**On the Use of Non-Experts for
Generation of High Quality Annotations of Seizure Events**

Vinit Shaha, Eva von Weltina, Tameem Ahsana, Iyad Obeida and Joseph Piconea

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

**Highlights:**

* A cost-effective method for efficiently creating large annotated datasets, which are critical to the development of machine learning technology, using undergraduate students is proposed.
* Intra-rater agreement for student annotators and inter-rater agreement with board-certified neurologists are high, indicating that a comparable quality of annotation can be achieved at a much lower cost and much faster turnaround time.
* This method was used to create a large publicly available seizure corpus known as the Temple University Hospital Seizure Detection Corpus (TUSZ).

***Abstract***

*Objective:* Engaging neurologists in the creation of seizure annotations with the level of detail necessary to conduct machine learning research is a slow, tedious and expensive process. This process is further complicated by inconsistent inter-rater agreement. The goal of this study was to demonstrate that undergraduate students can be trained to generate such data with acceptable levels of accuracy.

**Corresponding Author:** Vinit Shah, The Neural Engineering Data Consortium, ENGR 703A, Temple University, 1947 North 12th Street, Philadelphia, Pennsylvania, 19122, Tel: 215-204-4841, Fax: 215-204-5960, Email: vinitshah@temple.edu.

*Methods:* Student rater ability to create consistent, high quality annotations that meet the standards of expert raters is evaluated through a series of inter-rater agreement tests on the seizure sets collected from 3 different hospitals. Between 4 and 5 students are evaluated on a total of ~46 hours of data associated with 13 patients.

*Results:* Inter-rater agreement between neurologists and student annotators, using Cohen’s Kappa coefficient, is within the range of 0.53 – 1.00 indicating sufficiently strong agreement.

*Conclusions:* It is possible to create a database of high quality and of sufficient size to enable development of machine learning systems by engaging trained student annotators. This approach has proven to be more cost-effective and time-efficient than utilizing expert clinicians.

*Significance:* This approach has been used to produce the open source TUH EEG Seizure Detection Corpus, thereby enabling the application of sophisticated deep learning technologies requiring big data resources.

*Keywords*— electroencephalography, EEG, inter-rater agreement, machine learning

# Introduction

Electroencephalograms are the primary tool by which clinicians diagnose brain related illnesses such as epilepsy, non-epileptic seizures, and sleep disorders (Yamada and Meng, 2009). Seizures, which are seen most often in patients diagnosed with epilepsy, can occur in a convulsive or non-convulsive manner. In an ICU environment, approximately 90% of these seizures are clinically unrecognizable non-convulsive seizures which can only be diagnosed by continuous EEG (cEEG) monitoring (Hirsch, 2010). Though clinicians do periodically observe EEGs for the identification of such seizures, any delay in the treatment of non-convulsive seizures in ICU environments can be harmful and even deadly to patients (Hirsch and Kull, 2004; Wiebe, 2008). To aid in the speed and efficiency of the diagnosis and treatment process, automatic interpretation of EEGs has been studied over the past decade for application in clinically relevant software (Alotaiby et al., 2014; Gotman, 1982; Wilson et al., 2003). However, the development of such applications requires a large and, for many researchers, prohibitive amount of annotated EEG data.

Annotation of EEGs is ideally performed by certified neurologists who have received extensive training. In order to speed up the diagnosis process, experienced clinicians will rapidly skim through an EEG record and annotate any intervals in which interesting events occur using simple “start” and “stop” marks. As a result, clinicians will often miss some events, especially those that are subtle or brief. Additionally, these transcriptions cannot be directly used for technological development due to the insufficient level of spatial information conveyed by simple time stamp annotations. This annotation process is subjective and relies on clinical evidence including a record of push button events and of medication dosages. Due to inconsistencies in an individual’s judgement, poor inter-rater agreement (IRA) performance among neurologists is common on tasks such as detection of seizures and periodic discharges (Halford et al., 2015; Ronner et al., 2009).

Since the EEG is the primary tool used by neurologists for the diagnosis of neurological disorders, a significant portion of a typical neurologist’s professional life is assigned to EEG analysis. These professionals tend to have heavy clinical loads and little discretionary time to annotate data for research purposes. Contracting clinicians to create annotations using mechanisms such as Mechanical Turk has been unproductive and generally tends to be expensive. For example, in a recent NIH-funded research project on Cohort Retrieval, our attempts to hire such neurologists at rates of $75/hr. resulted in very little usable data. Furthermore, due to a lack of clear standards of interpretation and the creation of inexact annotations, agreement amongst these professionals was low, reflecting the dramatic differences in the way they annotate data. In this study, we show that it is possible to develop a large, standardized, and annotated dataset by training undergraduate student annotators in the interpretation of deidentified EEGs. This process is faster, notably less expensive, and can result in superior inter-rater reliability than can be achieved by employing neurologists in the development of a dataset of comparable size.

