Objective Evaluation Metrics for Automatic Classification of EEG Events

Vinit Shah[1], Meysam Golmohammadi[1, 2], Iyad Obeid[1], Joseph Picone[1]

[1] The Neural Engineering Data Consortium, Temple University
Philadelphia, Pennsylvania, USA,
[2] Internet Brands, El Segundo, California, USA

Abstract. The evaluation of machine learning algorithms in biomedical fields for applications involving sequential data lacks both rigor and standardization. Common quantitative scalar evaluation metrics such as sensitivity and specificity can often be misleading and not accurately integrate application requirements. Evaluation metrics must ultimately reflect the needs of users yet be sufficiently sensitive to guide algorithm development. For example, feedback from critical care clinicians who use automated event detection software in clinical applications has been overwhelmingly emphatic that a low false alarm rate, typically measured in units of the number of errors per 24 hours, is the single most important criterion for user acceptance. Though using a single metric is not often as insightful as examining performance over a range of operating conditions, there is, nevertheless, a need for a single scalar figure of merit.

In this chapter, we discuss the deficiencies of existing metrics for a seizure detection task and propose several new metrics that offer a more balanced view of performance. We demonstrate these metrics on a seizure detection task based on the TUH EEG Seizure Corpus. We introduce two promising metrics: (1) a measure based on a concept borrowed from the spoken term detection literature, Actual Term-Weighted Value, and (2) a new metric, Time-Aligned Event Scoring (TAES), that accounts for the temporal alignment of the hypothesis to the reference annotation. We demonstrate that state of the art technology based on deep learning, though impressive in its performance, still needs significant improvement before it will meet very strict user acceptance guidelines.

Keywords: Electroencephalogram, EEGs, scoring, evaluation metrics, machine learning, seizure detection

1 Introduction

Electroencephalograms (EEGs) are the primary means by which physicians diagnose and manage brain-related illnesses such as epilepsy, seizures and sleep disorders [1]. Automatic interpretation of EEGs by computer has been extensively studied for the past 40 years [2]-[6] with mixed results. Even though many research systems report impressive levels of accuracy in research publications, widespread adoption of commercial technology has yet to happen in clinical settings primarily due to the high false alarm rates of these systems [7]-[9]. 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.

There are in general two ways to evaluate machine learning technology: user acceptance testing [10][11] and objective performance metrics based on annotated reference data [12][13]. User acceptance testing is slow, time-consuming and expensive. It has never been a practical way to guide technology development because algorithm developers need rapid turnaround times on evaluations. Hence evaluations using objective performance metrics, such as sensitivity and specificity, are common in the machine learning field [14]-[16]. With this approach, it is very important to have a rich evaluation dataset and a performance metric that correlates well with user and application needs. The metric must have a certain level of granularity so that small differences in algorithms can be investigated and parameter optimizations can be evaluated. For example, in speech recognition applications, word error rate has been used for many years because it correlates well with user acceptance testing but provides the necessary level of granularity to guide technology development. Despite many years of research focused on finding better performance metrics [17][18], word error rate remains a valid metric for technology development and assessment in speech recognition.

Sequential pattern recognition applications, such as speech recognition, keyword search or EEG analysis, require additional considerations. Data are not simply assessed with an overall judgment (e.g., “did a seizure occur somewhere in this file?”). Instead, the locality of the hypothesis must be considered – to what extent did the start and end times of the hypothesis match the reference transcription. This is a complex issue since a hypothesis can partially overlap with the reference annotation, and a consistent mechanism for scoring such events must be adopted.

Unfortunately, there is no such standardization in the EEG literature. For example, Wilson et al. [4] advocates using a term-based metric involving sensitivity and specificity. A term was defined as a connection of consecutive decisions of the same type of event. A hypothesis is counted as a true positive when it overlaps with one or more reference annotations. A false positive corresponds to an event in which a hypothesis annotation does not overlap with any of the reference annotations. Kelly et al. [19] recommends using a metric that measures sensitivity and false alarms. A hypothesis is considered a true positive when time of detection is within two minutes of the seizure onset. Otherwise it is considered a false positive. Baldassano et al. [20] uses an epoch-based metric that measures false positive and negative rates as well as latency. The development, evaluation and ranking of various machine learning approaches is highly dependent on the choice of a metric.

