Type of the Paper: Article

Objective evaluation metrics for automatic classification of EEG events

Vinit Shah [1], Meysam Golmohammadi [1], Iyad Obeid [1] and Joseph Picone [1, *

[1] The Neural Engineering Data Consortium, Temple University 1947 North 12th Street, Philadelphia, Pennsylvania, 19122, USA.

*   Correspondence: joseph.picone@gmail.com; Tel.: 215-204-4841

Abstract:

Background: Common quantitative scalar evaluation metrics such as sensitivity and specificity can often be misleading for sequential data applications. There is a lack of standardization of scoring metrics in biomedical applications involving sequential data.

Methods: We analyze three popular scoring metrics and introduce two new metrics for a seizure detection task. We compare and contrast their results collected from machine learning models developed on the TUH EEG Corpus. Performance evaluated in terms of sensitivity and specificity does not address the time scales over which the scoring must occur. This is critical for sequential data applications. To address these issues, our proposed metrics introduce two classes of scoring metrics: term-based and time-aligned. We also compare these metrics using a more holistic view based on a Detection Error Trade-off curve.

Results: We show that time-aligned scoring is consistent with popular scoring approaches but provides more accurate assessments and diagnostics by comparing the degree of match. We also evaluate existing metrics adapted from the speech recognition community, where sequential scoring techniques are very mature.

Conclusion: The metrics proposed in the study are excellent candidates for standardizing scoring across the industry. These metrics will be used in several upcoming open-source seizure detection challenges.

Keywords: electroencephalograms; EEG; machine learning; evaluation metrics; scoring

1. Introduction

Electroencephalograms (EEGs) are the primary means by which physicians diagnose, evaluate and manage brain-related illnesses such as epilepsy, seizures and sleep disorders 1. Automatic interpretation of EEGs has been extensively studied in the past decade 2-6. However, even though many researchers report impressive levels of accuracy in publications, widespread adoption of commercial technology has yet to happen in clinical settings primarily due to the high false alarm (FA) rates of these systems 789. In this paper, we investigate the gap in performance between research and commercial technology and discuss how these perceptions are influenced by a lack of a standardized scoring methodology.

37 There are in general two types of ways to evaluate machine learning technology: user acceptance
38 testing 1011 and objective performance metrics based on annotated reference data 1213. User
39 acceptance testing is time-consuming and expensive. It has never been a practical way to guide
40 technology development because algorithm developers need rapid turnaround times on evaluations.
41 Hence evaluations using objective performance metrics, such as sensitivity and specificity, are
42 common in the machine learning field 141516. With this approach, it is very important to have a rich
43 evaluation dataset and a performance metric that correlates well with user and application needs.
44 The metric must have a certain level of granularity so that small differences in algorithms can be
45 investigated and parameter optimizations can be evaluated. For example, in speech recognition
46 applications, word error rate has been used for many years because it correlates well with user
47 acceptance testing but provides the necessary level of granularity to guide technology development.
48 Despite many years of research focused on finding better performance metrics 1718, word error rate
49 remains a valid metric for technology development and assessment.
50 Sequential pattern recognition applications, such as speech recognition, keyword search or EEG
51 analysis, require additional considerations. Data, typically organized in files on a computer, are not
52 simply assessed with an overall judgment (e.g., "did a seizure occur somewhere in this file?").
53 Instead, the locality of the hypothesis must be considered – to what extent did the start and end times
54 of the hypothesis match the reference transcription. This is a complex issue since a hypothesis can
55 partially overlap with the reference annotation, and a consistent mechanism for scoring such events
56 must be adopted. Unfortunately, there is no such standardization in the EEG literature. For example,
57 Wilson et al. 19 advocates using a term-based metric involving sensitivity and specificity. A term was
58 defined as a connection of consecutive decisions from the same type of event. A hypothesis is counted
59 as a true positive when it overlaps with one or more reference annotations. A false positive
60 corresponds to an event in which a hypothesis annotation does not overlap with any of the reference
61 annotations. Kelly et al. 20 recommends using a metric that measures sensitivity and FAs. A
62 hypothesis is considered a true positive when the time of detection is within two minutes of the
63 seizure onset. Otherwise it is considered a false positive. Baldassano et al. 21 uses an epoch-based
64 metric that measures false positive and negative rates as well as latency. The development, evaluation
65 and ranking of various machine learning approaches is highly dependent on the choice of a metric.
66 A large class of bioengineering problems, including seizure detection, involve prediction as well as
67 classification. In prediction problems, we are often concerned with how far in advance of an event
68 (or after the event has occurred) we can predict an outcome. Accuracy of prediction varies with
69 latency, so this type of performance evaluation adds some complexity to the process. Winterhalder
70 et al. 22 have studied this problem extensively and argue for a scoring based on long-term
71 considerations. In this paper, we are not concerned with these types of prediction problems. We are
72 focused mainly on assessing the accuracy of classification and assessing the proximity of these
73 classifications to the actual event.
74 Therefore, we analyze several popular scoring metrics and discuss their strengths and weaknesses
75 on sequential decoding problems. We introduce several alternatives, such as the Actual Term-
76 Weighted Value (ATWV) 2324 that have proven successful in other fields, and discuss their relevance
77 to EEG applications. We present a comparison of performance for several systems using these metrics
78 and discuss how this correlates with overall user acceptance.

79   **2. Materials and Methods**

80   Researchers in biomedical fields typically report performance in terms of sensitivity and specificity
81   25. In a two-class classification problem such as seizure detection, we can define four types of errors:

82   True Positives (TP):   the number of 'positives' detected correctly
83   True Negatives (TN):   the number of 'negatives' detected correctly
84   False Positives (FP):   the number of 'negatives' detected as 'positives'
85   False Negatives (FN):   the number of 'positives' detected as 'negatives'

86   Sensitivity (TP/(TP+FN)) and specificity (TN/(TN+FP)) are derived from these quantities. There are a
87   large number of auxiliary measures that can be calculated from these four basic quantities that are
88   used extensively in the literature. These are summarized concisely in 26. For example, in information
89   retrieval applications, systems are often evaluated using accuracy ((TP+TN)/(TP+FN+TN+FP)),
90   precision   (TP/(TP+FP)),   recall   (another   term   for   sensitivity)   and   F1   score
91   ((2·Precision·Recall)/(Precision + Recall)). However, none of these measures address the time scale on
92   which the scoring must occur, which is critical in the interpretation of these measures for many real-
93   time bioengineering applications.
94   In some applications, it is preferable to score every unit of time. With multichannel signals, such as
95   EEGs, scoring for each channel for each unit of time might be appropriate since significant events
96   such as seizures occur on a subset of the channels present in the signal. However, it is more common
97   in the literature to simply score a summary decision per unit of time that is based on an aggregation
98   of the per-channel inputs (e.g., a majority vote). We refer to this type of scoring as epoch-based 2728.
99   An alternative, that is more common in speech and image recognition applications, is term-
100   based 2429, in which we consider the start and stop time of the event, and each event identified in
101   the reference annotation is counted once. There are fundamental differences between the two
102   conventions. For example, one event containing many epochs will count more heavily in an epoch-
103   based scoring scenario. Epoch-based scoring generally weights the duration of an event more heavily
104   since each unit of time is assessed independently.
105   Time-aligned scoring is essential to sequential decoding problems. But to implement such scoring in
106   a meaningful way, there needs to be universal agreement on how to assess overlap between the
107   reference and the hypothesis. For example, Figure 1 demonstrates a typical issue in scoring. The
108   machine learning system correctly detected 5 seconds of a 10-sec event. Essentially 50% of the event
109   is correctly detected, but how that is reflected in the scoring depends on the specific metric. Epoch-
110   based scoring with an epoch duration of 1 sec would count 5 FN errors and 5 TP detections. Term-
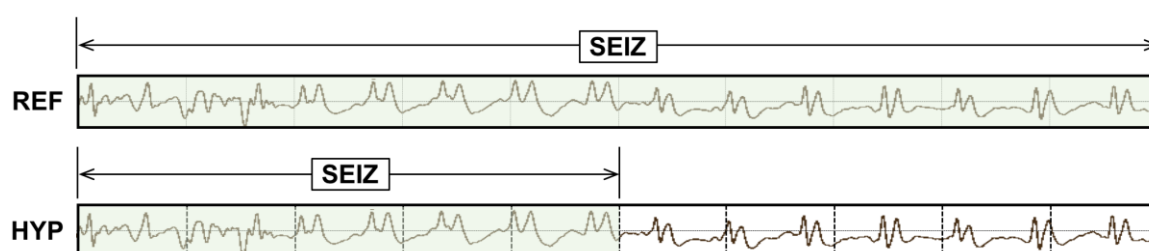

Figure 1. A typical situation where a hypothesis (HYP) has a 50% overlap with the reference (REF).

111   based scoring would potentially count this as a correct recognition depending on the way overlaps
112   are scored.

113   Term-based metrics score on an event basis and do not count individual frames. A typical approach
114   for calculating errors in term-based scoring is the Any-Overlap Method (OVLP) 3031. TPs are counted
115   when the hypothesis overlaps with reference annotation. FPs correspond to situations in which a
116   hypothesis does not overlap with the reference. The metric ignores the duration of the term in the
117   reference annotation. In Figure 2, we demonstrate two extreme cases for which the OVLP metric fails.
118   In each case, 90% of the event is incorrectly scored. In example 1, the system does not detect
119   approximately 9 seconds of a seizure event, while in example 2, the system incorrectly labels an
120   additional 9 seconds of time as seizure. OVLP is considered a very permissive way of scoring,
121   resulting in artificially high sensitivities. In Figure 2, the OVLP metric will score both examples as
122   100% TP.

123   It is very difficult to compare the performance of various systems when only two values are reported
124   (e.g. sensitivity and specificity) and when the prior probabilities vary significantly (in seizure
125   detection, the a priori probability of a seizure is very low, which means assessment of background
126   events dominate the error calculations). Often a more holistic view is preferred, such as a Receiver
127   Operating Characteristic (ROC) 15 or a Detection Error Trade-off (DET) curve 16. An ROC curve
128   displays the TP rate as a function of the FP rate while a DET curve displays the FN rate as a function
129   of the FP rate. When a single metric is preferred, the area under an ROC curve (AUC) 3233 is also an
130   effective way of comparing the performance. A random guessing approach to classification,
131   assuming equal priors for each class, will give an AUC of 0.5 while a perfect classifier will give an
132   AUC of 1.0.

