**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 that is further complicated by inconsistent inter-rater agreement. In this study, we 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:* The ability of undergraduate students 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. A small team of student annotators 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 high-quality database by engaging trained student annotators. This approach has proven to be significantly 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. Community feedback on the accuracy of the annotations has been overwhelmingly positive.

*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 & 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 or deadly to patients (Hirsch & Kull, 2004; Wiebe, 2008). To aid in the speed and efficiency of the diagnosis and treatment process, the development of automatic interpretation technology using state of the art machine learning approaches has been of great interest (Alotaiby et al., 2014; Gotman, 1982; Wilson et al., 2003). However, the development of such technology requires a large amount of annotated EEG data.

Annotations would ideally be performed by certified neurologists who have received extensive clinical 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. Clinicians often choose not to annotate some events deemed not clinically important, especially those that are subtle or brief. This is somewhat a result of their need to minimize the amount of time spent reading EEGs due to heavy caseloads. Unfortunately, such broad annotations are not useful for technology development due to the insufficient level of detail (e.g., machine learning systems typically need all relevant events marked on individual channels). Further, this annotation process is subjective and relies on clinical evidence including a record of push button events and of medication dosages. Not surprisingly, poor inter-rater agreement (IRA) among neurologists is common on tasks such as detection of seizures and periodic discharges (Halford et al., 2015; Ronner et al., 2009).

Since an EEG is still the primary tool used by neurologists for the diagnosis of neurological disorders, a significant portion of a typical neurologist’s professional life is spent interpreting EEGs. 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. Our participants indicated payment wasn’t the issue, but rather that finding time to do the work was the biggest challenge. We solicited a large number of medical practices internationally but received little usable data in return. Furthermore, due to a lack of clear standards of interpretation, agreement amongst these professionals was low, reflecting the dramatic differences in the way they annotate data.

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 complex models. Progress has been limited due to the lack of large open source corpora to support this type of research. To address this, we have used the methods described in this paper to develop one of the largest unencumbered open source repositories of annotated EEG data. We have over 2,000 subscribers to this resource and have received extremely positive feedback about the quality and quantity of the data (see *https://www.isip.piconepress.com/projects/tuh\_eeg* for more details).

In this study we evaluate the performance of undergraduate student annotators on identification of seizure events and show that it is possible to develop a large, standardized, and annotated dataset in a cost-effective manner. This process is faster, notably less expensive, and can result in superior inter-rater reliability. This work seeks to challenge the prevailing perspective that non-clinicians are incapable of adequately annotating EEG data. Instead, we demonstrate that properly trained undergraduates can in fact do such annotations at acceptable levels of accuracy.

# Method

In order to evaluate the accuracy of our student annotators, performance must be compared to that of board-certified neurologists. We have a long history of managing group annotation projects and conducting inter-rater reliability studies for other applications such as speech and image recognition (Houser et al., 2018; Deshmukh et al., 1998; Hamaker et al., 1998). Students were evaluated on test sets compiled from three different seizure detection corpora, each of which had been annotated by experts. The subsets used for this study were randomly sampled from EEG corpora collected at Duke University (DUSZ) (Swisher et al., 2015), Emory University (EUSZ) (Haider et al., 2016) and Temple University (TUSZ) (Shah et al., 2018). The first two corpora were originally developed to study the efficacy of qEEGs for ICU patients. The third corpus is an ongoing development that was the motivation for this study on inter-rater agreement. All three sets use digital scalp EEG recordings with electrode placements according to international 10-20 system.

These three corpora were selected because of their prominence in the literature and the variability of methods used to collect and annotate the data. Performance by each student annotator was compared to the reference annotations as well as other student annotators using Cohen’s Kappa statistic. For the first two corpora, we also had data on inter-rater agreement between expert annotators. The Kappa statistic was calculated using the results from two different methods for quantifying differences in annotations: the “any-overlap” and “epoch” metrics. A detailed analysis of these methods can be found in Shah et al. (2019).