A large class of bioengineering problems, including seizure detection, involve prediction as well as classification. In prediction problems, we are often concerned with how far in advance of an event we can predict an outcome. The accuracy of a prediction varies with latency. By convention, we refer to negative latency as prediction before the event has occurred. Positive latency means a system outputs a hypothesis after an event has occurred. It is not uncommon machine learning systems to have significant amounts of latency – often tens of seconds for EEG analysis. Similarly, prediction of a seizure before the seizure has occurred is an extremely valuable technology, especially if you can predict the onset of a seizure long in advance (e.g., ten of minutes). This gives healthcare providers a chance to perform a medical intervention as well as allows the patient to make necessary preparations for a medical emergency.

Measuring performance as a function of latency adds some complexity to the process. Winterhalder et al. [21] have studied this problem extensively and argue for a scoring based on long-term considerations. In this chapter, we are not concerned with these types of prediction problems. We are focused mainly on assessing the accuracy of classification of events and assessing the proximity of these classifications to the actual event. We refer to this as time-aligned scoring.

Therefore, in this chapter, we analyze several popular scoring metrics and discuss their strengths and weaknesses on sequential decoding problems. We introduce several alternatives, such as the Actual Term-Weighted Value (ATWV) [22][23] that have proven successful in other fields, and discuss their relevance to EEG applications. We present a comparison of performance for several systems using these metrics and discuss how this correlates with a proxy for overall user acceptance involving a combination of sensitivity and false alarm rate.

Comparing systems using a single operating point is, of course, not always correct. It is quite possible the systems are simply operating at different points on what is known as their Receiver Operating Characteristic (ROC) curve. This was a problem well-studied in the mid-1960's with the emergence of communication theory [15]. In machine learning, we often prefer to analyze systems using a Detection Error Tradeoff (DET) curve [23]-[25] . These curves provide a holistic view of performance but make it difficult to tune a system at a specific operating point. We will also briefly discuss holistic measures based on DET analysis.

2 Basic Error Types and Associated Derived Measures

Researchers in biomedical fields typically report performance in terms of sensitivity and specificity [26]. In a two-class classification problem such as seizure detection, we can define four types of errors:

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

False positives, also known as false alarms, play a very important role in sequential decoding applications since they tend to dominate performance considerations. Throughout this chapter, we will refer to the false alarm (FA) rate, which is simply the number of false positives divided by the total amount of data measured in units of time. We typically compute FAs/24 hrs. – the number of false alarms per day. This is a useful figure of merit for critical care applications in healthcare.

There are a large number of auxiliary measures that are used extensively in the literature that can be calculated from these four basic quantities. These are summarized concisely in [27]. For example, in information retrieval applications, systems are often evaluated using:

$$Sensitivity\ (Recall) = (TP/(TP+FN)) \tag{1}$$

$$Specificity\ (Selectivity) = (TN/(TN+FP)) \tag{2}$$

$$Accuracy = ((TP+TN)/(TP+FN+TN+FP)) \tag{3}$$

$$Precision = (TP/(TP+FP)) \tag{4}$$

More recently, integrated measures such as the F1 score and the Matthews correlation coefficient (MCC) [28] have become popular for tasks ranging from information retrieval to binary classification:

$$F1 = ((2 \times Precision \times Recall)/(Precision + Recall)) \tag{5}$$

$$MCC = \big((TP \times TN) - (FP \times FN)\big)\Big/\sqrt{\big((TP+FP)(TP+FN)(TN+FP)(TN+FN)\big)} \tag{6}$$

In the field of machine translation, the bilingual evaluation understudy (BLEU) metric, which measures the similarity between two strings of text, was one of the first objective evaluation metrics to claim a high correlation with human judgements of quality [29]. However, none of these measures address the time scale over which the scoring must occur, which is critical in the interpretation of these measures for many real-time bioengineering applications where the time alignment of the reference event and the hypothesized event are important, and spurious hypotheses play a critical role in overall system performance. Localization of events in time is integral to how healthcare providers use technology in clinical settings. Hence, evaluation metrics must take into account the accuracy of these localizations.

In some applications, it is preferable to score every unit of time. With multichannel signals, such as EEGs, scoring for each channel for each unit of time is appropriate since significant events such as seizures occur on a subset of the channels present in the signal. However, it is more common in the literature to simply score a summary decision per unit of time that is based on an aggregation of the per-channel inputs (e.g., a majority vote). We refer to this type of scoring as epoch-based [30][31].