133   The proper balance between sensitivity and FA rate is often application specific and has been studied
134   extensively in a number of research communities. For example, evaluation of voice keyword search
135   technology was carefully studied in the Spoken Term Detection (STD) evaluations conducted by
136   NIST 232434. These evaluations resulted in the introduction of a single metric, ATWV, to address
137   concerns about tradeoffs for the different types of errors that occur in voice keyword search systems.
138   Despite being popular in the voice processing community, ATWV has not been used in the
139   bioengineering community.

140   Therefore, in this paper, we compare and contrast five popular scoring metrics and one derived
141   measure:

EXAMPLE 1
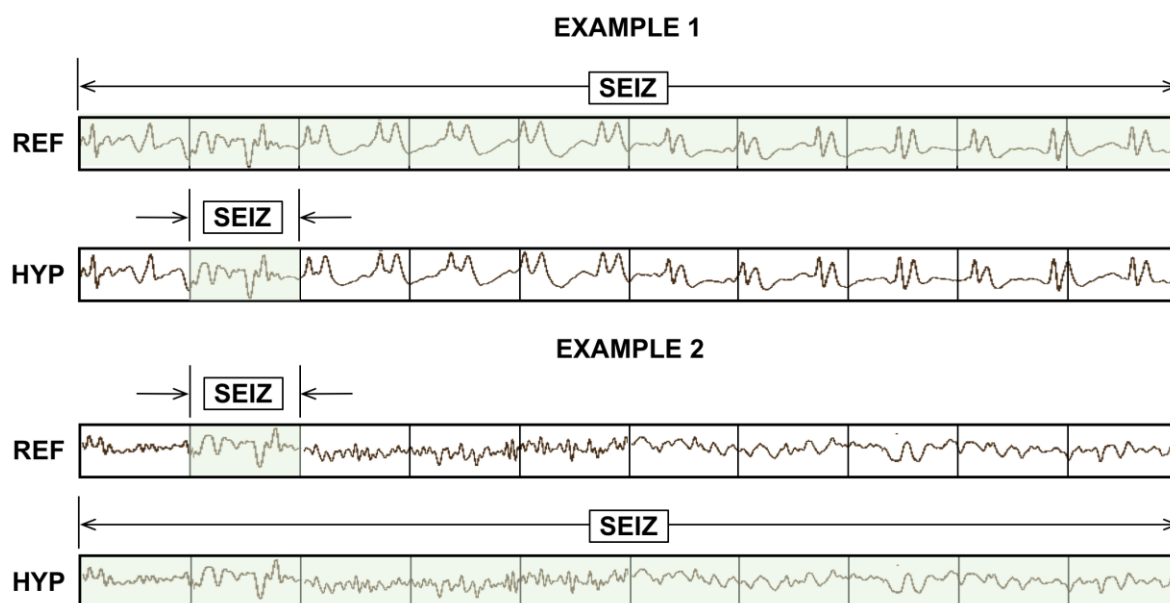

EXAMPLE 2

Figure 2. TP scores for the Any-Overlap method are 100% even though large portions of the event are missed.

(1) NIST Actual Term-Weighted Value (ATWV): based on NIST's popular scoring package (F4DE v3.3.1), this metric, originally developed for the NIST 2006 Spoken Term Detection evaluation, uses an objective function that accounts for temporal overlap between the reference and hypothesis using the detection scores assigned by the system.

(2) Dynamic Programming Alignment (DPALIGN): similar to the NIST package known as SCLite 35, this metric uses a dynamic programming algorithm to align terms. It is most often used in a mode in which the time alignments produced by the system are ignored.

(3) Epoch-Based Sampling (EPOCH): treats the reference and hypothesis as temporal signals, samples each at a fixed epoch duration, and counts errors accordingly.

(4) Any-Overlap (OVLP): assesses the overlap in time between a reference and hypothesis event, and counts errors using binary scores for each event.

(5) Time-Aligned Event Scoring (TAES): similar to (4), but considers the percentage overlap between the two events and weights errors accordingly.

(6) Inter-Rater Agreement (IRA): uses EPOCH scoring to estimate errors, and calculates Cohen's Kappa coefficient 36 using the measured TP, TN, FP and FN.

It is important to understand that each of these measures estimates TP, TN, FP and FN through some sort of error analysis. From these estimated quantities, traditional derived measures such as sensitivity and specificity are computed. As a result, we will see that sensitivity is a function of the underlying metric, and this is why it is important there be community-wide agreement on a specific metric.

We now briefly describe each of these approaches and provide several examples that illustrate their strengths and weaknesses. These examples are drawn on a compressed time-scale for illustrative purposes and were carefully selected because they are indicative of scoring metric problems we have observed in actual evaluation data collected from our algorithm research.

2.1. NIST Actual Term-Weighted Value (ATWV)

167   ATWV is a measure that balances sensitivity and FA rate. ATWV essentially assigns an application-
168   dependent reward to each correct detection and a penalty to each incorrect detection. A perfect
169   system results in an ATWV of 1.0, while a system with no output results in an ATWV of 0. It is
170   possible for ATWV to be less than zero if a system is doing very poorly (for example a high FA rate).
171   Experiments in voice keyword search have shown that an ATWV greater than 0.5 typically indicates
172   a promising or usable system for information retrieval by voice applications. We believe a similar
173   range is applicable to EEG analysis.

174   The metric accepts as input a list of N-tuples representing the hypotheses for the system being
175   evaluated. Each of these N-tuples consists of a start time, end time and system detection score. These
176   entries are matched to the reference annotations using an objective function that accounts for both
177   temporal overlap between the reference and hypotheses and the detection scores assigned by the
178   system being evaluated. These detection scores are often likelihood or confidence scores 23. The
179   probabilities of miss and FA errors at a detection threshold θ are computed using:

180
$$P_{Miss(kw,\theta)} = 1 - {N_{Correct(kw,\theta)}}\big/{N_{Ref(kw)}} \;, \tag{1}$$

181
$$P_{FA(kw,\theta)} = {N_{Spurious(kw,\theta)}}\big/{N_{NT(kw)}} \;, \tag{2}$$

182   where $N_{Correct(kw,\theta)}$ is the number of correct detections of terms with a detection score greater than
183   or equal to θ, $N_{Spurious(kw,\theta)}$ is the number of incorrect detections of terms with a detection score
184   greater than or equal to θ, and $N_{NT(kw)}$ is number of non-target trials for the term kw in the data. The
185   number of non-target trials for a term is related to the total duration of source signal in
186   seconds, T_{Source}, and is computed as $N_{NT(kw)} = T_{Source} - N_{Ref(kw)}$.

187   A term-weighted value is then computed that specifies a trade-off of misses and FAs. ATWV is
188   defined as the value of TWV at the system's chosen detection threshold. Using a predefined constant,
189   β, that was optimized experimentally (β = 999.9) 24, ATWV is computed using:

190
$$TWV_{(kw,\theta)} = 1 - P_{Miss(kw,\theta)} - \beta\, P_{FA(kw,\theta)} \;. \tag{3}$$

191   A standard implementation of this approach is available at 37. This metric has been widely used
192   throughout the human language technology community for 15 years. This is a very important
193   consideration in standardizing such a metric – researchers are using a common shared software
194   implementation that ensures there are no subtle implementation differences between sites or
195   researchers.

196   To demonstrate the features of this approach, consider the case shown in Figure 3. The hypothesis
197   for this segment consists of several short seizure events while the reference consists of one long event.
198   The ATWV metric will assign a TP score of 100% because the midpoint of the first event in the
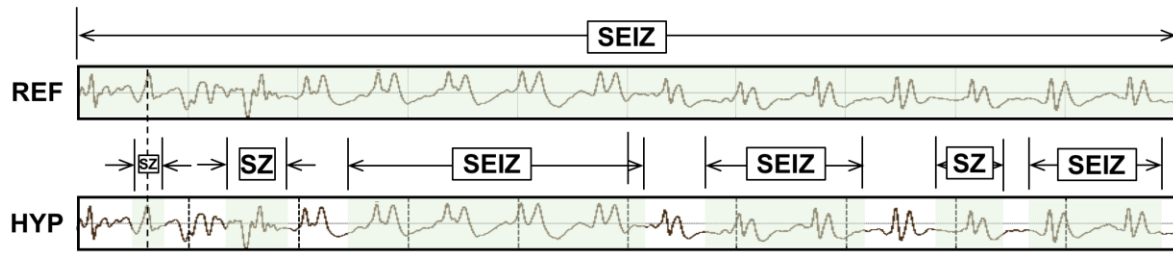
Figure 3. ATWV scores this segment as 1 TP and 5 FPs. The first event's mid-point overlaps with the reference. The remaining five of the fully overlapping hypothesis events are considered as FPs.

Figure 4. ATWV scores this segment as 0 TP and 3 FNs. The mid-point of the hypothesis event does not overlap with any of the reference events so all three reference events are considered as FNs.

199   hypothesis annotation is mapped to the long seizure event in the reference annotation. This is
200   somewhat generous given that 50% of the event was not detected. The remaining 5 events in the
201   hypothesis annotation are counted as false positives. The ATWV metric is relatively insensitive to the
202   duration of the reference event, though the 5 false positives will lower the overall performance of the
203   system. The important issue here is that the hypothesis correctly detected about 70% of the seizure
204   event, and yet because of the large number of false positives, it will be penalized heavily.
205   In Figure 4 we demonstrate a similar case in which the metric penalizes the hypothesis for missing
206   three seizure events in the reference. Approximately 50% of the segment is correctly identified. This
207   type of scoring penalizing repeated events that are part of a larger event in the reference might make
208   sense in an application like voice keyword search because in human language each word hypothesis
209   serves a unique purpose in the overall understanding of the signal. However, for a two-class event
210   detection problem such as seizure detection, such scoring too heavily penalizes the hypothesis for
211   splitting a long event into a series of short events.