Publicly available annotated EEG databases are scarce and under-representative of the diverse population of patients seen in real world clinical settings. For example, one of the most prominent databases available is CHB-MIT (Goldberger et al., 2000) which contains only 23 subjects. Emerging deep learning algorithms require large amounts of training data to support the development of robust models. Progress has been limited by the lack of availability of truly open source resources. Using the methods described in this paper, we have developed one of the largest unencumbered open source repositories of annotated data.

The subsets used for this study are randomly sampled from datasets collected at Duke University (DU) (Haider et al., 2016) and Emory University (EU) (Swisher et al., 2015), both of which were collected for the study of qEEG, and from the Temple University Hospital EEG Seizure Corpus (TUSZ) (Shah et al., 2018). These three corpora were independently evaluated because the modalities in which these databases were collected such as data sources, collection period, specific neurologists who defined “gold-standard” reference annotations, purpose of study, annotator experience, etc. varied with each set. In this study we evaluate the performance of trained undergraduate student annotators on identification of seizure events that have been annotated by trained neurologists, as well as the evaluation of inter-rater agreement (IRA) performance on all the datasets.

An example of a neurologist’s detailed annotations that we collected for our Temple University IRA test is shown in . For the sake of comparison, we define “gold-standard” annotations as those created by a group of neurologists and “aggregate-standard” as annotations created by a group of undergraduate students. Gold-standard annotations are defined by the agreement among two or more neurologists. Individual expert annotators’ annotations are not considered as a standard. Aggregate-standard annotations have been defined by conducting group meetings and discussions to agree and establish consent on marked annotations by the student annotators.

# Method

In order to evaluate the usefulness and validity of engaging student raters in the creation of an annotated seizure database, performance by students must be compared to that of board certified neurologists on the same data. Students were evaluated on test sets compiled from 3 different universities, each of which were annotated by neurologists with their own process and standards for annotation of seizure events. Performance by each student annotator, in addition to an agreed-upon reference, were compared to the gold-standard annotations. Cohen’s Kappa statistic, generated by two different scoring metrics namely Any-overlap and Epoch scoring method, was used to evaluate each comparison.

## Evaluation Data

The records used in this study were collected from three EEG seizure data sources from three different medical institutions to ensure a variety of seizure data and of local standards for annotation by expert raters. Creating data for this type of study is a challenge because there is a practical limit to the amount of data that can be annotated by our expert raters. Therefore, these three subsets were constructed to be as rich as possible and to contain a reasonable distribution of the types of events such as subtle or brief ones that cause problems for our annotators. A subset which were refer to as TUSZ-IRA, which is taken from the TUSZ Corpus (ref, XXXX), contains pruned records from 5 subjects with a total duration of 25,940 seconds of multichannel EEG data.

A second subset, which we refer to as DUSZ-IRA, was created from the DU Corpus, also contained records from 5 patients consisting of 72,001 seconds of multichannel data. This data was extracted from continuous EEG (cEEG) records and was not pruned. A third similar subset, referred to as the EUSZ-IRA set, was constructed from the Emory University dataset. EUSZ-IRA contains records from 3 patients consisting of 66,530 seconds of multichannel data. Both datasets were collected from critically ill ICU patients. Expert annotations in EUSZ-IRA were found to be less consistent than those from the DUSZ-IRA, leading us to select fewer patients from Emory University than from Duke University for use in evaluative tests.

Each inter-rater agreement test was performed independently such that trained student annotators were provided a combination of ictal and non-ictal files in TUSZ-IRA and ictal-only cEEG files from DUSZ-IRA and EUSZ-IRA. The gold-standard annotations of the DUSZ-IRA and EUSZ-IRA subsets were generated by neurologists at these respective institutions. In the DUSZ-IRA set these gold-standard annotations represent an agreement met by two neurologists (Swisher et al., 2015). In the EUSZ-IRA set these gold-standard annotations were annotated independently by three different neurologists (Haider et al., 2016). The TUSZ-IRA was originally distributed to 23 neurologists, four of whom completed and returned their annotations within the suggested timeline. According to gold-standard seizure annotations, there were 157 seizures of diverse focality, morphology, and duration between all IRA subsets. The distribution of number of seizures based on their duration are shown in . The seizures collected from TUSZ were within the range of 1 to 5 minutes. The majority of seizures in the DUSZ distribution have a duration within 30 seconds to 3 minutes. The majority of EUSZ seizures were 1 to 3 minutes long.