It is important to understand the type of annotation being used for this study since the concept of a detailed annotation varies across the community. To make it easy for experts, we allowed them to annotate images of the data. These handwritten notes were converted to digital information manually. An example of this process is shown in . The annotation we received from an expert annotator is shown along with the conversion of this information to a digital format. Allowing experts to annotate data using handwritten notes alleviated the need for the experts to learn a new tool or use an uncomfortable interface. Note that the annotators annotated each channel separately rather than making a global decision about an epoch. We refer to these types of annotations as channel-specific. This level of detail in an annotation is important for machine learning systems that use supervised learning to train their model parameters.

## Evaluation Data

The amount of data for this type of study is limited to what can be annotated by expert raters. Consequently, there are limits on the variety of conditions that can be sampled in such small data sets. A data set was created by selecting a subset of the evaluation data from the TUSZ Corpus (Shah et al., 2018). We refer to this as the TUSZ-IRA subset. It contains pruned EEG records from 5 subjects. There are 32 pruned files in this study with a total duration of 25,940 secs. There are 12 seizure events in this data set. Expert raters for this set were provided EEG screenshots of 10-second windows of the data, which was prepared using a popular bipolar double banana montage known as a Temporal Central Parasagittal (TCP) montage (Ferrell et al., 2019).

A second subset, which we refer to as DUSZ-IRA, was created from DUSZ by selecting data from 5 patients with a total duration of 72,001 secs. There are 63 seizure events in this data set. This data was extracted from continuous EEG (cEEG) records and was not pruned. A third subset, referred to as the EUSZ-IRA set, was selected from EUSZ and contains data from 3 patients with a total duration of 66,530 secs. There are 82 seizure events in this data. Both DUSZ and EUSZ were collected from critically ill patients in an intensive care unit (ICU).

Each inter-rater agreement test was performed independently. 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. These were blind tests ­– the students did not know they were annotating evaluation data during these tests. It was mixed in with their normal daily work. The gold-standard annotations for the DUSZ-IRA and EUSZ-IRA subsets were generated by neurologists at their respective institutions. In the DUSZ-IRA set, these gold-standard annotations represent an agreement between 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).

TUSZ-IRA was originally distributed to 23 neurologists, four of whom completed and returned their annotations. The data was originally arranged so that at least three neurologists would independently annotate each segment. Due to the low response rate, gold-standard annotations had to be constructed from segments that had been annotated by two or more neurologists. We also had our senior annotator and subject matter expert, who did not participate in this study, review these annotations. Analysis of these annotations revealed 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 a range of 1 to 5 minutes. The majority of seizures in DUSZ have a duration between 30 seconds and 3 minutes. The majority of EUSZ seizures were 1 to 3 minutes long.

## Non-expert Raters

These three data sets were annotated by a small group of five students who underwent two to three 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. All generally had some sort of STEM experience, but none had ever annotated data before. Training began with annotation of data for which we had accurate annotations (data that had been manually annotated by our experts and senior staff). Once students were performing acceptably on this data, we continually monitored their performance using a cross-validation approach (a small percentage of data was blindly annotated by all the students and compared). Students were given feedback and also participated in periodic group training sessions in which our experts reviewed their errors.

Not every student hired performs well as an annotator. After a few weeks we have a good idea of whether they will succeed at this task. After a couple of months (working 10 hours per week), they have usually become fairly accurate at this task. Within six months they usually reach an expert level of performance and will serve as trainers for a new generation of students.

Manually annotated data that is released for research has usually been annotated by at least two annotators and reviewed by an expert project leader. The total time spent on a 30-minute EEG file is about 15 minutes per annotator at a cost of about $10-$12/hour. When we include the time spent reviewing and organizing the data for release, we spend about 60 minutes of time per 30-minute EEG file, so the total cost is about $0.33/minute of data. Annotating a large corpus such as TUSZ averages about $1/minute when you include training costs and other overhead expenses.

These cost estimates are an approximation based on the average time spent on the files by the annotators. Some patients with severe forms of epilepsy (i.e. Lennox-Gastaut syndrome) are much harder to annotate and require multiple reviews and group discussion before reaching a consensus. On the other hand, normal records or milder forms of epilepsy patients (i.e. absence epilepsy) are much faster to annotate. Longer EEGs generally are faster to annotate because once baseline morphologies of the record are observed, it is easy to spot events of interest.

