File: AAREADME.txt

Database: TUH EEG Seizure Corpus (TUSZ)

Version: 2.0.3

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

Change Log:

v2.0.3 (20250401): Corrected a few corrupted files to match TUEG

v2.0.3 (20240207): Headers were modified. No change to the signal data.

v2.0.2 (20240113): Removed duplicate montages for two sessions in /eval:

eval/aaaaaqvx/s003\_2015\_08\_24/

eval/aaaaaqvx/s010\_2015\_08\_27/

03\_tcp\_ar\_a was retained and 01\_tcp\_ar was deleted.

Added a seizure event for:

dev/aaaaadkj/s002\_2007\_10\_22/02\_tcp\_le

v2.0.1 (20231004): A few problems with the start and end times of seizure

events were corrected, including boundaries that

exceeded the end of the file over overlapped on the same

channel. Most of these were related to issues with the

annotator tool. Several short gaps between two adjacent

seizure events were removed. There are 35 files that

changed. These are listed at the bottom of this file.

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

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:

Shah, V., von Weltin, E., Lopez. S., McHugh, J., Veloso, L.,

Golmohammadi, M., Obeid, I., and Picone, J. (2018). The Temple University

Hospital Seizure Detection Corpus. Frontiers in Neuroinformatics. 12:83.

doi: 10.3389/fninf.2018.00083

This publication can be retrieved from:

https://www.isip.piconepress.com/publications/journals\_refereed/2018/frontiers\_neuroscience/tuh\_eeg\_seizure

Our preferred reference for the TUH EEG Corpus, from which this

seizure corpus was derived, is:

Obeid, I., & Picone, J. (2018). The Temple University Hospital EEG Data

Corpus. In Augmentation of Brain Function: Facts, Fiction and Controversy.

Volume I: Brain-Machine Interfaces (1st ed., pp. 394–398). Lausanne,

Switzerland: Frontiers Media S.A.

The data in this release was based on v2.0.3 of the TUH EEG Corpus.

There are three main directories in this release: train, dev and eval.

The train directory contains data you are allowed to use for the

development of your technology. The dev data is disjoint from the

training set and should only be used for testing. Eval is a blind evaluation

set - you should never optimize parameters on this set.

The top-level directories: edf/dev, edf/eval and edf/train. Please see

the documentation for TUH EEG v2.0.3 to understand how the data is

structured.

There are three types of files in this release (older formats have

been obsoleted):

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

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

\*.csv\_bi: term-based annotations using only two labels (bckg and seiz)

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 (seiz) and background

(bckg). The multi-class annotations use all available seizure

types. These are described in the spreadsheet:

$TUSZ/v2.0.3/DOCS/seizures\_types\_v02.xlsx

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).

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 csv 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.

Finally, here are some basic descriptive statistics about the data.

The commands used to generate these numbers are (/dev is used as an

example) shown below. For the commands below, the

starting point was here:

/data/isip/data/tuh\_eeg\_seizure/v2.0.3/edf

( 1) Number of files:

nedc\_130\_[1]: find . -name "\*.edf" | wc

7361 7361 513353

nedc\_130\_[1]: find ./train -name "\*.edf" | wc

4664 4664 309828

nedc\_130\_[1]: find ./dev -name "\*.edf" | wc

1832 1832 117738

nedc\_130\_[1]: find ./eval -name "\*.edf" | wc

865 865 56343

nedc\_130\_[1]: find . -name "\*.csv" | wc

7361 7361 483909

nedc\_130\_[1]: find . -name "\*.csv\_bi" | wc

7361 7361 505992

( 2) Number of sessions:

nedc\_130\_[1]: find \* -mindepth 3 -maxdepth 3 | wc

1643 1643 46837

nedc\_130\_[1]: find train -mindepth 2 -maxdepth 2 | wc

1175 1175 36425

nedc\_130\_[1]: find dev -mindepth 2 -maxdepth 2 | wc

342 342 9918

nedc\_130\_[1]: find eval -mindepth 2 -maxdepth 2 | wc

126 126 3780

( 3) Number of patients:

nedc\_130\_[1]: find train -mindepth 1 -maxdepth 1 | wc

579 579 8685

nedc\_130\_[1]: find dev -mindepth 1 -maxdepth 1 | wc

53 53 689

nedc\_130\_[1]: find eval -mindepth 1 -maxdepth 1 | wc

43 43 602

( 4) Number of files with seizures:

nedc\_130\_[1]: find train -name "\*.csv" -exec grep -H "sz," {} \; | cut -d"/" -f5 | cut -d":" -f1 | sort -u | wc