It should be noted that our preliminary experiments on automated processing of this data (ref, XXXX) have shown that there are not significant differences in performance between these corpora. This led us to believe that evaluation on these three sets would be complementary.

## Non-expert Raters

All three sets were annotated by a group of 4 or 5 student employees who underwent 2-3 months of extensive training in the interpretation of EEGs and in the precise annotation of seizure events. Students began training early in their undergraduate career and came from several fields of study including neuroscience and biochemistry.

On the other hand, the board-certified EEG experts who participated in this study [... say something about the level of professional preparation of these people for TUSZ ... and assume the other studies used similar experts ...] have experience reviewing intensive care unit (ICU) EEGs rather than annotation for research purposes.

IRA tests were conducted on a bi-class basis, meaning that the only possible annotations used in these comparisons are ictal and non-ictal. Student raters have been trained to create multiclass annotations, indicating the type and spread of a seizure, but this aspect of annotation was not tested for purpose of this report.

Annotators were asked to simply mark onset and offset of the ictal events. No consideration was made to the start and end times of these events on individual channels. Channel-specific annotations are typically needed to support technology development for applications such as seizure detection (ref, XXXX). However, our clinical experts typically only mark events across all channels. We refer to this as a term-based annotation, and since our experts were only comfortable doing this, we used term-based annotations as the basis for this study.

Evaluations are made between student raters, between student raters and the aggregate-standard (generated via student rater discussion) annotations, between aggregate-standard annotations and gold-standard (neurologist) annotations, and, in the case of TUSZ-IRA, between gold-standard annotations. A total of 14 cEEG files (5 + 9 from DU and EU) and 32 pruned files (from the larger TUSZ corpus) were used for three sessions of IRA tests which, according to gold-standard annotations, contained a total seizure duration of 13,054 seconds (out of 164,471 total seconds). Though each annotator reviewed the EEG records independently, student annotators were allowed to use books, notes, and web resources as general references since this is how they normally do their work.

## Evaluation Metrics and Inter-Rater Agreement Analysis

Relative similarity between independently generated annotations can be evaluated using the Kappa statistic. We have used Cohen’s kappa coefficient (ref, XXXX) as our inter-rater agreement analysis tool. Cohen’s Kappa coefficients were calculated for each pair of raters at an intra-student level. The same statistic was used to evaluate IRA of the gold-standard and aggregate-standard annotations at an inter-expert level. The Cohen’s kappa coefficient can be calculated as:

 (1)

where is the relative observed agreement between raters (observed accuracy) and  is the hypothetical probability of chance agreement (expected accuracy). Values below 0 suggest no agreement, values in the range [0,0.20] indicate slight agreement, values in the range [0.20,0.60] indicate fair to moderate agreement, while values between [0.60,1.00] indicate substantial to complete agreement (Landis & Koch, 1977).

Intermediate variables in the calculation of Cohen’s Kappa such as observed accuracy and expected accuracy that are necessary for IRA analysis need to be computed from a measurement of the four types of classification errors: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Exactly how these quantities are estimated is an area of open research and discussed extensively in work by Ziyabari et al. (2018). We have estimated these quantities using two scoring metrics: Any-overlap (OVLP) scoring and Epoch based (EPOCH) scoring. The OVLP method considers a detection as any case of overlap between events. In the EPOCH scoring metric, a record gets divided into equally sized subsamples called an “epoch” (defined as 1 second for our analysis) and scoring is performed on each epoch independently. compares the functionality of the OVLP and EPOCH methods. The Kappa coefficient is calculated using each of these metrics and is cross-tabulated to represent intra-student and inter-expert analysis for all three seizure sets.

Identification of exact ictal onset and offset can be difficult in some ictal events, especially those which evolve very slowly. If the majority of a hypothesis annotation overlaps temporally with a reference annotation, give or take a few seconds at the very beginning and very end of the event, then this should be considered as a detection. OVLP addresses this concern by considering any overlap as a detection. This is in contrast to EPOCH scoring which evaluates agreement at every second. Though EPOCH scoring is not so forgiving of raters over or under annotating events, we have seen that, in comparisons on prolonged events with higher true detection scores (TP and TN), these detection scores will not be overwhelmed by the error scores (FP and FN).

