**The TUH EEG Seizure Corpus**

M. Golmohammadi1, V. Shah2, S. Lopez2, S. Ziyabari2, S. Yang2, J. Camaratta1, I. Obeid2 and J. Picone2

1. Biosignal Analytics, Inc., Philadelphia, PA, 19104

2. The Neural Engineering Data Consortium, Temple University, Philadelphia, PA, 19122

Corresponding Author: meysam@temple.edu

**Introduction:** Manual review of an EEG by a neurologist is time-consuming and tedious. Automatic seizure detection can reduce the time to diagnosis and enhance real-time applications such as ICU monitoring. These applications demand extremely low false alarm rates. Existing commercial systems perform poorly in critical care settings. The lack of big data resources that can be used to train sophisticated statistical models compounds the problem. A major goal of this study was to generate a large annotated corpus of seizure events that can support state of the art machine learning technology.

**Methods:** Using the TUH EEG Corpus, the world’s largest open-source clinical EEG corpus, we implemented a semi-automated strategy to label seizures by: (1) EEG reports were parsed using natural language processing techniques to locate sessions most likely to contain seizures. (2) Two seizure detection tools (Persyst and AutoEEG) were used to identify sessions with seizures. (3) Sessions, where both tools agreed with high confidence, were studied and divided into comprehensive training and evaluation subsets. (4) These subsets were manually annotated by a group of experts based on ACNS guideline.

**Results:** The current dataset includes 50 patients comprising 235 sessions for evaluation and 56 patients comprising 342 sessions for training. A hybrid machine learning system was developed on this data using a combination of hidden Markov models (HMMs) for sequential decoding and deep learning for postprocessing. Our system produced a sensitivity greater than 90% while maintaining a specificity below 10%.

Conclusions: The clinical use of existing seizure detection tools is limited due to poor performance, specifically a high false alarm rate. The existence of the seizure detection subset of the TUH EEG Corpus provides for the first time a sufficient amount of data to apply powerful machine learning algorithms. As a result, performance is now approaching that required for clinical acceptance.