212   *2.2. Dyamic Programming Alignment (DPALIGN)*

213   The DPALIGN metric essentially performs a minimization of an edit distance (the Levenshtein
214   distance) 12 to map the hypothesis onto the reference. DPALIGN determines the minimum number
215   of edits required to transform the hypothesis string into the reference string. Given two strings, the
216   source string X = [x1, x2, ..., xn] of length n, and target string Y = [y1, y2, ..., ym]   of length m, we
217   define $d_{i,j}$, which is the edit distance between the substring x1:xi and y1:yj, as:

218    $$d_{i,j} = \begin{cases} d_{i-1,j} + del \\ d_{i,j-1} + ins \\ d_{i-1,j-1} + sub \end{cases} ,$$ (4)

219   The quantities being measured here are often referred to as substitution (sub), insertion (ins) and
220   deletion (del) penalties. For this study, these three penalties are assigned equal weights of 1. A
221   dynamic programming algorithm is used to find the optimal alignment between the reference and

```
              Ref: bckg seiz bckg seiz bckg **** ****
              Hyp: bckg seiz bckg seiz bckg SEIZ BCKG
         (Hits: 5 Sub: 0 Ins: 2 Del: 0 Total Errors: 2)


              Ref: bckg seiz bckg seiz bckg SEIZ BCKG
              Hyp: bckg seiz bckg seiz bckg **** ****
         (Hits: 5 Sub: 0 Ins: 0 Del: 2 Total Errors: 2)
```

Figure 5. DPALIGN aligns symbol sequences based on edit distance and ignores time alignments.

222 hypothesis based on these weights. Though there are versions of this metric that perform time-
223 aligned scoring in which both the reference and hypothesis must include start and end times, this
224 metric is most commonly used without time alignment information.
225 The metric is best demonstrated using the two examples shown in Figure 5. In the first example, the
226 reference annotation has a series of 5 events, while the hypothesis contains two additional events.
227 The latter are counted as insertion errors. In the second example, the reference annotation now has
228 two additional events which are missing from the hypothesis. These are counted as deletion errors.
229 For convenience, lowercase symbols indicate correct detections while uppercase symbols indicate
230 errors. The asterisk symbol is used to denote deletion and insertion errors. Note that there is
231 ambiguity in these alignments. For example, it is not really clear which of the three seizure events in
232 the second example corresponded to each of the seizure events in the hypothesis. Nevertheless, this
233 ambiguity doesn't really influence the overall scoring. Though this type of scoring might at first seem
234 highly inaccurate since it ignores time alignments of the hypotheses, it has been surprisingly effective
235 in scoring machine learning systems in sequential data applications (e.g., speech recognition) 1235.

236 *2.3. Epoch-Based Sampling (EPOCH)*

237 Epoch-based scoring uses a metric that treats the reference and hypothesis as signals. These signals
238 are sampled at a fixed epoch duration. The corresponding label in the reference is compared to the
239 hypothesis. Similar to DPALIGN, substititions, deletions and insertion errors are tabulated with an
240 equal weight of 1 for each type of error. This process is depicted in Figure 6. Epoch-based scoring
241 requires that the entire signal be annotated, which is normally the case for sequential decoding
242 evaluations. It attempts to account for the amount of time the two annotations overlap, so it directly
243 addresses the inconsistencies demonstrated in Figure 3 and Figure 4.
244 One important parameter to be tweaked in this algorithm is the frequency with which we sample the
245 two annotations, which we refer to as the scoring epoch duration. It is ideally set to an amount of
246 time smaller than the unit of time used by the classification system to make decisions. For example,
247 the hypothesis in Figure 6 outputs decisions every 1 sec. The scoring epoch duration should be set
248 smaller than this. We use a scoring epoch duration of 0.25 sec for most of our work because our
249 analysis system epoch duration is typically 1 sec. We find in situations like this the results are not
250 overly sensitive to the choice of the epoch duration as long as it is below 1 sec. This parameter simply
251 controls how much precision one expects for segment boundaries.
252 Because EPOCH scoring samples the annotations at fixed time intervals, it is inherently biased to
253 weigh long seizure events more heavily. For example, if a signal contains one extremely long seizure
254 event (e.g., 1000 secs) and two short events (e.g., each 10 secs in duration), the accuracy with which
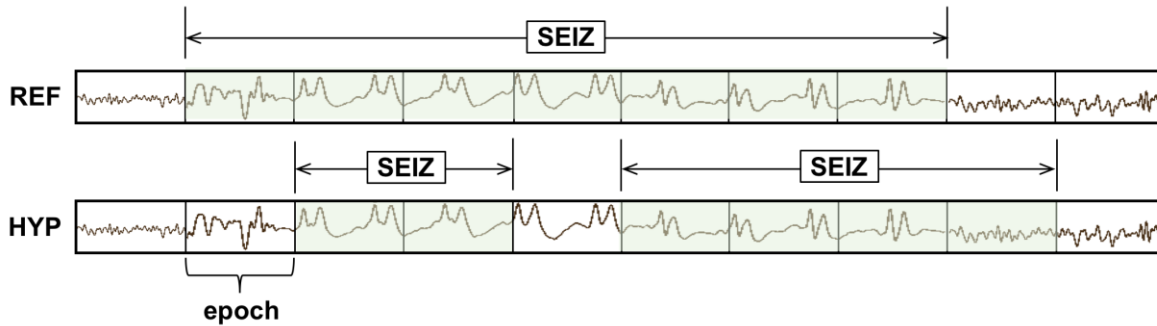
Figure 6. EPOCH scoring directly measures the similarity of the time-aligned annotations. TP, FN and FP are 5, 2 and 1 respectively. TPs are considered for epoch 3, 4 and 6-9. Epochs 2 and 5 are missed (counted as FNs). Epoch 9 adds an FP.

255 the first event is detected will dominate the overall scoring. Since seizure events can vary dramatically
256 in duration, this is a cause for concern.

257 *2.4. Any-Overlap Method (OVLP)*

258 We previously introduced the OVLP metric as a popular choice in the neuroengineering community
259 3031. OVLP is a more permissive metric that tends to produce much higher sensitivities. If an event
260 is detected in close proximity to a reference event, the reference event is considered correctly detected.
261 If a long event in the reference annotation is detected as multiple shorter events in the hypothesis,
262 the reference event is also considered correctly detected. Multiple events in the hypothesis annotation
263 corresponding to the same event in the reference annotation are not typically counted as FAs. Since
264 the FA rate is a very important measure of performance in critical care applications, this is another
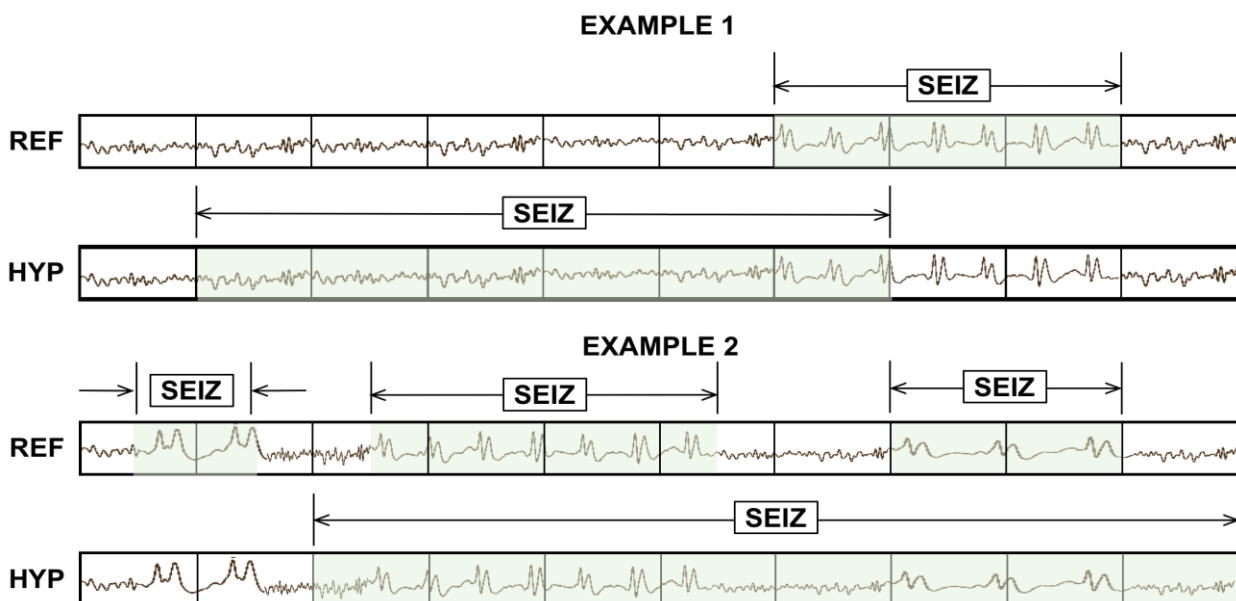265 cause for concern.

Figure 7. OVLP scoring is very permissive about the degree of overlap between the reference and hypothesis. The TP score for example 1 is 1 with no false alarms. In example 2, the system detects 2 out of 3 seizure events, so the TP and FN scores are 2 and 1 respectively.

266   The OVLP scoring method is demonstrated in Figure 7. It has one significant tunable parameter – a
267   guard band that controls the degree to which a misalignment is still considered as a correct match. In
268   this study, we use a fairly strict interpretation of this band and require some overlap between the two
269   events in time – essentially a guard band of zero. The guard band needs to be tuned based on the
270   needs of the application. Sensitivity generally increases as the guard band is increased.