It is worth noting that the EPOCH metric is biased towards longer events. Annotating the longest seizures will improve the results, through early or late onset and offset annotations can deteriorate the results. Though kappa statistics, including those created using the EPOCH method, are always symmetric, this is not the case when calculated using parameters generated by the OVLP method. presents an example of a seizure event in which the kappa statistic calculated from parameters generated using the OVLP method will be different depending upon which rater is taken to be the reference. In the first case of rater A versus rater B, three hypotheses annotations overlap with the first reference annotation, resulting in a single positive detection on that event. The alternate case of rater B versus rater A shows that a single hypothesis annotation overlaps with three reference annotations, resulting in three positive detections. For this reason, we always include full matrices when reporting kappa statistics generated using the OVLP method.

# Results

Performance of student raters on each of the three test sets was evaluated by comparing directly, annotations created by each student rater to those created by expert raters and by comparing aggregate-standard annotations created through student rater discussion to gold-standard neurologist annotations. In each set, any discrepancies seen between individual student annotators compared to gold-standard annotations are not reflected in aggregate-standard vs gold-standard comparisons; indicating the importance of discussion in the method for generation of high-quality annotations proposed herein.

## Temple University IRA Ser (TUSZ-IRA)

From the TUH EEG Seizure Detection Corpus, 5 patients containing 32 pruned (<1 hour long) files were selected with a total duration of 25,918 seconds for the TUSZ-IRA test. According to the most recent version of TUSZ (v1.2.0), there are 12 seizures in this subset. First, we consider the intra-expert neurologist’s annotations that were collected for the TUSZ-IRA subset. This test set was assigned to 4 neurologists. From 4 of these neurologists, one pair was assigned 14 identical files and another pair was assigned 18 identical files. Intra-expert level agreement of two pairs are shown in . The agreement between clinician annotator 1 (CAnn-1) and clinician annotator 2 (CAnn-2) on 14 files is almost perfect (~1) according to both OVLP and EPOCH methods. Similar evaluation on clinician annotator 3 (CAnn-3) and clinician annotator 4 (CAnn-4) shows very poor performance (~0.2) according to both IRA metrics. This is a result of CAnn-4 marking seizures parsimoniously whereas CAnn-3 marked seizures very generously.

TUSZ-IRA gold-standard annotations were established by taking aggregate markings of the intersecting annotations of neurologists for the first group. This approach wasn’t pragmatic for the second group due to low IRA. Here, clinician annotator CAnn-3’s annotations were taken as the ground-truth after review of these annotations by an independent expert.

 shows pairwise performance on the inter-expert level as gold-standard (Gold-St) vs. aggregate standard (Agg-St) as well on the intra-student level for student annotators (StAnn). The Gold-St vs. Agg-St pair shows almost perfect agreement between two groups with kappa values at 1.0 according to OVLP and 0.87 according to EPOCH. On the intra-student level, every StAnn pair’s IRA scores according to OVLP and EPOCH kappa values show substantial to perfect agreement. From the OVLP IRA results, it can be seen that StAnn1 and StAnn5 show perfect agreement on individual seizure events. Less agreement between StAnn1 and StAnn2 indicates that they have disagreement on seizure/non-seizure decisions as well as on the duration of seizures. Overall, the TUSZ-IRA test has the highest agreement among all the IRA tests conducted in this study.

##  Duke University IRA Test (DUSZ-IRA)

The Duke university IRA test set contains records related to 5 ICU-CEEG files associated with 5 patients for a total duration of 72,001 seconds. According to received gold-standard annotations from Duke University, these files contain 63 seizures for a total seizure duration of 5202 seconds. These gold-standard neurologist’s annotations were compared to 4 student raters’ annotations. Shows the pairwise agreement between individual student annotators, aggregate student annotations, and gold-standard annotations.

The IRA kappa values between individual student annotators and gold-standard annotations collected from DU range from moderate to substantial according to both scoring metrics where StAnn1 and StAnn4 show maximum agreement within the group as well as compared to gold-standard annotations. OVLP results generated DUSZ-IRA test has a very non-symmetric matrix which is representative of the example shown in . StAnn-2 and StAnn-3 performed poorly according to both gold-standard and aggregate-standard annotations.

## Emory University IRA Test (EUSZ-IRA)

The final subset in this study contained 9 ICU-CEEG files (duration: 66,530 seconds) related to 3 patients collected from Emory University Hospital. According to gold-standard annotations received from this institution, this subset contains 82 seizure events with a total duration of 5870 seconds. Gold-standard annotations of this test were agreed upon by at least three neurologists. shows the pairwise agreement between individual annotator pairs, aggregate-standard and gold-standard annotations.