The board-certified EEG experts who participated in this study had completed their neurology residency, passed the neurology boards and had years of experience in the field. More importantly, these neurologists had extensive experience reviewing ICU EEGs rather than annotation for research purposes.

IRA tests were conducted using a two-class system­ – ictal and non-ictal. Student annotators had been trained to create multiclass annotations in which they note the type and spread of a seizure, but this type of annotation was not tested in this study.

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 extremely useful for machine learning systems. However, our clinical experts typically only mark events across all channels. We refer to this as a “term-based annotation.” Since our experts were only comfortable doing this, we used term-based annotations as the basis for this study.

The analysis we present in this study includes comparisons between students, comparison to a consensus generated through group discussions between the students, and comparison to the expert gold-standard. A total of 14 cEEG files (5 + 9 from DU and EU) and 32 pruned files (from the TUSZ corpus) were used for three sessions of IRA tests which 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

The similarity between two annotations is most often evaluated using the Kappa statistic. We have used Cohen’s Kappa coefficient (McHugh M. L., 2012) as our inter-rater agreement analysis metric. A Kappa coefficient was calculated for each pair of student annotators. The same statistic was used to evaluate IRA between the gold-standard and consensus annotations. The Cohen’s Kappa coefficient can be calculated as:

$κ= \frac{p\_{0}-p\_{e}}{1-p\_{e}},$ (1)

where $p\_{o} $is the relative observed agreement between raters (observed accuracy) and $p\_{e}$ 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 coefficient such as observed accuracy and expected accuracy were computed from a measurement of the four types of classifications: 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, see Shah et al. (2019). We have estimated these quantities using two common scoring metrics: any-overlap (OVLP) and epoch-based (EPOCH) scoring. OVLP considers a detection as any amount of overlap between events. EPOCH divides a record into equally sized subsamples called an “epoch” (defined as 1 second for our analysis) and scoring is performed on each epoch independently. compares the OVLP and EPOCH methods. The Kappa coefficient is calculated for both 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 events, especially those which evolve slowly. If the majority of a hypothesis annotation overlaps temporally with a reference annotation, then this event is considered as a correct detection. (In practice, guard bands must be used because these overlaps will be off by tens of milliseconds due to the imprecision of the manual annotation process.) OVLP addresses this concern by considering any overlap as a detection. This is in contrast to EPOCH scoring which evaluates agreement at every 1-second epoch. OVLP is event based; EPOCH scoring is time-alignment based. Therefore, these two metrics span the range of possible scoring metrics described in Shah et al. (2019).

It is worth noting that the EPOCH metric is biased towards longer events. Correct annotation of the longest seizures will deliver the greatest improvement to the overall results. Though Kappa statistics, including those created using the EPOCH method, are typically symmetric, this is not the case when calculated using parameters 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 overlap with the first reference annotation, resulting in a single positive detection for that event. The alternate case of rater B versus rater A shows that a single hypothesis annotation overlaps with three reference events, resulting in three positive detections. For this reason, we always include full matrices when reporting Kappa statistics generated using the OVLP method.

Finally, we explore a variety of scoring metrics in Shah et al. (2019). The trends are similar for all of these metrics. With respect to seizure detection, we have not seen great variation in the Kappa statistic when using these different metrics.

# Results

In this section, we present the results of the IRA analysis for the three corpora previously described, beginning with TUSZ-IRA. According to the v1.2.0 annotations of TUSZ, there are 12 seizures in the subset of data selected. This set was assigned to four neurologists. Two of them were assigned 14 identical files while the other two were assigned 18 identical files. IRA for these two pairs is shown in . The agreement between clinical annotator 1 (C1) and annotator 2 (C2) on 14 files is almost perfect ($κ\~1)$ according to both OVLP and EPOCH methods. Similar evaluation of annotator 3 (C3) and 4 (C4) shows a very poor degree of agreement ($κ\~0.2) $according to both scoring metrics. This is a result of the fact that these annotators had significantly different approaches: C3 marked seizures very generously while C4 marked seizures very conservatively.

