File: \_AAREADME.txt

Database: TUH EEG Seizure Corpus

Version: 1.3.0

------------------------------------------------------------------------------------

Change Log:

(20180816) data was cleaned up and several annotation corrections were made.

several files were removed from the training set and development

test set. calibration start and end times for the eval data set are

recorded in \_SEIZURES.xlsx.

(20180415) cleaned up the data for release.

(20180323) made several annotation corrections. the evaluation (eval) set has

been renamed as the development test (dev\_test) set. a major

addition to this release is a record of calibration start and end

times in the beginning of the file that is included in

\_SEIZURES.xlsx.

(20171206) added new files to evaluation and training sets. files in the

montage 03\_tcp\_ar\_a have been added. patient numbers have been

changed to reflect TUH EEG v1.1.0.

(20170804) added annotations for the type of seizure event and replaced

rec files with tse and lbl files. added seizure annotations to

one file in the training set.

(20170701) includes the expanded training set, sub-one-second resolution

on the start/stop times of the annotations, expanded

documentation on the type of EEG session, and repair of

corrupted headers.

(20170615) several annotation errors were fixed. the annotation files

were reduced to biclass (seizure/no-seizure). a detailed

spreadsheet was included describing the data.

(20170427) changed the filenames to match our upcoming TUH EEG release

(20170426) naming conventions on the top-level directories and improved documentation.

----

This file contains some basic statistics about the TUH EEG Seizure Corpus, a corpus developed to motivate the development of high performance seizure detection algorithms using machine learning. This corpus is a subset of the TUH EEG Corpus and contains sessions that are known to contain seizure events. To balance the corpus, some sessions are provided that do not contain seizure events, so that the false alarm performance of a system can be tested.

When you use this specific corpus in your research or technology development, we ask that you reference the corpus using this

publication:

Golmohammadi, M., Shah, V., Lopez, S., Ziyabari, S., Yang, S., Camaratta, J., Obeid, J. and Picone, J. (2017). The TUH EEG Seizure Corpus. In American Clinical Neurophysiology Society (p. 1). Phoenix, Arizona,USA.

This publication can be retrieved from:

https://www.isip.piconepress.com/publications/conference\_presentations/2017/acns/tuh\_eeg\_seizures/

Our preferred reference for the TUH EEG Corpus, from which this seizure corpus was derived, is:

Obeid, I., & Picone, J. (2016). The Temple University Hospital EEG Data Corpus. Frontiers in Neuroscience, Section Neural Technology, 10, 196. <http://doi.org/http://dx.doi.org/10.3389/fnins.2016.00196> v1.3.0 of the TUH EEG Seizure Corpus was based on v1.1.0 of the TUH EEG Corpus.

There are two main directories in this release: dev\_test and train. The train directory contains data you are allowed to use for the development of your technology. The dev\_test data is disjoint from the training set and should only be used for testing.

The top-level directories:

edf/dev\_test/01\_tcp\_ar

edf/dev\_test/02\_tcp\_le

edf/dev\_test/03\_tcp\_ar\_a

edf/train/01\_tcp\_ar

edf/train/02\_tcp\_le

edf/train/03\_tcp\_ar\_a

refer to the appropriate channel configurations for the EEGs. 01\_tcp\_ar refers to an AR reference configuration, with annotations referencing a TCP format described below.

The pathname of a typical EEG file can be explained as follows:

Filename:

edf/dev\_test/01\_tcp\_ar/002/00000258/s002\_2003\_07\_21/00000258\_s002\_t000.edf

Components: edf: contains the edf data

dev\_test: part of the dev\_test set (vs.) train

01\_tcp\_ar: data that follows the averaged reference (AR) configuration, while annotations use the TCP channel configuration

002: a three-digit identifier meant to keep the number of subdirectories in a directory manageable. This follows the TUH EEG v1.1.0 convention.

00000258: official patient number that is linked to v1.1.0 of TUH EEG

s002\_2003\_07\_21: session two (s002) for this patient. The session was archived on 07/21/2003.