IRA between two groups of experts on the EUSZ-IRA test was substantial to almost perfect. The majority of the annotations created by student annotators are in nearly perfect agreement with respect to both gold and aggregate standards.

## Statistical Inference

For statistical inference, direct comparison of supersets containing all gold-standard annotations to those containing aggregate or individual annotator pairs are considered. All tests are performed based on the EPOCH scoring metric with an epoch duration of 1 second. (top left) shows the histogram of kappa scores based on individual files. To understand the normality of the distribution we performed a one-sample Kolmogorov-Smirnov test which yields a KS value of 0.685 (p-value < 0.001), indicating that the distribution can be considered as Gaussian. This same distribution is shown as a boxplot in the top right corner of the figure. It can be seen that overall IRA on most files is in the substantial to perfect range with some outliers in the moderate range. Whiskers of this plot spread from 0.62 in the lower range to 0.96 in the upper range. The bottom left figure shows individual boxplots for each IRA subset. Here, the second and third quartiles are in almost perfect agreement range for TUSZ and EUSZ. The DUSZ distribution has a greater range with its median value around ~0.7 and its lower whisker spreading to 0.5.

Performing a one-way ANOVA test on all student annotators rejects the null hypothesis with an F-value of 1.42 with p-value 0.239 for sensitivity and an F-value of 0.49 with p-value 0.684 for specificity calculated by the EPOCH scoring metric. From (bottom right), it can be seen that Pearson’s correlation coefficient calculated on time series for individual annotator’s annotations are highly correlated with respect to gold-standard annotations (p-value < 0.001). Both tests suggest that student annotators’ performance is very similar to each other as no one outperforms any other rater when compared to the gold-standard. These consistent results are an important factor in the justification of development by this group of standards for seizure annotation. The two IRA metrics being used, OVLP and EPOCH, are correlated with a ρ-value of 0.58 with p-value of 0.002.

# Discussion

Continuous EEG (cEEG), which contains hours to multiple days’ worth of data, is becoming a vital tool for correctly diagnosing patients with epilepsy and other brain related diseases. Hours’ worth of recordings for each patient under a neurologist’s care amounts to a very large volume of data that then must to be evaluated for diagnostic markers. Due to this high volume of data and the strain on their time, neurologists are accustomed to skimming quickly through the record for prominent events and clear state changes. This process, however, can lead to the neurologists missing brief and/or low amplitude, yet still clinically relevant, events. Annotations created in this process are simple and cursory. They are not suited for use in a database intended for technology development. Instead, if trained students, especially those with an interest in neuroscience and neurology, are employed to make these transcriptions, the process can be improved in terms of speed, detail, consistency, and cost. This study also suggests that group discussions and meetings among student annotators can contribute significantly to the establishment of standards for annotations and can aide in quality-control the annotated database.

The significance of this study is exemplified in the hinging of decisions made by neurologists as to the diagnosis and severity of epilepsy in clinical settings upon the identification of seizures in a patient’s EEG record. Unfortunately, identification of seizures, even according to ACNS terminology, is very subjective and uses rules that cannot be applied universally such as whether frequency evolution of PLDs/GPDs appearing in long bursts, several post-status epilepticus stages, and low frequency (1-3 Hz) spike and wave discharges lasting for more than 10 seconds should be considered as seizures. These ambiguities can result in differing opinions even among top experts in the field. Moreover, even after consensus, establishment of agreement on onset/offset of these events is nearly impossible. Our annotation team at the Neural Engineering Data Consortium (NEDC) has been collecting and discussing difficult EEG records and sharing them online with experienced readers; this forum is open to the public and all are welcome to post their thoughts and analysis. This is a good way to establish consent on morphologies, which can tend towards biases when discussed within small groups. The link to our FAQ queries can be found at:
https://www.isip.piconepress.com/projects/tuh\_eeg/faq/index.shtml

It is unclear why agreement between student annotators and neurologists was relatively low in the EUSZ-IRA subset. This could be attributed to experts considering several low amplitude morphologies as clinically irrelevant. Despite this, student annotators have managed to maintain a consensus within the group (on intra-expert level). Compared to gold-standard annotations, student annotators’ mean sensitivity and specificity on all subsets are 80.77% and 97.14% respectively. Lower sensitivity compared to high specificity suggests that annotators were more inclined to avoid false alarms.