271   *2.5. Time-Aligned Event Scoring (TAES)*

272   Though EPOCH scoring directly measures the amount of overlap between the annotations, there is a
273   possibility that this too heavily weighs single long events. Seizure events can vary in duration from
274   a few seconds to many hours. In some applications, correctly detecting the number of events is as
275   important as their duration. Hence, the TAES metric was designed as a compromise between these
276   competing constraints. The essential parameters for calculation of sensitivity and specificity such as
277   TP, TN and FP for TAES scoring metric are defined as follows:

278    $$TP = \frac{H_{stop} - H_{start}}{Ref_{dur}}, \qquad\qquad where\ R_{start} \le H \le R_{stop}\,, \qquad\qquad (5)$$

279    $$TN = \frac{1 - (TH_{stop} - TH_{start})}{Ref_{dur}}, \qquad\qquad where\ R_{start} \le H \le R_{stop}\,, \qquad\qquad (6)$$

280    $$FP = \begin{cases} \frac{H_{stop} - R_{stop}}{Ref_{dur}}, & if\ H_{stop} \ge R_{stop},\ H_{start} \ge R_{start}\ and\ H_{stop} - R_{stop} \le 1, \\ \frac{R_{start} - H_{start}}{Ref_{dur}}, & if\ R_{start} \ge H_{start}, R_{stop} \ge H_{stop}\ and\ R_{start} - H_{start} \le 1, \\ 1, & otherwise, \end{cases} \qquad (7)$$

281   where H and R represent the reference and hypothesis events respectively, and Ref_{dur} represents
282   the duration of the reference events.
283   TAES gives equal weight to each event, but it calculates a partial score for each event based on the
284   amount of overlap. The TP score is the total duration of a detected term divided by the total duration
285   of the reference term. The FN score is the fraction of the time the reference term was missed divided
286   by the total duration of the reference term. The FP score is the total duration of the inserted term
287   divided by total amount of time this inserted term was incorrect according to the reference
288   annotation. FPs are limited to a maximum of 1 per event. Therefore, like TP and FN, a single FP event
289   contributes only a fractional amount to the overall FP score if it correctly detects a portion of the same
290   event in the reference annotation (partial overlap). Moreover, if multiple reference events are detected
291   by a single long hypothesis event, all but the first detection are considered as FNs. These properties
292   of the metric help manage the tradeoff between sensitivity and FAs by balancing the contributions
293   from short and long duration events. An example of TAES scoring is depicted in Figure 8.

294   *2.6. Inter-Rater Agreement (IRA)*

295   Inter-rater agreement (IRA) is a popular measure when comparing the relative similarity of two
296   annotations. We refer to this metric as a derived metric since it is computed from error counts
297   collected using one of the other five metrics. IRA is most often measured using Cohen's Kappa
298   coefficient 36, which compares the observed accuracy with the expected accuracy. It is computed
299   using:

300     $$\kappa = \frac{p_0 - p_e}{1 - p_e},$$ (8)

301   where p_o is the relative observed agreement among raters and p_e is the hypothetical probability of
302   chance agreement.
303   The Kappa coefficient ranges between $\kappa = 1$ (complete agreement) and $-1 \leq \kappa \leq 0$ (no
304   agreement). It has been used extensively to assess inter-rater agreement for experts manually
305   annotating seizures in EEG signals. Values in the range of $0.5 \leq \kappa \leq 0.8$ are common for these types
306   of assessments 38. The variability amongst experts mainly involves fine details in the annotations,
307   such as the exact onset of a seizure. These kinds of details are extremely important for machine
308   learning and hence we need a metric that is sensitive to small variations in the annotations. For
309   completeness, we use this measure as a way of evaluating the amount of agreement between two
310   annotations.

311   *2.7. A Brief Comparison of Metrics*

312   A simple example of how these metrics compare on a specific segment of a signal is shown in Figure
313   9. A 10-sec section of an EEG signal is shown subdivided into 1-sec segments. The reference has three
314   isolated events. The system being evaluated outputs one hypothesis that starts in the middle of the
315   first event and continues through the remaining two events. ATWV scores the system as 1 TP and 2
316   FNs since it assigns the extended hypothesis event to the center reference event and leaves the other
317   two undetected. The ATWV score is 0.33 for seizure events, 0.25 for background events, resulting in
318   an average ATWV of 0.29. The sensitivity and FA rates for seizure events for this metric are 33% and
319   0 per 24 hrs. respectively. DPALIGN scores the system the same way since time alignments are
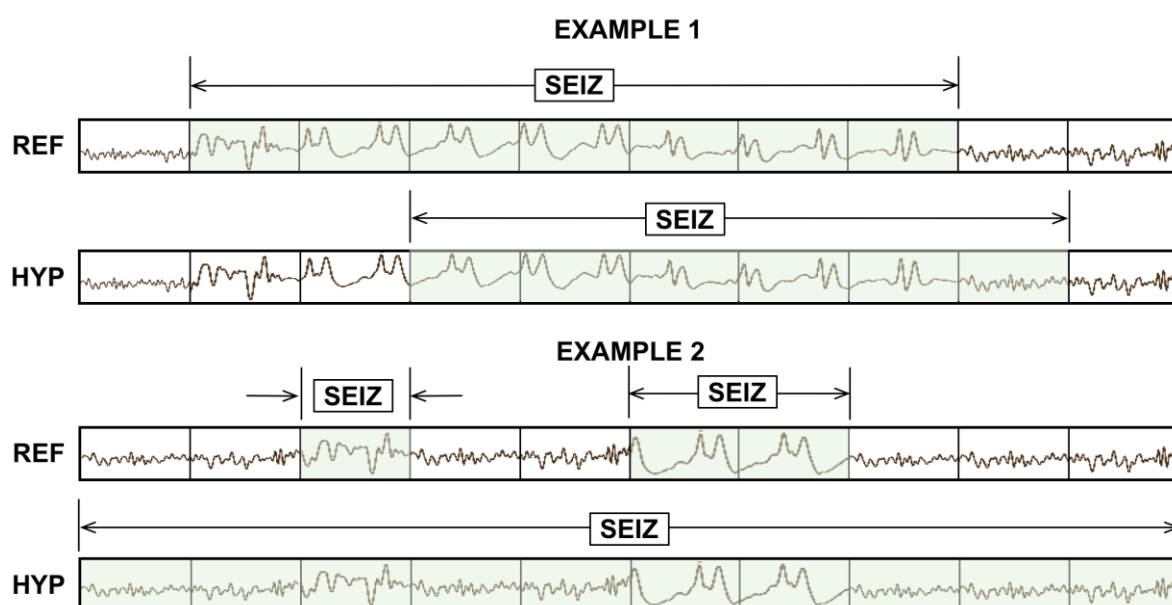

Figure 8. TAES scoring accounts for the amount of overlap between the reference and hypothesis. TAES scores example 1 as 0.71 TP, 0.29 FN and 0.14 FP. Example 2 is scored as 1 TP (due to event at epoch 3), 1 FN for the second reference event spanning epochs 6 and 7, and 1 FP for the remaining epochs that were incorrectly detected as seizures.

Figure 9. An example that summarizes the differences between scoring metrics.

320 ignored and the first event in each annotation are matched together, leaving the other two events
321 undetected.

322 The EPOCH method scores the alignment 5 TP, 3 FP and 1 FN using a 1-sec epoch duration because
323 there are 4 epochs for which the annotations do not agree and 5 epochs where they agree. The
324 sensitivity is 83.33% and the FA rate per 24 hrs. is very high because of the 3 FPs. The OVLP method
325 scores the segment as 3 TP and 0 FP because detected events have partial to full overlap with all the
326 reference events, giving a sensitivity of 100% with an FA rate of 0. TAES scores this segment as 0.5
327 TP and 2.5 FN because the first event is only 50% correct and there are FN errors for the 5th to 7th
328 and 9th epochs (an example of multiple overlapping reference events), giving a sensitivity of 16.66%
329 and a corresponding high FA rate.

330 IRA for seizure events evaluated using Cohen's Kappa statistic is 0.09 for this example because there
331 are essentially 4 errors for 6 seizure events. IRAs below 0.5 indicate a poor match between the
332 reference and the hypothesis.

333 It is difficult to conclude from this example which of these measures are most appropriate for EEG
334 analysis. However, we see that ATWV and DPALIGN generally produce similar results. The EPOCH
335 metric produces larger counts because it samples time rather than events. OVLP produces a high
336 sensitivity while TAES produces a low sensitivity but a relatively higher FA rate.

337 **3. Results**

338 This To demonstrate the differences between these metrics on a realistic task, we have evaluated a
339 range of machine learning systems on a seizure detection task based on the TUH EEG Seizure Corpus
340 39. An overview of the corpus is given in Table 1. This is the largest open source corpus of its type. It
341 consists of clinical data collected at Temple University Hospital, and represents a very challenging
342 machine learning task because it contains a rich
343 variety of common real-world problems found in
344 clinical data (e.g., patient movement). There are 50
345 patients in the evaluation corpus, making it large
346 enough to accurately assess fine differences in
347 algorithm performance.

348 A general architecture for the five machine
349 learning systems evaluated is shown in Figure 10.
350 An EEG signal is input using a European Data
351 Format (EDF) file. The signal is converted to a
352 sequence of feature vectors. LFCC (Linear

Table 1. The TUH EEG Seizure Corpus (v1.1.1)

Description	Train	Eval
Patients	196	50
Sessions	456	230
Files	1,505	984
No. Seizure Events	870	614
Seizure (secs)	51,140	53,930
Non-Seizure (secs)	877,821	547,728
Total (secs)	928,962	601,659

353 Frequency Cepstral Coefficient) features are created using a 0.1 sec frame duration and a 0.2 second
354 analysis window for each channel 40. We use first 7 cepstral coefficients along with their first and
355 second derivatives. A group of frames are classified into an event on a per-channel basis using
356 combination of deep learning networks. The deep learning system essentially looks across multiple
357 epochs, which we refer to as the temporal context, and multiple channels, which we refer to as the
358 spatial context since each channel is associated with a location of an electrode on a patient's head.
359 There are a wide variety of algorithms that can be used to produce a decision from these inputs. Even
360 though seizures occur on a subset of the channels input to such a system, we focus on a single decision
361 made across all channels at each point in time.

362 The five systems selected were carefully chosen because they represent a range of performance that
363 is representative of state of the art on this task and because these systems exhibit different error
364 modalities on this task. The performance of these systems is sufficiently close so that the impact of
365 these different scoring metrics becomes apparent. The systems selected were:

366 (1) HMM/SdA: a hybrid system consisting of a hidden Markov model (HMM) decoder and a
367 postprocessor that uses a Stacked Denoising Autoencoder (SdA). An N-channel EEG was
368 transformed into N independent feature streams using a standard sliding window based
369 approach. The hypotheses generated by the HMMs were postprocessed using a second stage
370 of processing that examines the temporal and spatial context. We apply a third pass of
371 postprocessing that uses a stochastic language model to smooth hypotheses involving
372 sequences of events so that we can suppress spurious outputs. This third stage of
373 postprocessing provides a moderate reduction in the false alarm rate.