872 872 20056

nedc\_130\_[1]: find dev -name "\*.csv" -exec grep -H "sz," {} \; | cut -d"/" -f5 | cut -d":" -f1 | sort -u | wc

325 325 7475

nedc\_130\_[1]: find eval -name "\*.csv" -exec grep -H "sz," {} \; | cut -d"/" -f5 | cut -d":" -f1 | sort -u | wc

195 195 4485

( 5) Number of sessions with seizures:

nedc\_130\_[1]: find train -name "\*.csv" -exec grep -H "sz," {} \; | cut -d"/" -f2,3 | sort -u | wc

352 352 6688

nedc\_130\_[1]: find dev -name "\*.csv" -exec grep -H "sz," {} \; | cut -d"/" -f2,3 | sort -u | wc

114 114 2147

nedc\_130\_[1]: find eval -name "\*.csv" -exec grep -H "sz," {} \; | cut -d"/" -f2,3 | sort -u | wc

63 63 1197

( 6) Number of patients with seizures:

nedc\_130\_[1]: find train -name "\*.csv" -exec grep -H "sz," {} \; | cut -d"/" -f2 | sort -u | wc

208 208 1872

nedc\_130\_[1]: find dev -name "\*.csv" -exec grep -H "sz," {} \; | cut -d"/" -f2 | sort -u | wc

45 45 405

nedc\_130\_[1]: find eval -name "\*.csv" -exec grep -H "sz," {} \; | cut -d"/" -f2 | sort -u | wc

34 34 306

( 7) Total number of seizure events (measured using \*.csv\_bi):

nedc\_130\_[1]: find train -name "\*.csv\_bi" -exec grep -H seiz {} \; | wc

2421 2421 233397

nedc\_130\_[1]: find dev -name "\*.csv\_bi" -exec grep -H seiz {} \; | wc

1075 1075 101655

nedc\_130\_[1]: find eval -name "\*.csv\_bi" -exec grep -H seiz {} \; | wc

469 469 44540

( 8) Total duration:

nedc\_130\_[1]: find train -name "\*.csv" -exec grep duration {} \; | awk '{ sum+=$4} END {print sum}'

3277229

nedc\_130\_[1]: find dev -name "\*.csv" -exec grep duration {} \; | awk '{ sum+=$4} END {print sum}'

1567972

nedc\_130\_[1]: find eval -name "\*.csv" -exec grep duration {} \; | awk '{ sum+=$4} END {print sum}'

459713

( 9) Total size of the corpus (/train + /dev + /eval): 81,492 Mbytes (81.4 Gbytes)

nedc\_130\_[1]: cd /data/isip/data/tuh\_eeg\_seizure/

nedc\_130\_[1]: du -sBM v2.0.3

81537M v2.0.3

(10) Total duration of seizure events:

nedc\_130\_[1]: find train -name "\*.csv\_bi" -exec grep -H "seiz," {} \; | cut -d"," -f2,3 | sed -e "s/,/ /g" | awk '{ sum +=($2-$1)} END {print sum}'

175125

nedc\_130\_[1]: find dev -name "\*.csv\_bi" -exec grep -H "seiz," {} \; | cut -d"," -f2,3 | sed -e "s/,/ /g" | awk '{ sum +=($2-$1)} END {print sum}'

71310.6

nedc\_130\_[1]: find eval -name "\*.csv\_bi" -exec grep -H "seiz," {} \; | cut -d"," -f2,3 | sed -e "s/,/ /g" | awk '{ sum +=($2-$1)} END {print sum}'

27246.7

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

If you have any additional comments or questions about the data,

please direct them to help@nedcdata.org.

Best regards,

Joe Picone