# Conclusion

This study is the result of a long-term effort by this group to develop an annotated seizure database of sufficient size, accuracy, and detail to facilitate the development of new technologies. We present this method to justify this effort and to bolster confidence in the quality and usefulness of this database and of the student annotators that have been essential in its creation. This study holds greater significance than many previously conducted IRA tests in neurology due to following reasons:
(1) All annotations were collected and annotated in a digital format; (2) The study was performed on routine EEG records (TUSZ) as well as cEEG records (DUSZ, EUSZ) from different institutions which combined are representative of real world clinical practices; (3) Instead of asking for absence or presence of ictal events in a particular segment, this study additionally emphasizes exact onset and offset of every single ictal event.(4) This study uses two separate scoring metrics, one on an event basis and other on an epoch basis, for detection of seizure events. This gives us two separate perspectives; specificity in terms of seizure duration, and base sensitivity to the event. (5) This study proposes a method for the continued and future generation of refined annotated data sources for use in the development of new clinical technology.

The tool our student annotators used to create, review, and edit annotations is called the “demo tool” which was developed by our team at the NEDC. This EEG viewing and annotation tool (Capp et al., 2017) is now available at no cost to the public and can be found and downloaded at
 https://www.isip.piconepress.com/projects/tuh\_eeg/downloads/nedc\_demo/

Out original intention was to create a larger volume of annotation data than what is currently present in the TUSZ-IRA test set. As previously mentioned, we distributed a portion of our Temple University Seizure Detection Corpus (TUSZ) to more than 20 neurologists but only 4 of them could deliver us annotations. This illustrates the difficulty of engaging expert neurophysiologists in such studies and in general for detailed annotation process. The time commitment necessary to create precise, detailed annotations is simply not conducive to the busy schedule that neurologists contend with.

Finally, this study emphasizes the vital role that a large, detailed, and affordable dataset must play in the path forward of rapid technology development, and the usefulness of student annotators in this pursuit. It in no way attempts to question or challenge the abilities of expert neurophysiologists.

# Acknowledgements

The authors would like to thank Dr. Mercedes Jacobson of Temple University Hospital and Dr. Steven Tobochnik of New York Presbyterian Hospital at Columbia University for enabling the creation of this corpus and their instruction in the interpretation of clinical EEG data.

# ReferenceS

Alotaiby, T., Alshebeili, S., Alshawi, T., Ahmad, I., & Abd El-Samie, F. (2014). EEG seizure detection and prediction algorithms: a survey. *EURASIP Journal on Advances in Signal Processing*, *2014*(1), 1–21. doi: 10.1186/1687-6180-2014-183.

Capp, N., Krome, E., Obeid, I., & Picone, J. (2017). Rapid Annotation of Seizure Events Using an Extensible Visualization Tool. In *IEEE Signal Processing in Medicine and Biology Symposium* (p. 1). Philadelphia, Pennsylvania, USA. doi: 10.1109/SPMB.2017.8257043.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., … Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation*, *101*(23), e215–e220. doi: 10.1161/01.CIR.101.23.e215.

Gotman, J. (1982). Automatic recognition of epileptic seizures in the EEG. *Electroencephalography and Clinical Neurophysiology*, *54*(5), 530–540. [doi](https://doi): 10.1016/0013-4694(82)90038-4.

Haider, H. A., Esteller, R. D., Hahn, C. B., Westover, M. J., Halford, J. W., Lee, J. M., & LaRoche, S. (2016). Sensitivity of quantitative EEG for seizure identification in the intensive care unit. *Neurology*, *87(9)*, 935–944. doi: 10.1212/WNL.0000000000003034.

Halford, J. J., Shiau, D., Desrochers, J. A., Kolls, B. J., Dean, B. C., Waters, C. G., … LaRoche, S. M. (2015). Inter-Rater Agreement on Identification of Electrographic Seizures and Periodic Discharges in ICU EEG Recordings. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *126*(9), 1661–1669. doi: 10.1016/j.clinph.2014.11.008.

Hirsch, L. (2010). Urgent Continuous EEG (cEEG) Monitoring Leads to Changes in Treatment in Half of Cases. *Epilepsy Currents*, *10(4)*, 82–85. doi: 10.1111/j.1535-7511.2010.01364.x.

Hirsch, L. J., & Kull, L. L. (2004). Continuous EEG monitoring in the intensive care unit. *American Journal of Electroneurodiagnostic Technology*, *44*(3), 137–158. doi: 10.1046/j.1528-1157.43.s.3.7.x.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33(1)*, 159–74. issn: 0006-341X.

Ronner, H. E., Ponten, S. C., C.J., S., & Uitdehaag, B. M. J. (2009). Inter-observer variability of the EEG diagnosis of seizures in comatose patients. *European Journal of Epilepsy*, *18(4)*, 257–263. doi: 10.1016/j.seizure.2008.10.010.

