Q1: (ARTICLE TYPE) Does this manuscript conform to the definition below of Data Report articles? If not, please contact the Frontiers Editorial Office (editorial.office@frontiersin.org).

Data Reports present a description of research data sets. They should include the methods used to collect the data and information on how readers may interpret the dataset. In contrast, they should not contain detailed analyses or new scientific insights. They should also include a link for direct access to the data.

Reviewer 1 | 06 Mar 2018 | 23:22

#1

Yes

Add comment

MANUSCRIPT LENGTH

**Comment:** We have provided links to the database. That will be further explained in Q14, where the reviewer mistakenly concluded the database wasn’t publicly available. Over 1,000 sites have correctly accessed the data through our web site.

Q 2

Data Report articles should not exceed 3,000 words and may contain no more than two figure/table. Do you deem that any part of the manuscript should be shortened? If so, please specify.

Reviewer 1 | 06 Mar 2018 | 23:22

#1

Yes

The manuscript contains four figures but two are duplication. The two should be removed. In addition, regarding the figures there are some issues:

1. the information presented by the figures is limited and using tables or other forms of data representation, authors could have been able to present more information on the databases. for instances, the

2. abbreviations used in the figures should be written in full terms in the captions. for insstance the disease types.

**Comment:** The manuscript contains two figures – not four figures. It is compliant with the guidelines for the journal provided to authors. The first figure represents data on the number of events of each type of seizure. The second figure represents some aggregate statistics on the age and duration of the data files. These are not duplicates but completely different things.

The abbreviations used in these figures are explained in the text. Due to space and word count limitations, we have to be concise about these things.

LANGUAGE AND GRAMMAR

Q 3

Is the language, specifically the grammar, of sufficient quality?

Reviewer 1 | 06 Mar 2018 | 23:22

#1

No

Manuscript needs language revising and editing. Some instances of badly constructed sentences:

1. The term "TUH EEG Seizure Corpus (TUSZ)", the abbreviation is correct? what does TUH stand for?

2. punctuation need revising and polishing. for instance, please use comma before words like while, respectively, ...

**Comment:** The reviewer is suggesting stylistic changes that are a matter of personal preference. The current manuscript passes Microsoft Word’s grammar checker and follows language conventions normally found in professional journals.

The acronym TUSZ has been used for some time to represent this data as it was being developed. It is widely recognized in the community and has not been a problem for over 1,000 users. All our data products for better or for worse use 4-letter acronyms. “TU” represents “Temple University” – a widely used and acknowledged acronym. “SZ” is short for “SeiZure”. We are not sure why the reviewer has a problem with this since it is fairly clear where this acronym comes from. It is just an acronym. As long as it is defined in the text there is no problem.

Q 4

Should the authors send this manuscript to an expert in English editing and scientific writing?

Reviewer 1 | 06 Mar 2018 | 23:22

#1

No

**Comment:** No corrections needed.

Q 5

Do the authors use standardized vocabulary or ontologies to describe the data?

No answer given.

Add comment

**Comment:** No corrections needed.

RELEVANCE

Q 6

Is the presented data relevant to the field of research of the Journal?

Reviewer 1 | 06 Mar 2018 | 23:22

#1

Yes

**Comment:** No corrections needed.

S2. REQUIRED INFORMATION

DATASET

Q 7

Does the Data Report include the name of the data set, the name of the database/repository where the dataset has been submitted and a link to the data set?

Reviewer 1 | 06 Mar 2018 | 23:22

#1

No

in introduction the name of dataset is given as "TUH EEG Seizure Corpus (TUSZ)". the dataset name TUSZ seems not to comply with the full term.

**Comment:** We explained above where the name, which is an acronym, comes from.

Q 8

Do the authors describe in sufficient details how the data was acquired?

Reviewer 1 | 06 Mar 2018 | 23:22

#1

No

at line 62 "All three algorithms were tested on 100 reports (50 ictal and 50 non-ictal), with NegEx performing slightly better than the Naïve Bayes and SVM classifiers." why authors did not report the performance (sensitivity and specificity )of the three methods and compare them? reporting the values would give a better understanding to readers.

In this regard, why authors compared the performance of the three techniques with 100 cases as it is very low to yield a reliable comparison of the performance.

line 69 regarding the accuracy of NegEx, authors write the approach correctly classified 99% of the reports in the pilot study. However, the next sentence says "NegEx technique identified 844 sessions of as seizure onset event, but the actual ones or manual annotation is 174 sessions' It says that the difference between the calculated numbers (844) with the actual events (174) is high, then how the accuracy was 99%? please advise.

How authors evaluated the accuracy of the three methods of AutoEEG, NegEx, Persyst software for seizure event?

line 76, "This system detected seizures with high confidence in 1,466 files out of 31,645 files" please report the values?

line 77 "Files for which both systems agreed on a seizure were given the highest priority for annotation". which both techniques?

in line 101, "If there was insufficient evidence to classify the type of seizure, then an event was defined as either

102 “generalized non-specific” or “focal non-specific” depending on the focality". How focality was determined?