374 Standard three state left-to-right HMMs with 8 Gaussian mixture components per state were
375 used for sequential decoding. We divide each channel of an EEG into 1-second epochs, and
376 further subdivide these epochs into a sequence of frames. Each epoch is classified using an
377 HMM trained on the subdivided epoch, and then these epoch-based decisions are
378 postprocessed by additional statistical models in a process similar to the language modeling
379 component of a speech recognizer.

380 The output of the epoch-based decisions was postprocessed by a deep learning system. The
381 SdA network has three hidden layers with corruption levels of 0.3 for each layer. The number
382 of nodes per layer are: first layer = 800, second layer = 500, third layer = 300. The parameters for
383 pre-training are: learning rate = 0.5, number of epochs = 150, batch size = 300. The parameters
384 for fine-tuning are: learning rate = 0.1, number of epochs = 300, batch size = 100. The overall
385 result of the second stage is a probability vector of dimension two containing a likelihood that
386 each label could have occurred in the epoch. A soft decision paradigm is used rather than a
387 hard decision paradigm because this output is smoothed in the third stage of processing.

388 (2) HMM/LSTM: an HMM decoder postprocessed by a Long Short-Term Memory (LSTM)
389 network. Like the HMM/SdA hybrid approach previously described, the output of the HMM
390 system is a vector of dimension 2 × number of channels (22) × the window length (7). Therefore,
391 we also use PCA before LSTM in this approach to reduce the dimensionality of the data to 20.
392 For this study, we used a window length of 41 for LSTM, and this layer is composed of one
393 hidden layer with 32 nodes. The output layer nodes in this LSTM level use a sigmoid function.
394 The parameters of the models are optimized to minimize the error using a cross-entropy loss
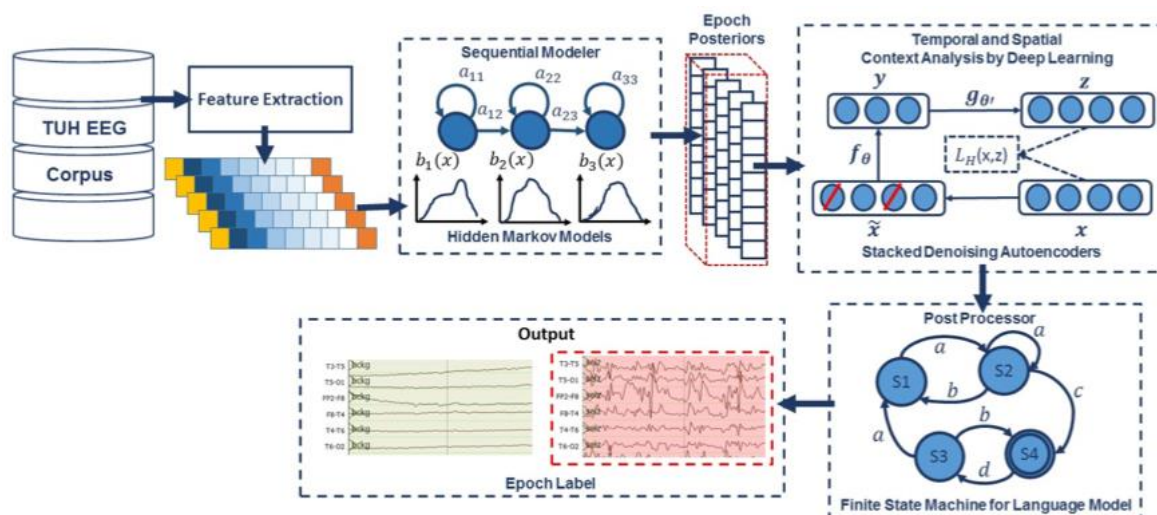
Figure 10. A typical hybrid machine learning model architecture that incorporates spatial and temporal features

395	function. Adaptive Moment Estimation (Adam) is used in the optimization process.
396	(3) IPCA/LSTM: a preprocessor based on Incremental Principal Component Analysis (IPCA)
397	followed by an LSTM decoder. The EEG features are delivered to an IPCA layer for spatial
398	context analysis and dimensionality reduction. A batch size of 50 is used in IPCA and the
399	output dimension is 25. The output of IPCA is delivered to a LSTM for classification. We used
400	a one-layer LSTM with a hidden layer size of 128 and batch size of 128 is used along with Adam
401	optimization and a cross–entropy loss function.
402	(4) CNN/MLP: a pure deep learning-based approach that uses a Convolutional Neural Network
403	(CNN) decoder and a Multi-Layer Perceptron (MLP) postprocessor. The network contains six
404	convolutional layers, three max pooling layers and two fully-connected layers. A rectified
405	linear unit (ReLU) non-linearity is applied to the output of every convolutional and fully-
406	connected layer.
407	(5) CNN/LSTM: a pure deep learning-based architecture that uses a combination of CNN and
408	LSTM networks. In this architecture, we integrate 2D CNNs, 1D CNNs and LSTM networks to
409	better exploit long-term dependencies. Exponential Linear Units (ELU) are used as the
410	activation functions for the hidden layers. Adam is used in the optimization process along with
411	a mean squared error loss function.

412 Comprehensive details about the architectures are available in 4040. The details of these systems are
413 not critical to this study. We selected these systems because we needed a range of typical system
414 performance that would expose differences in the scoring metrics. What is most important is how the
415 range of performance is reflected in these metrics.

416 A comparison of the performance of the different architectures is presented in Table 2. Though the
417 relative rankings of these systems not surprisingly vary with the metric, the ranking of these systems
418 is accurately represented by the overall trends in Table 2. HMM/SdA generally performs the poorest
419 of these systems, delivering a respectable sensitivity but at a high FA rate. CNN/LSTM typically
420 delivers highest performance and has a low FA rate, which is very important in this type of
421 application.   section may be divided by subheadings. It should provide a concise and precise

<div align="center">**Table 2.** Performance vs. scoring metric</div>

Metric	Measure	HMM/SdA	HMM/LSTM	IPCA/LSTM	CNN/MLP	CNN/LSTM
ATWV	Sensitivity	30.35%	26.73%	24.73%	29.52%	30.34%
	Specificity	61.38%	68.93%	64.51%	65.87%	93.15%
	FAs/24 hrs	98	75	94	94	11
	ATWV	-0.8392	-0.8469	-0.4628	-0.7971	0.1737
OVLP	Sensitivity	35.35%	30.05%	32.97%	39.09%	30.83%
	Specificity	73.35%	80.53%	77.57%	76.84%	96.86%
	FAs/24 hrs	77	60	73	77	7
DPALIGN	Sensitivity	44.11%	33.77%	35.77%	43.35%	32.46%
	Specificity	66.87%	72.99%	69.59%	71.49%	95.17%
	FAs/24 hrs	86	66	81	77	8
TAES	Sensitivity	17.29%	22.84%	22.12%	31.58%	12.48%
	Specificity	66.04%	70.41%	66.64%	64.75%	95.24%
	FAs/24 hrs	82	68	83	91	8
EPOCH	Sensitivity	20.71%	50.46%	51.02%	65.03%	9.784%
	Specificity	98.22%	94.82%	94.09	91.55%	99.84%
	FAs/24 hrs	1418	4133	4711	6738	126

422 description of the experimental results, their interpretation as well as the experimental conclusions
423 that can be drawn.

4. Discussion

425 Evaluating systems from a single operating point is always a bit tenuous. Therefore, in Figure 11, we
426 provide DET curves for the systems and in Table 3 we provide AUCs for these DET curves calculated
427 using OVLP and TAES for comparison. This is due to our emphasis on using OVLP and TAES metrics
428 for seizure detection-like applications. The DET curves were derived from output from the OVLP
429 scoring metric. The shapes of the DET curves do not change significantly with the scoring metric
430 though the absolute numbers vary similarly to what we see in Table 2. AUC values from Table 3 also
431 follow similar trends but the differences between the AUC and TAES metrics is less pronounced
432 compared to the differences between AUC and OVLP. It is clear from this data that CNN/LSTM
433 performance is significantly different from the other systems. This is primarily because of its low FA
434 rate. For this particular application, sensitivity drops rapidly as the FA rate is lowered. Therefore,
435 comparing a single data point for each system is dangerous because the systems are most likely
436 operating at different points on a DET curve if the sensitivities are significantly different. We find
437 tuning these systems to have a comparable FA rate is important when comparing two systems only
438 based on sensitivity.
439 In Table 2 we can examine the sensitivity of the different metrics by looking at the variation in
440 sensitivity. For example, for HMM/SdA, we see the lowest sensitivities are produced by TAES and
441 EPOCH scoring, while the highest sensitivities are produced by OVLP and DPALIGN. This makes
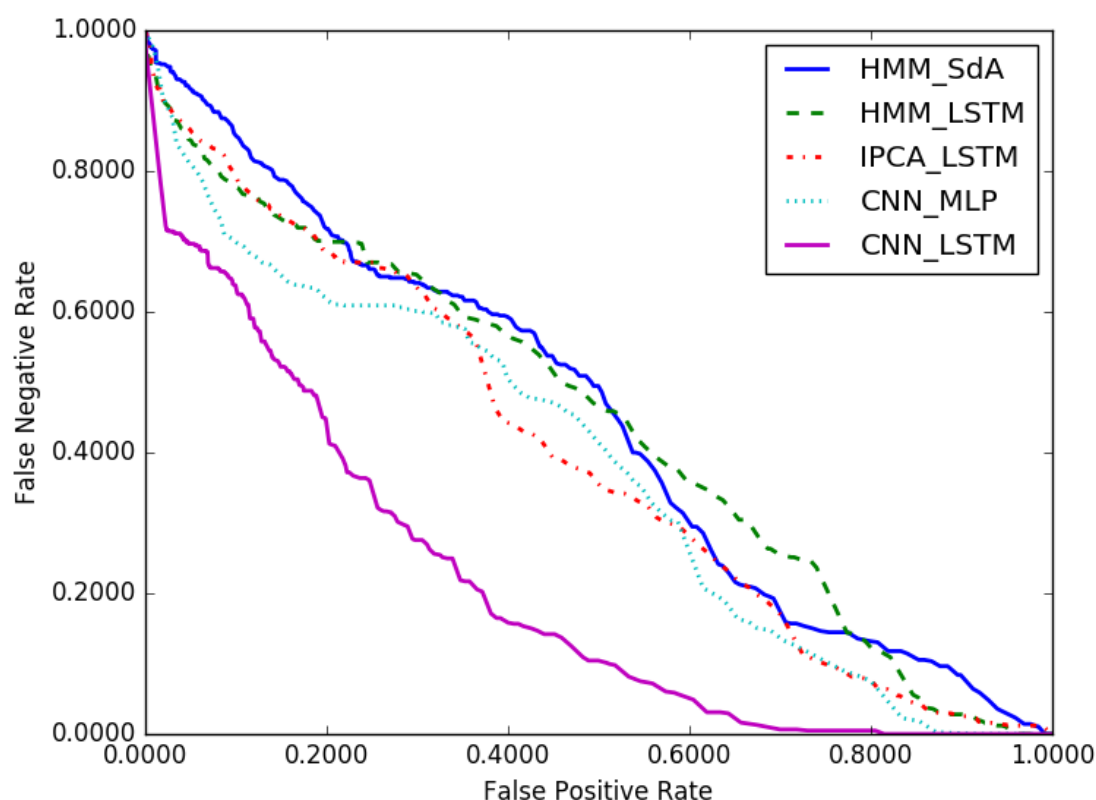442 sense because OVLP and DPALIGN are very forgiving of time alignment errors, while TAES and