Shah, V., von Weltin, E., Lopez, S., McHugh, J. R., Veloso, L., Golmohammadi, M., … Picone, J. (2018). The TUH EEG Seizure Corpus. Submitted to *Frontiers in Neuroscience*. *https://www.isip.piconepress.com/publications/unpublished/journals/2018/frontiers\_neuroscience/tuh\_eeg\_seizure/*.

Swisher, C. B., White, C. R., Mace, B. E., & Dombrowski, K. E. (2015). Diagnostic Accuracy of Electrographic Seizure Detection by Neurophysiologists and Non-Neurophysiologists in the Adult ICU Using a Panel of Quantitative EEG Trends. *Journal of Clinical Neurophysiology*, *32*(4), 324–330. doi: 10.1097/WNP.0000000000000144.

Wiebe, S. (2008). Review of The Johns Hopkins atlas of digital EEG: An interactive training guide. *The Canadian Journal of Neurological Sciences / Le Journal Canadien Des Sciences Neurologiques*, *35*(2), 269. doi: 10.1001/archneurol.2007.44.

Wilson, S. B., Scheuer, M. L., Plummer, C., Young, B., & Pacia, S. (2003). Seizure detection: correlation of human experts. *Clinical Neurophysiology*, *114*(11), 2156–2164. doi: 10.1016/S1388-2457(03)00212-8

Yamada, T., & Meng, E. (2017). *Practical guide for clinical neurophysiologic testing: EEG*. (E. Meng & R. L. (Online Service), Eds.). Philadelphia, Pennsylvania, USA: Lippincott Williams & Wilkins. doi: 10.1111/j.1468-1331.2009.02936.x.

Ziyabari, S., Shah, V., Golmohammadi, M., Obeid, I., & Picone, J. (2018). Objective Evaluation Metrics for Automatic Classification of EEG Events. Submitted to the *Journal of Neural Engineering*. *https://www.isip.piconepress.com/publications/unpublished/journals/2018/iop\_jne/metrics/*.

# LIST OF TABLES

Table 3. Pairwise comparison between student annotators and gold-standard annotations on the DUSZ-IRA set

Table 4. Pairwise comparison between student annotators and gold-standard annotations on the EUSZ-IRA set

**Table 1. Neurologist’s pairwise agreements on TUSZ IRA subset**

|  |  |
| --- | --- |
| **OVLP** | **Hyp** |
| CAnn-1 | CAnn-2 |
| Ref | CAnn-1 | 1.000 | 1.000 |
| CAnn-2 | 1.000 | 1.000 |
| **EPOCH** | Hyp |
| CAnn-1 | CAnn-2 |
| Ref | CAnn-1 | 1.000 | 0.98 |
| CAnn-2 | 0.98 | 1.000 |
| **OVLP** | Hyp |
| CAnn-3 | CAnn-4 |
| Ref | CAnn-3 | 1.000 | 0.2 |
| CAnn-4 | 0.2 | 1.000 |
| **EPOCH** | Hyp |
| CAnn-3 | CAnn-4 |
| Ref | CAnn-3 | 1.000 | 0.159 |
| CAnn-4 | 0.159 | 1.000 |

**Table 2. TUSZ-IRA test's pairwise kappa value between student annotators and gold-standard annotation**

|  |  |
| --- | --- |
| **OVLP** | **Hyp** |
| Gold-St | Agg-St | AtAnn1 | StAnn2 | StAnn3 | StAnn4 | StAnn5 |
| Ref | Gold-St | 1.000 | 1.000 | 0.949 | 0.837 | 0.815 | 0.945 | 0.949 |
| Agg-St | 1.000 | 1.000 | 0.949 | 0.837 | 0.815 | 0.945 | 0.949 |
| StAnn1 | 0.949 | 0.949 | 1.000 | 0.895 | 0.865 | 0.895 | 1.000 |
| StAnn2 | 0.836 | 0.836 | 0.894 | 1.000 | 0.762 | 0.886 | 0.894 |
| StAnn3 | 0.818 | 0.818 | 0.867 | 0.767 | 1.000 | 0.767 | 0.867 |
| StAnn4 | 0.945 | 0.945 | 0.894 | 0.886 | 0.762 | 1.000 | 0.894 |
| StAnn5 | 0.949 | 0.949 | 1.000 | 0.895 | 0.865 | 0.895 | 1.000 |

|  |  |
| --- | --- |
| **EPOCH** | **Hyp** |
| Gold-St | Agg-St | AtAnn1 | StAnn2 | StAnn3 | StAnn4 | StAnn5 |
| Ref | Gold-St |  | 0.870 | 0.829 | 0.727 | 0.881 | 0.831 | 0.886 |
| Agg-St |  |  | 0.812 | 0.828 | 0.933 | 0.846 | 0.932 |
| StAnn1 |  |  |  | 0.764 | 0.856 | 0.824 | 0.869 |
| StAnn2 |  |  |  |  | 0.835 | 0.858 | 0.832 |
| StAnn3 |  |  |  |  |  | 0.869 | 0.959 |
| StAnn4 |  |  |  |  |  |  | 0.865 |
| StAnn5 |  |  |  |  |  |  |  |