Note that TUSZ-IRA gold-standard annotations were established by constructing a consensus using the annotations from C1 and C2. In contrast, this approach couldn’t be used for the files reviewed by C3 and C4 due to the low IRA between them. Instead, the annotations were reviewed by an independent expert who concluded that C3’s annotations were mostly accurate and could serve as the basis for a gold-standard, with only minor changes.

 shows the pairwise agreement between the gold-standard (Gold) vs. the consensus (Cons) vs. the student annotators (St1 – St5). The Gold vs. Cons pair shows almost perfect agreement between the two groups with Kappa values at 1.0 according to OVLP and 0.87 according to EPOCH. Agreement between students is also fairly high. From the OVLP IRA results, it can be seen that St1 and St5 show perfect agreement on individual seizure events. There was less agreement between St1 and St2. An analysis of this indicated that there was some disagreement on seizure/non-seizure decisions as well as on duration.

Next, we provide an analysis for DUSZ in which shows the pairwise agreement between various combinations of annotator groups. The IRA Kappa values between individual student annotators and the gold standard range from moderate to substantial according to both scoring metrics. St1 and St4 show maximum agreement within the group and compare most favorably with the gold-standard annotations. OVLP results generated for DUSZ-IRA are non-symmetric due to the type of misalignment described in . St2 and St3 performed poorly according to both the gold-standard and the consensus annotations.

In we present IRA results for the EUSZ data set. The EUSZ corpus was created from the patients admitted between 2008 and 2010 for treatment at Emory University Hospital. A majority of these records contain widespread periodic discharges and rhythmic delta activities along with electrographic seizures. According to Haider et al. (2016), the data was challenging to annotate. However, our student annotators haves performed well during our IRA tests. Student annotators performed at the same level of agreement according to both gold-standard and consensus annotations.

We also conducted an analysis of statistical significance. Direct comparison of sets containing all gold standard annotations to those containing consensus or individual annotator pairs were 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 $(p < 0.001)$. Both tests suggest that student annotators’ performance is 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) is a vital tool for correctly diagnosing patients with epilepsy and other brain related diseases. These recordings, ranging in duration from hours to days, contain an overwhelming amount of data that requires manual interpretation. Neurologists are accustomed to quickly skimming 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. Cursory annotations created in this way are not suited for use in the development of advanced machine learning technology. Trained undergraduate students obviously provide a much lower cost option for generating these annotations. This study demonstrates that a group of well-trained students can annotate EEG data with an acceptable level of accuracy.

Identification of seizures, even according to ACNS guidelines, is subjective and uses rules that cannot be applied universally. For example, some common rules advocated in the ACNS guidelines are vague in regards to (1) whether the frequency evolution of periodic lateralized discharges (PLDs) and generalized periodic discharges (GPDs) appearing in long bursts, (2) several post-status epilepticus stages and “gray zone” Ictal Interictal Continuum (IIC) phases, and (3) whether low frequency (1-3 Hz) spike and wave discharges lasting for more than 10 seconds, should be considered as seizures. These rules can result in annotation differences, even among experts. Moreover, even after there is consensus on an event, establishing an agreement on the precise onset/offset of these events is quite difficult.

One integral part of our work to reach consensus on such ambiguous data has been to develop an FAQ where examples can be discussed. This forum is open to the public and all are welcome to post their thoughts and analysis. This is a good way to establish consistency in our annotations. The link to our FAQ is:
*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. 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 deep learning technology. We present this work to document the process used to validate the performance of our annotators. This study is significant because the study was performed on a variety of EEG records from different institutions. Further, we have assessed the ability of our annotators to accurately identify the onset and offset of every ictal event. We have used two separate scoring metrics as well to provide alternate perspectives on the results.