An alternative, that is more common in speech and image recognition applications, is term-based [23][32], in which we consider the start and stop time of the event, and each event identified in the reference annotation is counted once. There are fundamental differences between the two conventions. For example, one event containing many epochs will count more heavily in an epoch-based scoring scenario. Epoch-based scoring generally weights the duration of an event more heavily since each unit of time is assessed independently.

Time-aligned scoring is essential to the evaluation of sequential decoding systems. But to implement such scoring in a meaningful way, there needs to be universal agreement on how to assess overlap between the reference and the hypothesis. For example, Figure 1 demonstrates a typical issue in scoring. The machine learning system correctly detected 5 seconds of a 10-sec event. Essentially 50% of the event is correctly detected,

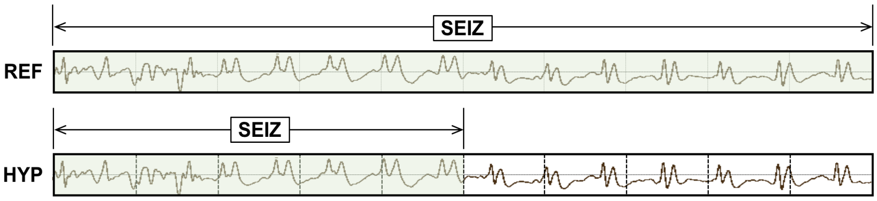

Figure 1. A hypothesis (HYP) has a 50% overlap with the reference (REF).

but how that is reflected in the scoring depends on the specific metric. Epoch-based scoring with an epoch duration of 1 sec would count 5 FN errors and 5 TP detections. Term-based scoring would potentially count this as a correct recognition depending on the way overlaps are scored.

Term-based metrics score on an event basis and do not count individual frames. A typical approach for calculating errors in term-based scoring is the Any-Overlap Method (OVLP) [5][33]. TPs are counted when the hypothesis overlaps with reference annotation. FPs correspond to situations in which a hypothesis does not overlap with the reference. The metric ignores the duration of the term in the reference annotation. In Figure 2, we demonstrate two extreme cases for which the OVLP metric fails. In each case, 90% of the event is incorrectly scored. In Example 1, the system does not detect approximately 9 seconds of a seizure event, while in Example 2, the system incorrectly labels an additional 9 seconds of time as seizure. OVLP is considered a very permissive way of scoring, resulting in artificially high sensitivities. In Figure 2, the OVLP metric will score both examples as 100% TP. These kinds of significant differences in scoring, and in the interpretation of the results, necessitate a deeper look at the characteristics of several popular evaluation metrics, and motivate the need for industry-wide standardized scoring. That is the focus of this book chapter.

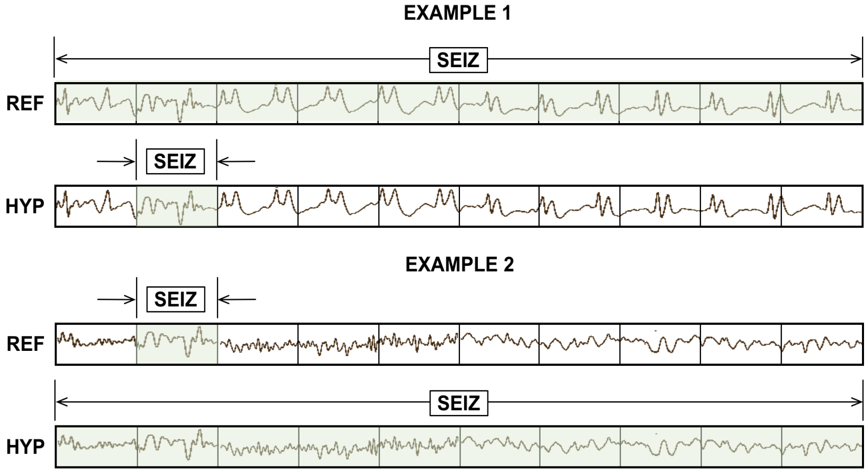

Figure 2. TP scores for the Any-Overlap method are 100% even though large portions of the event are missed.