**Comment:** This paper is not intended to be a scientific study of the accuracy of these NLP techniques. There are several important reasons for this including (1) we did not attempt to develop sophisticated algorithms; (2) we do not have the entire database annotated so that we could evaluate the accuracy of these algorithms in terms of quantities like precision and recall; and (3) we used off the shelf technology for these algorithms.

We included this information in the paper because many people have asked us about how effective techniques to triage the data using NLP techniques was in finding files that have seizures. Seizures occur so infrequently you have to be very intelligent about what files you choose to annotate. The results quotes are stated as being very informal and intended to give readers a rough estimate of how effective these techniques were. We make it very clear this is not the main focus of this paper.

We clearly explain that the algorithms gave high performance on a small set of 100 reports. However, these are clearly not representative of the entire data set. Since we manually annotated a large chunk of the data, we further explain that the performance on the larger data set was much lower than 99%. This is to be expected since the 100 report set is small. We provide this information simply because it is informative for the readers to know how well these off the shelf algorithms perform. It is not intended to be a scientific study of this topic. In the end, the only effective way to ensure the integrity of the data was to manually annotate it.

Q 9

Do the authors provide information on the data collection period?

Reviewer 1 | 06 Mar 2018 | 23:22

#1

No

The time period is not reported in the manuscript.

**Comment:** We explained in the paper that this is an ongoing data collection process. We explained that the data was extracted from the first release of the TUH EEG Corpus, which covers the years 2002-2015. We then explained the strategy for culling data from this time period.

Q 10

The Data Report should include an overview of the data files and their formats. Has this information been included?

Reviewer 1 | 06 Mar 2018 | 23:22

#1

Yes

**Comment:** No corrections needed.

Q 11

Does the Data Report include clear information on how readers may access and interpret the data set as well as reuse the data?

Reviewer 1 | 06 Mar 2018 | 23:22

#1

No

I tried to download the dataset but failed.

also there is not database named "TUSZ or TUH EEG Seizure Corpus

There are several data bases in the given link and for all of them user and pass is required. There was no link or instruction in the page for readers on how to register to receive the dataset.

**Comment:** We explained that the data is available from this page:

<https://www.isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml>

Users are requested to sign up so that we can track the number of users and provide them with update information. The sign up process is automatic. Once you are signed up, users have access to a variety of sets of the data. Instructions how to sign up are provided on the page – go to this link:

<https://www.isip.piconepress.com/projects/tuh_eeg/html/request_access.shtml>

TUSZ is available here as noted in the paper:

<https://www.isip.piconepress.com/projects/tuh_eeg/downloads/tuh_eeg_seizure/>

When you access this link you will be prompted for a username and password that you receive when you sign up. The process is completely automated ­– there is no review of the information. This is the only database of this scope and size that is openly available this way.

We have had over 1,000 sites successfully complete this process. If the reviewer has any problems following these instructions, please contact us.

REFERENCES

Q 12

Is prior relevant work properly and fully cited?

Reviewer 1 | 06 Mar 2018 | 23:22

#1

No

Some sentences need citations:

at line 33 "Previous attempts to employ panels of experts or use crowdsourcing strategies were not productive (ref ?).

at line 39 .... o those of expert neurologists and shown to have a high inter-rater agreement (ref?)

**Comment:** We have added citations for this.

Q 13

Are references to the protocols or methods used to collect the data included?

Reviewer 1 | 06 Mar 2018 | 23:22

#1

No

in the introduction and in the Method sections, some of the references to the referred methods or protocols in this study are not presented,

**Comment:** It is unclear what the reviewer is specifically referring to. We have added citations to hopefully makes things more clear.

Add comment

S3. OTHER COMMENTS

Q 14

Please add here any further comments on this manuscript.

Reviewer 1 | 06 Mar 2018 | 23:22

#1

A main question on the overall performance of this dataset to be used in any ML or AI techniques:

In line 112: authors wrote "The training and evaluation sets contain 265 and 50 subjects respectively. In ML or AI we usually divide the training and evaluation set as 70 to 30 percent to reach a reliable outcome. in this database, authors seem not to observe this criterion, although the training set is higher than evaluation set. The evaluation set should contain a greater portion. Also, how authors ensured the training and evaluation sets contain the good representative of the whole data with a reliable distribution?

**Comment:** Again, we must disagree with the reviewer. The split between training and evaluation is often very arbitrary and discipline specific. For the past 35 years we have used 80/20, 50/50 and many other splits in a wide range of well-publicized scientific studies. The exact split actually depends on the number of seizure events in the files and the types of events. We have tried to reasonably balance these things, but this is not a perfect science, because the nature of the seizures vary widely.

We can tell you that the results we have achieved using this data are holding up on different partitions of this data, new data we have recently acquired, and studies done on other databases (e.g., the Emery and Duke datasets). So we are very comfortable with the partitioning was done.

I tried to download the dataset but failed.

also there is not database named "TUSZ or TUH EEG Seizure Corpus

There are several data bases in the given link and for all of them user and pass is required. There was no link or instruction in the page for readers on how to register to receive the dataset.

The manuscript needs major revision.

**Comment:** This was addressed in Q11. We will be happy to help the reviewer locate the correct pages and information. The database is clearly documented on these pages.