Figure 11. A comparison of DET curves (FNR Vs. FPR) for all 5 machine learning architectures

443 EPOCH penalize time alignment errors heavily. We see
444 similar trends for CNN/LSTM though the range of
445 differences between the three highest scoring metrics is
446 smaller. We also see that the five algorithms are ranked
447 similarly by each scoring metric. HMM/SdA consistently
448 scores the lowest and CNN/LSTM consistently scores the
449 highest. The other three systems are very similar in their
450 performance.
451 The ATWV scores for all algorithms are extremely low.
452 The ATWV scores are below 0.5 which indicates that

Table 3. A comparison of AUC based on the OVLP and TAES metrics

Algorithm	AUC (OVLP)	AUC (TAES)
HMM/SdA	0.44	0.72
HMM/LSTM	0.44	0.71
IPCA/LSTM	0.39	0.72
CNN/MLP	0.38	0.65
CNN/LSTM	**0.21**	**0.56**

453 overall performance is poor. However, the ATWV score for CNN/LSTM is significantly higher than
454 the other four systems. ATWV attempts to reduce the information contained in a DET curve to a
455 single number, and does a good job reflecting the results shown in Figure 11. The DET curves for
456 HMM/LSTM and HMM/SdA overlap considerably for an FP rate between 0.25 and 1.0, and this is a
457 primary reason why their ATWV scores are similar. However, for the seizure detection application
458 we are primarily interested in the low FP rate region, and in that range, HMM/LSTM and IPCA/LSTM
459 perform similarly.
460 While sensitivity and specificity are commonly used metrics in the bioengineering community, from
461 Table 2 and Figure 11 we see that the FA rate also plays a major role in determining the usability of a
462 system. A commonly used metric in the machine learning community that is somewhat intuitive is
463 accuracy. The accuracy of the five systems is shown in Table 4. Accuracy weights all types of errors
464 as equally important. This is acceptable if the dataset is balanced. However, for many bioengineering
465 applications, such as seizure detection, the target class, or class of interest, occurs infrequently. We

Table 4. Accuracy vs. scoring metric

Metric	HMM/SdA	HMM/LSTM	IPCA/LSTM	CNN/MLP	CNN/LSTM
ATWV	54.0%	54.0%	52.1%	54.9%	70.7%
OVLP	65.1%	66.5%	65.6%	66.9%	78.9%
DPALIGN	61.5%	60.2%	59.2%	62.9%	73.6%
TAES	56.6%	57.3%	55.4%	57.2%	69.7%
EPOCH	92.3%	91.5%	90.8 %	89.5%	91.5%

Table 5. F_1 score vs. scoring metric

Metric	HMM/ SdA	HMM/LSTM	IPCA/LSTM	CNN/MLP	CNN/LSTM
ATWV	0.24	0.28	0.24	0.28	0.42
OVLP	0.31	0.33	0.34	0.38	0.45
DPALIGN	0.35	0.36	0.35	0.42	0.45
TAES	0.16	0.26	0.24	0.31	0.19
EPOCH	0.29	0.47	0.46	0.49	0.14

see that CNN/LSTM is significantly more accurate than the other four systems, but that the differences between these remaining four systems is minimal when using accuracy as a metric.

Another popular metric that attempts to aggregate performance into a single data point, and is popular in the information retrieval communities, is the F1 score. These scores for the five systems are shown in Table 5. We see there are significant variations in F1 scores. For example, for the TAES and EPOCH metrics, which stress time alignments, the best performing system is not CNN/LSTM. F1 scores do not adequately emphasize FAs for applications such as seizure detection.

We generally prefer operating points where performance in terms of sensitivity, specificity and FAs is balanced. The ATWV metric explicitly attempts to balance these by assigning a reward to each correct detection and a penalty to each incorrect detection. None of the conventional metrics described here consider the fraction of a detected event that is correct. This is the inspiration behind the development of TAES scoring. TAES scoring requires the time alignments to match, which is a more stringent requirement than, for example, OVLP. Consequently, the sensitivity produced by the TAES and EPOCH metrics tends to be lower.

Finally, comparing results across these five metrics can provide useful diagnostic information and provide insight into the system's behavior. For example, the IPCA/LSTM and HMM/LSTM systems have relatively higher sensitivities according to the EPOCH metric, indicating that these systems tend to detect longer seizure events. Conversely, since the CNN/LSTM system has relatively low sensitivities according to the TAES and EPOCH metrics, it can be inferred that this system misses longer seizure events. Similarly, if the sensitivity was relatively high for TAES and relatively low for EPOCH, it would indicate that the system tends to detect a majority of smaller to moderate events precisely regardless of the duration of an event. Similarly, a comparison of ATWV scores with other metrics gives diagnostic information such as whether a system accurately detects the onset and end of an event or whether the system splits long events into multiple short events. Examining the ensemble of scores can be revealing for these six metrics.

To understand the pairwise statistical difference between these evaluation metrics and deep

Table 6. Correlation of the scoring metrics (for sensitivity)

Metric	ATWV	DPALIGN	OVLP	TAES	EPOCH
ATWV	---	**0.87 (p < 0.001)**	**0.92 (p < 0.001)**	0.71 (p < 0.001)	*0.50 (p < 0.001)*
DPALIGN		---	**0.90 (p < 0.001)**	0.69 (p < 0.001)	*0.48 (p < 0.001)*
OVLP			---	0.78 (p < 0.001)	0.62 (p < 0.001)
TAES				---	0.87 (p < 0.001)
EPOCH					---

Table 7. Correlation of the scoring metrics (for specificity)

Metric	ATWV	DPALIGN	OVLP	TAES	EPOCH
ATWV	---	0.49 (p < 0.001)	0.45 (p < 0.001)	0.54 (p < 0.001)	*0.32 (p < 0.001)*
DPALIGN		---	**0.94 (p < 0.001)**	**0.89 (p < 0.001)**	*0.38 (p < 0.001)*
OVLP			---	**0.95 (p < 0.001)**	0.44 (p < 0.001)
TAES				---	0.56 (p < 0.001)
EPOCH					---

architectures, we have performed several tests: Kolmogorov-Smirnov (KS), Pearson's R (correlation coefficient) and Z-test. These tests were performed to evaluate results of hybrid deep learning architectures on the basis of sensitivity and specificity. Each individual patient from the TUSZ dataset was evaluated separately. Outliers were removed by rejecting all input values collected from patients which have no seizures and from those for which deep learning systems detected no seizures.

Prior to performing tests for evaluating statistically differences, such as a z-test, t-test or ANOVA, it must first be determined whether or not the group sample, in our case individual metric's score on per patient evaluation, is normally distributed. We performed KS tests on each separate evaluation metric and confirmed that the group distribution is indeed Gaussian. The KS values for normal distributions collected range from 0.61 – 0.71 for sensitivity and 0.99 – 1.00 for specificity with the p-values equal to zero. We then evaluate the correlation coefficient (Pearson's R) between pairs of metrics.

Correlations for each pair of scoring metrics are shown in Table 6 (for sensitivity) and Table 7 (for specificity). It can be seen that the pairs ATWV-EPOCH and DPALIGN-EPOCH, have minimum correlation (~0.5). The pairwise correlations between OVLP, ATWV and DPALIGN are much higher. The EPOCH method has a low correlation with all other metrics but TAES. This makes sense because the EPOCH method scores events on a constant time scale instead of on individual events. TAES takes into account the duration of the overlap, so it is the closest method to EPOCH in this regard. Since OVLP and TAES both score overlapping events independently, we also expect these two methods to be correlated (sensitivity: 0.78; specificity: 0.95). ATWV on the other hand has fairly low correlations with the other metrics for specificity because of its stringent rules for FPs when there are multiple overlapping events. The overall highest correlation is between ATWV and OVLP for sensitivity, and OVLP and TAES for specificity. All the correlation values (Pearson's R) collected in these tables are statistically significant with the p-values < 0.001.

To understand the statistical significance of each system, we perform two-tailed Z-tests on all the recognition system pairs for sensitivity as shown in Table 8 and for specificity as shown in Table 9.