3 Evaluation Metrics

The proper balance between sensitivity and FA rate is often application specific and has been studied extensively in a number of research communities. For example, evaluation of voice keyword search technology was carefully studied in the Spoken Term Detection (STD) evaluations conducted by NIST [22][23][34]. These evaluations resulted in the introduction of a single metric, ATWV, to address concerns about tradeoffs for the different types of errors that occur in voice keyword search systems. Despite being popular in the voice processing community, ATWV has not been used in the bioengineering community.

Therefore, in this chapter, we present a detailed comparison of five important scoring metrics popular in a wide range of machine learning communities, including ATWV. These are briefly described below:

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.

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 measures such as sensitivity and specificity are computed, as shown in Eqs. (1)-(6). 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 also include two derived measures in our analysis:

6. *Inter-Rater Agreement (IRA):* uses EPOCH scoring to estimate errors, and calculates Cohen's Kappa coefficient [31] using the measured TP, TN, FP and FN.
7. *Area Under the Curve (AUC):* Reduces a ROC or DET curve to a single scalar figure of merit by measuring the area encompassed by the curve.

IRA is popular for comparing the variability in human annotations when manually annotating reference data. We consider this a derived measure because it relies on one of the first five measures to estimate errors. Similarly, AUC relies on the generation of a ROC or DET curve, which in turn depend on one of the first five measures to estimate errors.

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 timescale for illustrative purposes and were carefully selected because they demonstrate the strengths and weaknesses of the algorithms we are evaluating.

3.1 NIST Actual Term-Weighted Value (ATWV)

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

The metric accepts as input a list of N-tuples representing the hypotheses for the system being evaluated. Each of these N-tuples consists of a start time, end time and system detection score. These entries are matched to the reference annotations using an objective function that accounts for both temporal overlap between the reference and hypotheses and the detection scores assigned by the system being evaluated. These detection scores are often likelihood or confidence scores [22]. The probabilities of miss and FA errors at a detection threshold θ are computed using:

$$P_{Miss(kw,\theta)} = 1 - {}^{N_{Correct(kw,\theta)}}\!/_{N_{Ref(kw)}} \,, \tag{7}$$

$$P_{FA(kw,\theta)} = {}^{N_{Spurious(kw,\theta)}}\!/_{N_{NT(kw)}} \,, \tag{8}$$

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

A term-weighted value is then computed that specifies a trade-off between misses and FAs. ATWV is defined as the value of TWV at the system's chosen detection threshold. Using a predefined constant, β, that was optimized experimentally ($\beta = 999.9$) [23], ATWV is computed using:

$$TWV_{(kw,\theta)} = 1 - P_{Miss(kw,\theta)} - \beta\, P_{FA(kw,\theta)} \,. \tag{9}$$

A standard implementation of this approach is available at [35].

This metric has been widely used throughout the human language technology community for almost 20 years. This is a very important consideration in standardizing such a metric – researchers are using a common shared software implementation that ensures there are no subtle implementation differences in scoring software implementation between sites or researchers. There are always numerous parameters associated with this type of software and the only ways to make sure algorithms are producing identical results are (1) the existence of a common (open source) software package or (2) the distribution of a detailed set of regression tests that establish the equivalency of the implementations. The former has been a standard methodology for 40 years in the human language technology community, but the bioengineering communities have not quite achieved this level of standardization yet.

To demonstrate the features of this approach, consider the case shown in Figure 3. The hypothesis for this segment consists of several short seizure events while the reference consists of one long event. The ATWV metric will assign a TP score of 100% because the midpoint of the first event in the hypothesis annotation is mapped to the long seizure event in the reference annotation. This is somewhat generous given that 50% of the event was not detected. The remaining 5 events in the hypothesis annotation are counted as false positives. The ATWV metric is relatively insensitive to the duration of the reference event, though the 5 false positives will lower the overall performance of the system. The important issue here is that the hypothesis correctly detected about 70% of the seizure event, and yet because of the large number of false positives, it will be penalized heavily.

In Figure 4 we demonstrate a similar case in which the metric penalizes the hypothesis for missing three seizure events in the reference. Approximately 50% of the segment is correctly identified. This type of scoring penalizing repeated events that are part of a larger event in the reference might make sense in an application like voice keyword search because in human language each word hypothesis serves a unique purpose in the overall understanding of the signal. However, for a two-class event detection problem such as seizure detection, such scoring too heavily penalizes the hypothesis for splitting a long event into a series of short events.