00000258\_s002\_t000.edf: the actual EEG file. These are split into a series of files starting with t000.edf, t001.edf, ... These represent pruned EEGs, so the original EEG is split into these segments, and uninteresting parts of the original recording were deleted (common in clinical practice).

The easiest way to access the annotations is through the spreadsheet provided (\_SEIZURES\_\*.xlsx). This contains the start and stop time of each seizure event in an easy to understand format. Convert the file to .csv if you need a machine-readable version.

There are six types of files in this release:

\*.edf: the EEG sampled data in European Data Format (edf)

\*.txt: the EEG report corresponding to the patient and session

\*.tse: term-based annotations using all available seizure type classes

\*.tse\_bi: same as \*.tse except bi-class annotations (seizure/background)

\*.lbl: event-based annotations using all available seizure type classes

\*.lbl\_bi: same as \*.lbl except bi-class annotations (seizure/background)

**Event-based annotations** are per-channel. This means the annotation contains, in addition to a start and stop time, a channel index. Seizures often can be observed on one or more channels and then spread to other channels. Event-based annotations capture this.

**Term-based annotations** use one label that applies to all channels. These are most useful for machine learning research in which we tend to worry only about the overall classification of a segment and are not concerned about individual channels.

**Bi-class annotations use two labels**: seizure and background. The multi-class annotations use all available seizure types. There are described in the spreadsheet (\_SEIZURES\_\*.xlsx).

Time-synchronous event (TSE) files use a simple format that looks like this: 0.0000 490.0000 bckg 1.0000

The fields are: start time in secs, stop time in secs, label and probability (by default, set to 1.0).

Label files (LBL) are more complicated and essentially describe a graph that can represent a hierarchical annotation (e.g., FNSZ and GNSZ map to SEIZ). They contain the start and stop times, a channel index, a level index and probabilities for each class or symbol.

Clinical EEGs use a variety of channel configurations. In the larger TUH EEG Corpus, there are over 40 different channel configurations. In this subset, **there are two type of EEGs: averaged reference (AR) and linked ears reference (LE).** Fortunately, all files in this subset contain the standard channels you would expect from a 10/20 configuration, and all files can be converted to a TCP montage (which is what we use internally for our processing).

What is somewhat confusing is that some patients have sessions listed under both 01\_tcp\_ar and 02\_tcp\_le. There are 50 unique patients in the development test set and 266 patients in the training set. But a find command will return slightly higher numbers:

find dev\_test -mindepth 2 -maxdepth 2 | wc

65 65 1570

find train -mindepth 2 -maxdepth 2 | wc

303 303 7675

because some patients appear in multiple montages:

ls -1 -d \*/\*/\*/\*/00002991

edf/train/01\_tcp\_ar/029/00002991

edf/train/02\_tcp\_le/029/00002991

edf/train/03\_tcp\_ar\_a/029/00002991

ls -1 -d \*/\*/\*/\*/00002297

edf/dev\_test/02\_tcp\_le/022/00002297

edf/dev\_test/03\_tcp\_ar\_a/022/00002297

To learn more about this, please consult the following publication:

Lopez, S., Gross, A., Yang, S., Golmohammadi, M., Obeid, I., & Picone, J. (2016). An Analysis of Two Common Reference Points for EEGs. In IEEE Signal Processing in Medicine and Biology Symposium (pp. 1–4). Philadelphia, Pennsylvania, USA. Available at: https://www.isip.piconepress.com/publications/conference\_proceedings/2016/ieee\_spmb/montages/.