518   Cells in these tables contain entries that consist of the sensitivity/specificity differences between the
519   systems and a binary classification value (Yes/No) based on extracted p-values from the Z-test with
520   95% confidence. Here again, the data was prepared by scoring systems on individual patients. Prior

Table 8. Significance calculated for scoring metrics using Z-tests for α-value 0.05 (for sensitivity)

ATWV (Abs. sensitivity difference (%), Significant/Non-significant)					
ML Systems (Sens.)	**CNN-LSTM**	**CNN-MLP**	**HMM-LSTM**	**HMM-SDA**	**IPCA-LSTM**
CNN-LSTM(30.34%)	---	(00.82%) Y	(03.61%) Y	(00.01%) Y	(05.61%) Y
CNN-MLP(29.52%)		---	(02.79%) N	(00.83%) N	(04.79%) N
HMM-LSTM(26.73%)			---	(03.62%) N	(02.00%) N
HMM-SDA(30.35%)				---	(05.62%) N
IPCA-LSTM(24.73%)					---
DPALIGN (Abs. sensitivity difference)					
ML Systems (Sens.)	**CNN-LSTM**	**CNN-MLP**	**HMM-LSTM**	**HMM-SDA**	**IPCA-LSTM**
CNN-LSTM(32.46%)	---	(10.89%) Y	(01.31%) Y	(11.65%) Y	(03.31%) Y
CNN-MLP(43.35%)		---	(09.58%) N	(00.76%) N	(07.58%) N
HMM-LSTM(33.77%)			---	(10.34%) N	(02.00%) N
HMM-SDA(44.11%)				---	(08.34%) N
IPCA-LSTM(35.77%)					---
EPOCH (Abs. sensitivity difference)					
ML Systems (Sens.)	**CNN-LSTM**	**CNN-MLP**	**HMM-LSTM**	**HMM-SDA**	**IPCA-LSTM**
CNN-LSTM(09.78%)	---	(55.25%) N	(40.68%) N	(10.93%) Y	(41.24%) N
CNN-MLP(65.03%)		---	(14.57%) Y	(44.32%) Y	(14.01%) N
HMM-LSTM(50.46%)			---	(29.75%) Y	(00.56%) N
HMM-SDA(20.71%)				---	(30.31%) Y
IPCA-LSTM(51.02%)					---
OVLP (Abs. sensitivity difference)					
ML Systems (Sens.)	**CNN-LSTM**	**CNN-MLP**	**HMM-LSTM**	**HMM-SDA**	**IPCA-LSTM**
CNN-LSTM(30.83%)	---	(08.26%) Y	(02.14%) Y	(04.52%) Y	(02.14%) Y
CNN-MLP(39.09%)		---	(09.04%) N	(03.74%) N	(06.12%) N
HMM-LSTM(30.05%)			---	(05.30%) N	(02.92%) N
HMM-SDA(35.35%)				---	(02.38%) N
IPCA-LSTM(32.97%)					---
TAES (Abs. sensitivity difference)					
ML Systems (Sens.)	**CNN-LSTM**	**CNN-MLP**	**HMM-LSTM**	**HMM-SDA**	**IPCA-LSTM**
CNN-LSTM(12.48%)	---	(19.10%) N	(10.36%) N	(04.81%) Y	(09.64%) N
CNN-MLP(31.58%)		---	(08.74%) N	(14.29%) Y	(09.46%) N
HMM-LSTM(22.84%)			---	(05.55%) Y	(00.72%) N
HMM-SDA(17.29%)				---	(04.83%) Y
IPCA-LSTM(22.12%)					---

521 to performing Z-tests, the Gaussianity of each sample was evaluated using a KS test. All the samples
522 were confirmed as normal with p-values < 0.001.
523 From Table 8, it can be observed that, aside from the EPOCH and TAES scoring metrics, the

Table 9. Significance calculated for scoring metrics using Z-tests for α-value 0.05 (for specificity)

ATWV (Abs. specificity difference (%), Significant/Non-significant)					
ML Systems (Spec.)	**CNN-LSTM**	**CNN-MLP**	**HMM-LSTM**	**HMM-SDA**	**IPCA-LSTM**
CNN-LSTM(93.15%)	---	(27.28%) Y	(24.22%) Y	(31.77%) Y	(28.64%) Y
CNN-MLP(65.87%)		---	(03.06%) N	(04.49%) N	(01.36%) N
HMM-LSTM(68.93%)			---	(07.55%) Y	(04.42%) N
HMM-SDA(61.38%)				---	(03.13%) N
IPCA-LSTM(64.51%)					---
DPALIGN (Abs. specificity difference (%), Significant/Non-significant)					
ML Systems (Spec.)	**CNN-LSTM**	**CNN-MLP**	**HMM-LSTM**	**HMM-SDA**	**IPCA-LSTM**
CNN-LSTM(95.17%)	---	(23.68%) Y	(22.18%) Y	(28.30%) Y	(25.58%) Y
CNN-MLP(71.49%)		---	(01.50%) N	(04.62%) Y	(01.90%) N
HMM-LSTM(72.99%)			---	(06.12%) Y	(03.40%) N
HMM-SDA(66.87%)				---	(02.72%) Y
IPCA-LSTM(69.59%)					---
EPOCH (Abs. specificity difference (%), Significant/Non-significant)					
ML Systems (Spec.)	**CNN-LSTM**	**CNN-MLP**	**HMM-LSTM**	**HMM-SDA**	**IPCA-LSTM**
CNN-LSTM(99.84%)	---	(08.29%) N	(05.02%) N	(01.62%) N	(05.75%) N
CNN-MLP(91.55%)		---	(03.27%) N	(06.67%) N	(02.54%) N
HMM-LSTM(94.82%)			---	(03.40%) N	(00.73%) N
HMM-SDA(98.22%)				---	(04.13%) N
IPCA-LSTM(94.09%)					---
OVLP (Abs. specificity difference (%), Significant/Non-significant)					
ML Systems (Spec.)	**CNN-LSTM**	**CNN-MLP**	**HMM-LSTM**	**HMM-SDA**	**IPCA-LSTM**
CNN-LSTM(96.86%)	---	(20.02%) Y	(16.33%) Y	(23.51%) Y	(19.29%) Y
CNN-MLP(76.84%)		---	(03.69%) N	(03.49%) Y	(00.73%) N
HMM-LSTM(80.53%)			---	(07.18%) Y	(02.96%) N
HMM-SDA(73.35%)				---	(04.22%) Y
IPCA-LSTM(77.57%)					---
TAES (Abs. specificity difference (%), Significant/Non-significant)					
ML Systems (Spec.)	**CNN-LSTM**	**CNN-MLP**	**HMM-LSTM**	**HMM-SDA**	**IPCA-LSTM**
CNN-LSTM(95.24%)	---	(31.21%) Y	(24.83%) Y	(29.20%) Y	(28.60%) Y
CNN-MLP(64.03%)		---	(06.38%) N	(02.01%) Y	(02.61%) N
HMM-LSTM(70.41%)			---	(04.37%) Y	(03.77%) N
HMM-SDA(66.04%)				---	(00.60%) Y
IPCA-LSTM(66.64%)					---

524   differences between the CNN-LSTM system and all the other systems are statisticially significant
525   (rejecting the null hypothesis with p-values < 0.05). On the other hand, the EPOCH and TAES metrics
526   fail to reject the null hypothesis for CNN-LSTM. According to these metrics, the performance of

527 HMM-SDA is statistically different from the other systems, confirming its poor performance. This
528 can also be observed from EPOCH/TAES results of Table 2.

529 Table 9 indicates a different trend than Table 8. The EPOCH metric fails to reject null-hypothesis for
530 all the systems. Since specificity is calculated from TN and FP values, for an evaluation set 167 hours
531 in duration and an epoch size 0.25, a few thousand seconds of FPs do not make any significant
532 difference in terms of specificity. This can also be directly observed in Table 2 where the specificity
533 of all systems according to the EPOCH metric is always greater than 90%. The huge difference
534 between the duration of background and seizure events is the primary reason for such high
535 specificities. On the other hand, the OVLP and TAES metrics completely agree with each other's Z-
536 test results for specificity.

537 **5. Conclusions**

538 Standardization of scoring metrics is an extremely important step for a research community to take
539 in order to make progress on machine learning problems such as automatic interpretation of EEGs.
540 There has been a lack of standardization in most bioengineering fields. Popular metrics such as
541 sensitivity and specificity do not completely characterize the problem and neglect the importance
542 that FA rate plays in achieving clinically acceptable solutions. In this paper, we have compared
543 several popular scoring metrics and demonstrated the value of considering the accuracy of time
544 alignments in the overall assessment of a system. We have proposed the use of a new metric, TAES
545 scoring, which is consistent with popular scoring approaches such as OVLP, but provides more
546 accurate assessments by producing fractional scores for recognition of events based on the degree of
547 match in the time alignments. We have also demonstrated the efficacy of an existing metric, ATWV,
548 that is popular in the speech recognition community.

549 We have also not discussed the extent to which we can tune these metrics by weighting various types
550 of errors based on feedback from clinicians and other 'customers' of the technology. Optimization of
551 the metric is a research problem in itself, since many considerations, including usability of the
552 technology and a broad range of applications, must be involved in this process. Our informal
553 attempts to optimize ATWV and OVLP for seizure detection have not yet produced significantly
554 different results than what was presented here. Feedback from clinicians has been consistent that FA
555 rate is perhaps the single most important measure once sensitivity is above approximately 75%. As
556 we move more technology into operational environments we expect to have more to contribute to
557 this research topic.

558 Finally, the Python implementation of these metrics is available at the project web site:
559 https://www.isip.piconepress.com/projects/tuh_eeg/downloads/nedc_eval_eeg.   Readers   are
560 encouraged to refer to the software for detailed questions about the specific implementations of these
561 algorithms and the tunable parameters available.

580

References

1. 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. https://doi.org/10.1111%2Fj.1468-1331.2009.02936.x.

2. Scheuer, M. L., Bagic, A., & Wilson, S. B. (2017). Spike detection: Inter-reader agreement and a statistical Turing test on a large data set. Clinical Neurophysiology, 128(1), 243–250. https://doi.org/10.1016/j.clinph.2016.11.005.

3. Gadhoumi, K., Lina, J.-M., Mormann, F., & Gotman, J. (2016). Seizure prediction for therapeutic devices: A review. Journal of Neuroscience Methods, 260 (Supplement C), 270–282. https://doi.org/10.1016/j.jneumeth.2015.06.010.

4. 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. https://doi.org/10.1016/S1388-2457(03)00212-8.