Our 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 four of them could deliver annotations. This illustrates the difficulty of engaging expert neurophysiologists in such studies. The time commitment necessary to create precise, detailed annotations is simply prohibitive for experts who carry heavy clinical workloads. We have demonstrated that undergraduate students can accurately annotate EEG data at a cost that is typically 10x lower than what it would take to hire clinical experts to do this work.

The tool our student annotators used to create, review, and edit annotations is available as open source code at *https://www.isip.piconepress.com/projects/tuh\_eeg/downloads/nedc\_demo/* (Capp et al., 2017). The TU data described in this study is available at *https://www.isip.piconepress.com/projects/tuh\_eeg* as well. These rapidly growing resources are having a significant impact on our ability to research and develop automatic interpretation technology.

# Acknowledgements

Research reported in this publication was 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 material is also based in part upon work supported by the National Science Foundation under Grant No. IIP-1827565. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

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. We would also like to thank Dr. Jocelyn Cheng of NYU Langone Comprehensive Epilepsy Center, Dr. Dipti Zutshi of Wayne University, Dr. Maromi Nei of Jefferson Comprehensive Epilepsy Center and Dr. Colin Ellis from Hospital of the University of Pennsylvania for serving as expert annotators.

# 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.

Ferrell, S., Mathew, V., Ahsan, T., & Picone, J. (2019). The Temple University Hospital EEG Corpus: Electrode Location and Channel Labels. Philadelphia, Pennsylvania, USA. *https://www.isip.piconepress.com/publications/reports/2019/tuh\_eeg/electrodes*.

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.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276–282. *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/*.

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., Golmohammadi, M., Obeid, I., & Picone, J. (2018). Objective evaluation metrics for automatic classification of EEG events. *Journal of Neural Engineering*, 1–21. *https://www.isip.piconepress.com/publications/unpublished/journals/2018/iop\_jne/metrics/* (in review).

Shah, V., von Weltin, E., Lopez, S., McHugh, J. R., Veloso, L., Golmohammadi, M., … Picone, J. (2018). The Temple University Hospital Seizure Detection Corpus. *Frontiers in Neuroinformatics*, *12*, 83. *https://doi.org/10.3389/fninf.2018.00083*.

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*.

# LIST OF TABLES

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

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

|  |  |
| --- | --- |
| **OVLP** | **Hyp** |
| C1 | C2 |
| Ref | C1 | 1.000 | 1.000 |
| C2 | 1.000 | 1.000 |
| **EPOCH** | Hyp |
| C1 | C2 |
| Ref | C1 | 1.000 | 0.98 |
| C2 | 0.98 | 1.000 |
| **OVLP** | Hyp |
| C3 | C4 |
| Ref | C3 | 1.000 | 0.2 |
| C4 | 0.2 | 1.000 |
| **EPOCH** | Hyp |
| C3 | C4 |
| Ref | C3 | 1.000 | 0.159 |
| C4 | 0.159 | 1.000 |

Table . Kappa values for the TUSZ-IRA subset

|  |  |
| --- | --- |
| **OVLP** | **Hyp** |
| Gold | Cons | St1 | St2 | St3 | St4 | St5 |
| Ref | Gold | 1.000 | 1.000 | 0.949 | 0.837 | 0.815 | 0.945 | 0.949 |
| Cons | 1.000 | 1.000 | 0.949 | 0.837 | 0.815 | 0.945 | 0.949 |
| St1 | 0.949 | 0.949 | 1.000 | 0.895 | 0.865 | 0.895 | 1.000 |
| St2 | 0.836 | 0.836 | 0.894 | 1.000 | 0.762 | 0.886 | 0.894 |
| St3 | 0.818 | 0.818 | 0.867 | 0.767 | 1.000 | 0.767 | 0.867 |
| St4 | 0.945 | 0.945 | 0.894 | 0.886 | 0.762 | 1.000 | 0.894 |
| St5 | 0.949 | 0.949 | 1.000 | 0.895 | 0.865 | 0.895 | 1.000 |

|  |  |
| --- | --- |
| **EPOCH** | **Hyp** |
| Gold | Cons | St1 | St2 | St3 | St4 | St5 |
| Ref | Gold |  | 0.870 | 0.829 | 0.727 | 0.881 | 0.831 | 0.886 |
| Cons |  |  | 0.812 | 0.828 | 0.933 | 0.846 | 0.932 |
| St1 |  |  |  | 0.764 | 0.856 | 0.824 | 0.869 |
| St2 |  |  |  |  | 0.835 | 0.858 | 0.832 |
| St3 |  |  |  |  |  | 0.869 | 0.959 |
| St4 |  |  |  |  |  |  | 0.865 |
| St5 |  |  |  |  |  |  |  |