The channel number in .lbl and .lbl\_bi files refers to the channels defined using a standard ACNS TCP montage. This is our preferred way of viewing seizure data. The montage is defined as follows:

montage = 0, FP1-F7: EEG FP1-REF -- EEG F7-REF

montage = 1, F7-T3: EEG F7-REF -- EEG T3-REF

montage = 2, T3-T5: EEG T3-REF -- EEG T5-REF

montage = 3, T5-O1: EEG T5-REF -- EEG O1-REF

montage = 4, FP2-F8: EEG FP2-REF -- EEG F8-REF

montage = 5, F8-T4 : EEG F8-REF -- EEG T4-REF

montage = 6, T4-T6: EEG T4-REF -- EEG T6-REF

montage = 7, T6-O2: EEG T6-REF -- EEG O2-REF

montage = 8, A1-T3: EEG A1-REF -- EEG T3-REF

montage = 9, T3-C3: EEG T3-REF -- EEG C3-REF

montage = 10, C3-CZ: EEG C3-REF -- EEG CZ-REF

montage = 11, CZ-C4: EEG CZ-REF -- EEG C4-REF

montage = 12, C4-T4: EEG C4-REF -- EEG T4-REF

montage = 13, T4-A2: EEG T4-REF -- EEG A2-REF

montage = 14, FP1-F3: EEG FP1-REF -- EEG F3-REF

montage = 15, F3-C3: EEG F3-REF -- EEG C3-REF

montage = 16, C3-P3: EEG C3-REF -- EEG P3-REF

montage = 17, P3-O1: EEG P3-REF -- EEG O1-REF

montage = 18, FP2-F4: EEG FP2-REF -- EEG F4-REF

montage = 19, F4-C4: EEG F4-REF -- EEG C4-REF

montage = 20, C4-P4: EEG C4-REF -- EEG P4-REF

montage = 21, P4-O2: EEG P4-REF -- EEG O2-REF

**For example, channel 1 is a difference between electrodes F7 and T3, and represents an arithmetic difference of the channels (F7-REF)-(T3-REF), which are channels contained in the EDF file. For files in the 02\_tcp\_le montage the channels are named as EEG P4-LE. All channel derivations are the same. For files in the 03\_tcp\_ar\_a montage the derivations EEG A1-REF and EEG A2-REF are not included.**

A spreadsheet is provided that classify each seizure by type. This spreadsheet contains a legend that explains these fields. Finally, here are some basic descriptive statistics about the data:

DEVELOPMENT TEST SET

total files: 1013

total sessions: 238

total patients: 50

files with seizures: 284

sessions with seizures: 107

patients with seizures: 39

total number of seizures: 685

total seizure duration: 60882.8694 secs (9.9200%)

total background duration: 552349.1306 secs

total duration: 613232.0000 secs

total duration of files with seizures: 234007.0000 secs (38.1500%)

-----------------------------

TRAINING SET

total files: 1987

total sessions: 580

total patients: 264

files with seizures: 412

sessions with seizures: 194

patients with seizures: 128

total number of seizures: 1247

total seizure duration: 89125.9710 secs (7.5000%)

total background duration: 1098164.0290 secs

total duration: 1187290.0000 secs

total duration of files with seizures: 309542.0000 secs (26.0700%)

-----------------------------

If you have any additional comments or questions about the data, please direct them to help@nedcdata.org.

Best regards,

Sean Ferrell

NEDC Data Resources Development Manager

=====================================================================

Comparison to v1.2.1:

DEVELOPMENT TEST SET

Files and Sessions:

files: 1015 files containing seizures: 273

sessions: 239 sessions containing seizures: 102

patients: 50 patients containing seizures: 38

Signal Data:

seizures: 58,322.3671 secs ( 9.45%)

background: 558,779.6329 secs ( 90.55%)

total: 617,102.0000 secs (100.00%)

files w/seizures: 225,017.0000 secs ( 36.46% of the total data)

TRAINING SET

Files and Sessions:

files 1989 files containing seizures: 384

sessions 581 sessions containing seizures: 181

patients 264 patients containing seizures: 118

Signal Data:

seizures: 78,838.0892 secs ( 6.63%)

background: 1,109,474.9108 secs ( 93.37%)

total: 1,188,313.0000 secs (100.00%)

files w/seizures: 288,650.0000 secs ( 24.29% of the total data)