We would like to thank the reviewers for their thoughtful and careful reading of the paper. Our point-by-point responses are below.

REFEREE REPORT(S):   
Referee: 1   
  
**Major:**

* **Can the authors please address why overfitting is not an issue?  In particular, how was training and testing done exactly? I see that there are the "Train" and "Eval" sets in Table 1.  It seems that the authors used all the "Train" data and then tested only once on the "Eval" data with each method.  I think the paper would be much stronger and more robust if training and testing sets were done many times with cross-validation so that one could know how well each metric and classification system works across lots of iterations of cross-validation.  This would allow one to make error bars and compute statistical significance between the different methods, which would lend a much more complete view of differences in performance.  If cross-validation isn't possible to do (due to time constraints or some other thing I'm misunderstanding about the dataset), it would help greatly for the authors to address this issue directly in the paper.**
* We agree if this paper were about machine learning, measuring performance using a cross-validation approach would be important. However, this paper is about how to score a given set of results. We have selected a variety of systems for this analysis so there are a diversity of results to consider. The scoring process, however, is deterministic, so cross-validation type approaches aren’t really relevant. We do agree that it is useful to have a variety of results that exercise all the special cases that an algorithm might need to see. We have tried to get other researchers to evaluate their systems on this data and share results. That is always difficult as people are still overly protective of their results. But we can tell you that many several sites we collaborate with closely are using this software and finding its performance consistent with our findings and useful to their research.

We have also helped people use this software on other applications (e.g., acoustic signal processing) recently and the results have also been consistent with our findings.

* The TUH EEG Seizure dataset was developed in a way that evaluation set was considered as a held-out set. The evaluation set was designed to be a good representation of the diversity of EEG morphologies and patient types. This includes seizure types, male to female ratio, age etc.. Results on the evaluation set have always been consistent with the dev test and training data. A good example of this can be found here:

<https://www.isip.piconepress.com/publications/unpublished/books/2019/springer/ieee_spmb_2018/dpath/paper_v13.pdf>

The optimization process for the scoring algorithm has been performed on a wide range of data sets. It has not been tuned to the specific evaluation set shown in this paper. In fact, the parameters that need to be adjusted are fairly basic and are adjusted based on subject matter expertise and common sense ­­– not to optimize performance on an evaluation set.

* The focus of this study was to introduce a need for a standardized scoring process and to discuss the pros and cons of specific popular scoring approaches. The machine learning work done here has already been published. A good summary of that work is here:

<https://www.isip.piconepress.com/publications/book_sections/2019/springer/deep_learning/paper_v19.pdf>

In previous reviews we were asked to summarize the systems. To accommodate these reviewers, we present these models very briefly just to give reader an idea of the model architecture. We didn’t intend this to be a major focus of the paper. We mention this in the second last paragraph of Section 3: “Comprehensive details about the architectures are available in [40][41]. The details of these systems are not critical to this study.”

* An important point in this paper is that the comparison of performance with these metrics can be used to understand the strengths and weakness of an algorithm. We talk about his briefly in the paper.

**Minor:**

* **The figure captions should be much more verbose.  As is, they are very short and it isn't always evident what is happening in the figure (e.g. Figures 9 and 10).**
* We have updated the captions to be more descriptive in the updated version. This is somewhat of a stylistic issue ­– we usually use long captions but felt this journal prefers short captions with the descriptions in the body of the document. When publications use a two-column format it make it easier if captions are short.
* **Please describe in more detail how the "signal is converted to a sequence of feature vectors."  What exactly does each entry in a vector represent?  How long is the sequence?  How many timepoints in the EEG signal does one feature vector represent?**
* The feature extraction process is not really a focus of this paper. We did add some information explaining this. The process is completely described in the references we provide to this work.
* **The link to the code does not work:**

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

* To access this link, and all of the resources described in this paper, you must register here:

https://www.isip.piconepress.com/projects/tuh\_eeg/html/downloads.shtml

An automated process will send you a username and password that you can use to access the resources. Our sponsors require that we track users of these resources for obvious reasons. We currently have over 2,000 registered users.

* **It might help the reader to bold or italicize the cells with the best performance in the various tables of performance.**
* This is not as straightforward as it might seem when you consider all the metrics. The best-performing architecture in terms of absolute numbers tends to depend somewhat on the scoring metric used at a particular operating point. For example, consider a comparison between CNN-MLP and HMM-SdA systems for the ATWV and TAES metrics. Performance of both systems is almost equal (with HMM-SdA possibly performing slightly better) according to ATWV but on the other hand, CNN-MLP’s performance using TAES is significantly better. It is easy to judge CNN-MLP as a better performing system than HMM-SdA but only considering TAES.
* Other tables where these comparisons are more informative, such as the correlation between scoring metrics, have been updated based on reviewer’s suggestions.
* **Is there any reason why ROC AUC isn't used for the other metrics (e.g. ATWV) in Table 3?  It would be nice to explain the reason for only showing the ROC area for OVLP and TAES metrics in the paper (unless the reason is already there and I missed it), or to actually include these comparisons also in Table 3, if possible.**
* We have only provided results for AUC for the OVLP and TAES metrics because OVLP is the most popular metric in the community for classification of the annotated EEG events. Our scoring software actually produces AUC for all the metrics represented.
* We have tried to make a point in this paper that DET curves are still important. This is particularly true for EEG analysis because the false alarm rate is so crucial, and performance in the region of low false alarms is unpredictable. Depending on the operating point chosen, the ranking of algorithms can change significantly. AUC averages across all operating points, which means it will be dominated by operating points that are not often relevant to clinical applications. We actually look at AUC over a limited region where the false alarm rate is low. Since the TAES metric is new, we felt contrasting OVLP and TAES was appropriate. Adding AUC for the other metrics does not really add much to the paper in our opinion.

REFEREE REPORT(S):   
Referee: 2   
  
**The manuscript makes extensive comparisons with another scoring methods and implements different machine learning methods. However there is very little  discussion how this particular methodology will improve a real world clinical applications. It is not  improving real time detection of seizure.**

* Real time seizure detection is not the focus of the study here. Development of ML algorithms is an iterative process. Other fields, such as speech recognition, have for years been conducting common evaluations using standardized scoring. The bioengineering field has not yet embraced this approach. We are attempting to change this, and this paper is a strong first step in that direction. Making the software freely available also helps.
* The techniques we have presented can be used to gain insight into an algorithm’s performance.
* It is also important to note that often the differences in algorithm performance are in the noise compared to the differences between scoring algorithms. So, if researcher X has implemented algorithm Y and scored it using algorithm Z, it is difficult for another researcher to directly compare results unless they have access to the specific scoring algorithm used (and the data of course).

**The automatic classification may help offline review of patient recordings by epileptologist but for this application analysis very important is how the method trained on different patients seizures would perform on new patients. Presented in this papers methodology may be useful in comparing machine learning algorithms but with a little relation to practical application in neuroengineering.**

* Again, the focus of this paper is scoring, not technology development.
* This is a data related issue. We have processed almost two decades of clinical EEG recordings to develop the TUH EEG Seizure Corpus. In our current database, there are 642 patients. These patients should cover most morphologies that have been studied in neurology. If ML algorithms can learn base features of such morphologies, they can be easily be applied on new patients. In fact, the training and evaluation sets used in our study, which are publicly available, are split in a way that there is no overlap between patients. This is an important issue to understand ­– the subject pool for the evaluation set is disjoint from the training data. We have followed this approach for decades.