**Table 3. Pairwise comparison between student annotators and gold-standard annotations on the DUSZ-IRA set**

|  |  |
| --- | --- |
| **OVLP** | **Hyp** |
| Gold-St | Agg-St | AtAnn1 | StAnn2 | StAnn3 | StAnn4 |
| Ref | Gold-St | 1.000 | 0.729 | 0.729 | 0.327 | 0.419 | 0.766 |
| Agg-St | 0.745 | 1.000 | 0.974 | 0.458 | 0.556 | 0.862 |
| StAnn1 | 0.745 | 0.974 | 1.000 | 0.487 | 0.586 | 0.862 |
| StAnn2 | 0.428 | 0.536 | 0.570 | 1.000 | 0.582 | 0.556 |
| StAnn3 | 0.468 | 0.575 | 0.605 | 0.512 | 1.000 | 0.574 |
| StAnn4 | 0.780 | 0.861 | 0.861 | 0.475 | 0.555 | 1.000 |

|  |  |
| --- | --- |
| **EPOCH** | **Hyp** |
| Gold-St | Agg-St | AtAnn1 | StAnn2 | StAnn3 | StAnn4 |
| Ref | Gold-St |  | 0.736 | 0.686 | 0.531 | 0.576 | 0.708 |
| Agg-St |  |  | 0.921 | 0.627 | 0.679 | 0.833 |
| StAnn1 |  |  |  | 0.641 | 0.664 | 0.826 |
| StAnn2 |  |  |  |  | 0.657 | 0.616 |
| StAnn3 |  |  |  |  |  | 0.647 |
| StAnn4 |  |  |  |  |  |  |

**Table 4. Pairwise comparison between student annotators and gold-standard annotations on the EUSZ-IRA set**

|  |  |
| --- | --- |
| **OVLP** | **Hyp** |
| Gold-St | Agg-St | AtAnn1 | StAnn2 | StAnn3 | StAnn4 |
| Ref | Gold-St | 1.000 | 0.834 | 0.798 | 0.833 | 0.823 | 0.852 |
| Agg-St | 0.838 | 1.000 | 0.923 | 0.882 | 0.967 | 0.912 |
| StAnn1 | 0.796 | 0.920 | 1.000 | 0.865 | 0.909 | 0.895 |
| StAnn2 | 0.838 | 0.882 | 0.869 | 1.000 | 0.882 | 0.890 |
| StAnn3 | 0.837 | 0.969 | 0.917 | 0.888 | 1.000 | 0.906 |
| StAnn4 | 0.849 | 0.908 | 0.894 | 0.884 | 0.896 | 1.000 |

|  |  |
| --- | --- |
| **EPOCH** | **Hyp** |
| Gold-St | Agg-St | AtAnn1 | StAnn2 | StAnn3 | StAnn4 |
| Ref | Gold-St |  | 0.779 | 0.752 | 0.740 | 0.739 | 0.758 |
| Agg-St |  |  | 0.967 | 0.825 | 0.825 | 0.855 |
| StAnn1 |  |  |  | 0.818 | 0.827 | 0.836 |
| StAnn2 |  |  |  |  | 0.823 | 0.853 |
| StAnn3 |  |  |  |  |  | 0.848 |
| StAnn4 |  |  |  |  |  |  |

# LIST OF Figures

Seizure Duration Distribution of the TUSZ, DUSZ, and EUSZ IRA sets.

Fig. 5. Student Annotators IRA Results Comparison with Neurologists’ Gold Standard Annotations Based on the EPOCH Scoring Metric

# Figures



Figure . Example of a neurologist’s descriptive annotation of TUSZ data



Figure 2 Seizure duration distribution of TUSZ, DUSZ and DUSZ IRA subsets





Figure Functionality of Overlap Scoring metric (Top) and Epoch scoring metric (Bottom)





Figure An example of inconsistent parameters generated by OVLP scoring method





Figure Student annotators IRA results comparison with neurologists’ gold standard annotations based on EPOCH scoring metric