1.0.1)

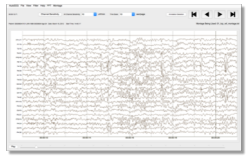
- Aids in the development of technology that can differentiate between post-ictal and transient slowing.
- Slowing can be a focal or generalized decrease in frequency and is either a part of a seizure or an independent event.



- This corpus has been used to study common error modalities in automated seizure detection.

The NEDC Demo System (v0.4.1)

- A visualization tool developed to rapidly annotate EEG signals.
- Written in Python using the PyQt toolkit and is easily customized to support specific annotation tasks.
- User can annotate EEG signals on a per-channel basis.
- Can be used on most popular operating systems including Windows, Linux, and Mac systems.



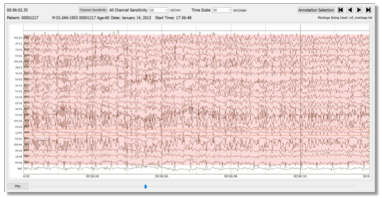
- A cohort retrieval system has been integrated into this viewer that allows users to query the TUH EEG Corpus for sessions that match particular search criteria.

NEDC Eval EEG (v1.2.0)

- A standardized scoring package that is an important piece of any common evaluation framework.
- Implements a variety of popular scoring metrics based on common measures such as sensitivity, specificity, and Cohen's kappa statistic.
- Includes a new metric based on time-aligned and epoch scoring, providing a more balanced view of performance.

The TUH EEG Seizure Corpus (v1.3.0)

- A subset of TUH EEG (TUSZ) developed towards automatic seizure detection through machine learning research.
- Every EEG file has been manually annotated for seizure events including channel number, start and stop times, and type of seizure.
- A record of calibration start and end times that occur in the beginning of an EEG recording are now included.
- A held-out blind evaluation dataset was developed which will be used in an upcoming Kaggle-style challenge hosted by IBM.



	No. of Patients	No. of Sessions	No. of Seizures
Training	264	580	1,247
Dev test	50	238	685

The TUH EEG Epilepsy Corpus (v1.0.0)

- Supports the study of characteristics of patients diagnosed as epileptic.
- The patients are sorted into two classes: (1) epileptic and (2) non-epileptic.
- Patients were sorted into these classes based on medications listed during the recording, clinical history of the patient, and EEG signal features associated with epilepsy.
- Some keywords that indicate seizure behavior include: spike and wave, sharp wave, sharp waves, and spike.

The TUH EEG Events Corpus (v1.0.1)

- This corpus is used to develop a six-way classification system of each one-second epoch.
- The system detects three events of clinical interest: (1) spike and/or sharp waves, (2) periodic lateralized epileptiform discharges, and (3) generalized epileptiform discharges.
- Three events used to model background are also detected: (1) eye movement, (2) artifacts, and (3) background.

	No. of Patients	No. of Files	No. of Events
Training	290	359	547
Evaluation	80	159	240

Summary

- The TUH EEG Corpus and its variants promote the development of machine learning technology for automatic interpretation of EEGs.
- These corpora and supporting tools are open source and freely available at https://www.isip.piconepress.com/projects/tuh_eeg/downloads/.
- Future plans include:

Database	Version	Date
TUH EEG	v2.0.0	January, 2019
TUSZ	v1.7.0	January, 2019
TUH EEG	v3.0.0	March, 2019
TUSZ	v1.8.0	March, 2019

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.
- This research was also supported in part by a grant from the Temple University College of Engineering Research Experience for Undergraduates program and the Pennsylvania Commonwealth Universal Research Enhancement Program (PA CURE).