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

|  |  |
| --- | --- |
| **OVLP** | **Hyp** |
| Gold | Cons | St1 | St2 | St3 | St4 |
| Ref | Gold | 1.000 | 0.729 | 0.729 | 0.327 | 0.419 | 0.766 |
| Cons | 0.745 | 1.000 | 0.974 | 0.458 | 0.556 | 0.862 |
| St1 | 0.745 | 0.974 | 1.000 | 0.487 | 0.586 | 0.862 |
| St2 | 0.428 | 0.536 | 0.570 | 1.000 | 0.582 | 0.556 |
| St3 | 0.468 | 0.575 | 0.605 | 0.512 | 1.000 | 0.574 |
| St4 | 0.780 | 0.861 | 0.861 | 0.475 | 0.555 | 1.000 |

|  |  |
| --- | --- |
| **EPOCH** | **Hyp** |
| Gold | Cons | St1 | St2 | St3 | St4 |
| Ref | Gold |  | 0.736 | 0.686 | 0.531 | 0.576 | 0.708 |
| Cons |  |  | 0.921 | 0.627 | 0.679 | 0.833 |
| St1 |  |  |  | 0.641 | 0.664 | 0.826 |
| St2 |  |  |  |  | 0.657 | 0.616 |
| St3 |  |  |  |  |  | 0.647 |
| St4 |  |  |  |  |  |  |

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

|  |  |
| --- | --- |
| **OVLP** | **Hyp** |
| Gold | Cons | St1 | St2 | St3 | St4 |
| Ref | Gold | 1.000 | 0.834 | 0.798 | 0.833 | 0.823 | 0.852 |
| Cons | 0.838 | 1.000 | 0.923 | 0.882 | 0.967 | 0.912 |
| St1 | 0.796 | 0.920 | 1.000 | 0.865 | 0.909 | 0.895 |
| St2 | 0.838 | 0.882 | 0.869 | 1.000 | 0.882 | 0.890 |
| St3 | 0.837 | 0.969 | 0.917 | 0.888 | 1.000 | 0.906 |
| St4 | 0.849 | 0.908 | 0.894 | 0.884 | 0.896 | 1.000 |

|  |  |
| --- | --- |
| **EPOCH** | **Hyp** |
| Gold | Cons | St1 | St2 | St3 | St4 |
| Ref | Gold |  | 0.779 | 0.752 | 0.740 | 0.739 | 0.758 |
| Cons |  |  | 0.967 | 0.825 | 0.825 | 0.855 |
| St1 |  |  |  | 0.818 | 0.827 | 0.836 |
| St2 |  |  |  |  | 0.823 | 0.853 |
| St3 |  |  |  |  |  | 0.848 |
| St4 |  |  |  |  |  |  |

# LIST OF Figures

Figure 1. Example of a neurologist’s descriptive annotation

Figure 2. Seizure duration distribution for the TUSZ, DUSZ and DUSZ IRA subsets

Figure 3. A demonstration of the Overlap (top) and Epoch (bottom)

Figure 4. An example of inconsistencies in the OVLP scoring metric

Figure 5. Student annotators’ IRA results

# Figures





Figure . Example of a neurologist’s descriptive annotation (Top) and student annotators’ channel specific annotation (Bottom) of TUSZ dataset



Figure . Seizure duration distribution for the TUSZ, DUSZ and DUSZ IRA subsets





Figure . A demonstration of the Overlap (top) and Epoch (bottom) scoring metrics





Figure . An example of inconsistencies in the OVLP scoring metric





Figure . Student annotators’ IRA results