5. Gotman, J., Flanagan, D., Zhang, J., & Rosenblatt, B. (1997). Automatic seizure detection in the newborn: Methods and initial evaluation. Electroencephalography and Clinical Neurophysiology, 103(3), 356–362. https://doi.org/10.1016/S0013-4694(97)00003-9.

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

7. Cvach Maria, M. (2014). Managing hospital alarms. Nursing Critical Care, 9(3), 13–27. https://doi.org/10.1097/01.CCN.0000446255.81392.b0.

8. Bridi, A. C., Louro, T. Q., & Da Silva, R. C. L. (2014). Clinical Alarms in intensive care: implications of alarm fatigue for the safety of patients. Revista Latino-Americana de Enfermagem, 22(6), 1034. https://doi.org/10.1590/0104-1169.3488.2513.

9. Hu, P. (2015). Reducing False Alarms in Critical Care. Presented at the Working Group on Neurocritical Care Informatics, Neurocritical Care Society Annual Meeting. Scottsdale, Arizona, USA. Not available online.

10. Hambling, B. (2013). User Acceptance Testing A step-by-step guide. (P. van Goethem, Ed.), User Acceptance Testing. Swindon, United Kingdom: BCS Learning & Development Limited. https://www.amazon.com/User-Acceptance-Testing-Step-Step/dp/1780171676.

11. Banchs, R., Bonafonte, A., & Perez, J. (2006). Acceptance Testing of a Spoken Language Translation System. Proceedings of LREC (p. 106). Genoa, Italy. http://www.lrec-conf.org/proceedings/lrec2006/pdf/60_pdf.pdf.

12. Picone, J., Doddington, G., & Pallett, D. (1990). Phone-mediated word alignment for speech recognition evaluation. IEEE Transactions on Acoustics, Speech and Signal Processing, 38(3), 559–562. https://doi.org/10.1109/29.106877.

13. Michel, M., Joy, D., Fiscus, J. G., Manohar, V., Ajot, J., & Barr, B. (2017). Framework for Detection Evaluation (F4DE). https://github.com/usnistgov/F4DE.

14. Altman, D. G., & Bland, J. M. (1994). Diagnostic Tests 1: Sensitivity And Specificity. British Medical Journal, 308(6943), 1552. https://doi.org/10.1136/bmj.308.6943.1552.

15. Wozencraft, J. M., & Jacobs, I. M. (1965). Principles of Communication Engineering. New York City, New York, USA: Wiley. https://books.google.com/books/about/Principles_of _communication_engineering.html?id=4ORSAAAAMAAJ.

16. Martin, A., Doddington, G., Kamm, T., Ordowski, M., & Przybocki, M. (1997). The DET curve in assessment of detection task performance. Proceedings of Eurospeech (pp. 1895–1898). Rhodes, Greece. https://doi.org/10.1.1.117.4489.

17. Wang, Y.-Y., Acero, A., & Chelba, C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (pp. 577–582). Saint Thomas, Virgin islands: IEEE. https://doi.org/10.1109/ASRU.2003.1318504.

18. Mostefa, D., Hamin, O., & Choukri, K. (2006). Evaluation of Automatic Speech Recognition and Speech Language Translation within TC-STAR: Results from the first evaluation campaign. Proceedings of the International Conference on Language Resources and Evaluation (pp. 149–154). Genoa, Italy. http://lrec-conf.org/proceedings/lrec2006/pdf/813_pdf.pdf.

634   19.   Wilson, S.B., Scheuer, M.L., Plummer, C., Young, B., & Pacia, S. (2003) Seizure detection: correlation of
635   human experts. Clinical Neurophysiology, 114(11), 2156–64. https://doi.org/ 10.1016/S1388-2457(03)00212-
636   8.
637   20.   Kelly, K.M., Shiau, D.S., Kern, R.T., Chien, J.H., Yang, M.C.K., Yandora, K.A. Valeriano, J. P.,   Halford, J.
638   J., & Sackellares, J. C. (2010). Assessment of a scalp EEG-based automated seizure detection system. Clinical
639   Neurophysiology, 121(11), 1832–43. https://doi.org/10.1016/j.clinph.2010.04.016.
640   21.   Baldassano, S., Wulsin, D., Ung, H., Blevins, T., Brown, M-G., Fox, E., & Litt, B. (2016). A novel seizure
641   detection algorithm informed by hidden Markov model event states. Journal of Neural Engineering,13(3),
642   36011. https://doi.org/10.1088/1741-2560/13/3/036011.
643   22.   Winterhalder, M., Maiwald, T., Voss, H. U., Aschenbrenner-Scheibe, R., Timmer, J., & Schulze-Bonhage, A.
644   (2003). The seizure prediction characteristic: a general framework to assess and compare seizure prediction
645   methods. Epilepsy and Behavior, 4(3), 318–325. https://doi.org/10.1016/S1525-5050(03)00105-7.
646   23.   Wegmann, S., Faria, A., Janin, A., Riedhammer, K., & Morgan, N. (2013). The TAO of ATWV: Probing the
647   mysteries of keyword search performance. Proceedings of the IEEE Workshop on Automatic Speech
648   Recognition   and   Understanding   (ASRU).   Olomouc,   Czech   Republic: IEEE.
649   https://doi.org/10.1109/ASRU.2013.6707728.
650   24.   Fiscus, J., Ajot, J., Garofolo, J., & Doddingtion, G. (2007). Results of the 2006 Spoken Term Detection
651   Evaluation. In Proceedings of the SIGIR 2007 Workshop: Searching Spontaneous Conversational Speech
652   (pp. 45–50). Amsterdam, Netherlands. https://www.nist.gov/ publications/results-2006-spoken-term-
653   detection-evaluation.
654   25.   Japkowicz, N., & Shah, M. (2014). Evaluating Learning Algorithms: a classification perspective. New York
655   City, New York, USA: Cambridge University Press. https://doi.org/10.1017/CBO9780511921803.
656   26.   Confusion matrix. (n.d.). Retrieved October 31, 2017. https://en.wikipedia.org/wiki/ Confusion_matrix.
657   27.   Liu, A., Hahn, J. S., Heldt, G. P., & Coen, R. W. (1992). Detection of neonatal seizures through computerized
658   EEG   analysis.   Electroencephalography   and   Clinical   Neurophysiology,   82(2),   32–37.
659   https://doi.org/10.1016/0013-4694(92)90179-L.
660   28.   Navakatikyan, M. A., Colditz, P. B., Burke, C. J., Inder, T. E., Richmond, J., & Williams, C. E. (2006). Seizure
661   detection algorithm for neonates based on wave-sequence analysis. Clinical Neurophysiology, 117(6),
662   1190–1203. http://dx.doi.org/10.1016/j.clinph.2006.02.016.
663   29.   Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2017). The Microsoft 2017 Conversational
664   Speech Recognition System. https://arxiv.org/abs/1708.06073.
665   30.   Gotman, J., Flanagan, D., Zhang, J., & Rosenblatt, B. (1997). Automatic seizure detection in the newborn:
666   Methods and initial evaluation. Electroencephalography and Clinical Neurophysiology, 103(3), 356–362.
667   https://doi.org/10.1016/S0013-4694(97)00003-9.
668   31.   Wilson, S. B., Scheuer, M. L., Plummer, C., Young, B., & Pacia, S. (2003). Seizure detection: correlation of
669   human   experts.   Clinical   Neurophysiology,   114(11),   2156–2164.   https://doi.org/10.1016/S1388-
670   2457(03)00212-8.
671   32.   Mason, S. J., & Graham, N. E. (2002). Areas beneath the relative operating characteristics (ROC) and relative
672   operating levels (ROL) curves: Statistical significance and interpretation. Quarterly Journal of the Royal
673   Meteorological Society, 128(584), 2145–2166. https://doi.org/10.1256/003590002320603584.
674   33.   Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic
675   Test   Evaluation.   Caspian   Journal   of   Internal   Medicine,   4(2),   627–635.
676   http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/.
677   34.   Fiscus, J. G., & Chen, N. (2013). Overview of the NIST Open Keyword Search 2013 Evaluation Workshop.
678   Bethesda, Maryland, USA. http://ws680.nist.gov/publication/get_pdf.cfm?pub_id= 914517.
679   35.   National Institute of Standards and Technology (NIST) (2017). Speech Recognition Scoring Toolkit.
680   https://github.com/usnistgov.
681   36.   McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia Medica, 22(3), 276–282. Retrieved
682   from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/.
683   37.   National Institute of Standards and Technology (NIST) (2017). Framework for Detection Evaluation.
684   https://github.com/usnistgov/F4DE.
685   38.   Halford, J. J., Shiau, D., Desrochers, J. A., Kolls, B. J., Dean, B. C., Waters, C. G., … LaRoche, S. M. (2015).
686   Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG

687   recordings. Clinical Neurophysiology : Official Journal of the International Federation of Clinical
688   Neurophysiology, 126(9), 1661–9. https://doi.org/10.1016/j.clinph.2014.11.008.

689   39.   Golmohammadi, M., Shah, V., Lopez, S., Ziyabari, S., Yang, S., Camaratta, J., … Picone, J. (2017). The TUH
690   EEG Seizure Corpus. In Proceedings of the American Clinical Neurophysiology Society Annual Meeting
691   (p.   1).   Phoenix,   Arizona,   USA: American   Clinical   Neurophysiology   Society.
692   https://www.isip.piconepress.com/publications/conference_presentations/2017/acns/tuh_eeg_seizures.

693   40.   Golmohammadi, M., Shah, V., Obeid, I., & Picone, J. (2019). Deep Learning Approaches for Automatic
694   Analysis of EEGs. In S.-M. Chan & W. Pedrycz (Eds.), Deep Learning: Algorithms and Applications (1st
695   ed., p. 76). New York, New York, USA: Springer-Verlag. https://doi.org/10.3389/fnhum.2019.00076.

696   41.   Golmohammadi, M., Ziyabari, S., Shah, V., Obeid, I., & Picone, J. (2018). Deep Architectures for Spatio-
697   Temporal Modeling: Automated Seizure Detection in Scalp EEGs. In Proceedings of the IEEE International
698   Conference on Machine Learning and Applications (ICMLA) (pp. 1–6). Orlando, Florida, USA.
699   https://doi.org/http://dx.doi.org/10.1109/ICMLA.2018.00118.

700