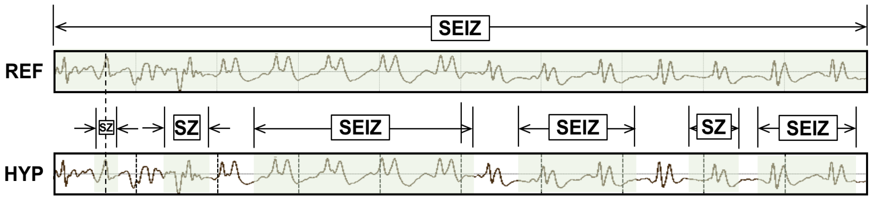

Figure 3. ATWV scores this segment as 1 TP and 5 FPs.

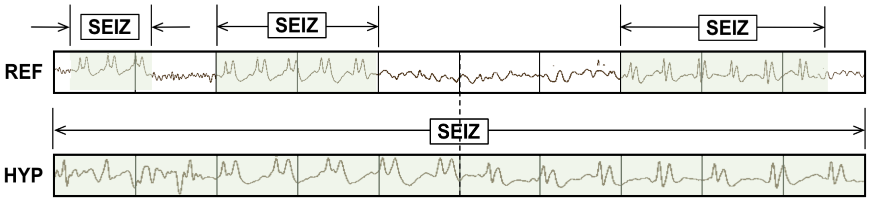

Figure 4. ATWV scores this segment as 0 TP and 3 FN events.

3.2 Dynamic Programming Alignment (DPALIGN)

The DPALIGN metric essentially performs a minimization of an edit distance (the Levenshtein distance) [12] to map the hypothesis onto the reference. DPALIGN determines the minimum number of edits required to transform the hypothesis string into the reference string. Given two strings, the source string $X = [x_1, x_2, \dots, x_n]$ of length n, and target string $Y = [y_1, y_2, \dots, y_m]$ of length m, we define $d_{i,j}$, which is the edit distance between the substring $x_1 : x_i$ and $y_1 : y_j$, as:

$$d_{i,j} = \begin{cases} d_{i-1,j} + del \\ d_{i,j-1} + ins \\ d_{i-1,j-1} + sub \end{cases}, \tag{10}$$

The quantities being measured here are often referred to as substitution (sub), insertion (ins) and deletion (del) penalties. For this study, these three penalties are assigned equal weights of 1. A dynamic programming algorithm is used to find the optimal alignment between the reference and hypothesis based on these weights. Though there are versions of this metric that perform time-aligned scoring in which both the reference and hypothesis must include start and end times, this metric is most commonly used without time alignment information.

The metric is best demonstrated using the two examples shown in Figure 5. In the first example, the reference annotation has a series of 7 events, while the hypothesis contains 5 events. The hypothesis substitutes background for the second seizure event, omits the third seizure event and the last background event. Hence, there are a total of three errors: two deletions and one substitution. In the second example, the reference annotation and hypothesis have been swapped to demonstrate the symmetry of the error calculations. The hypothesis generated two insertions and one substitution.

In practice, there are often multiple alignments that makes sense based only on the labels. As long as the algorithm is consistent about its choices, scoring will be fine. To accurately resolve such ambiguities, the actual endpoints of the hypotheses must be compared to the endpoints in the reference annotations. NIST distributes the ability to score this way, often referred to as time-aligned scoring, in their open source package [35]. But this scoring mode is a little more complicated from a data interfacing point of view and has not been historically as popular.

```
   Ref: bckg seiz SEIZ SEIZ bckg seiz bckg
   Hyp: bckg seiz BCKG **** bckg seiz ****
(Hits: 4 Sub: 1 Ins: 0 Del: 2 Total Errors: 3)

   Ref: bckg seiz BCKG **** bckg seiz ****
   Hyp: bckg seiz SEIZ SEIZ bckg seiz bckg
(Hits: 4 Sub: 1 Ins: 2 Del: 0 Total Errors: 3)
```

Figure 5. DPALIGN aligns symbol sequences based on edit distance, ignoring the actual time alignments present in the reference annotation and the system output.

Though this type of scoring might at first seem highly inaccurate since it ignores time alignments of the hypotheses, it has been surprisingly effective in scoring machine learning systems in sequential data applications (e.g., speech recognition) [12][16][23].

3.3 Epoch-Based Sampling (EPOCH)

Epoch-based scoring uses a metric that treats the reference and hypothesis as signals. These signals are sampled at a fixed epoch duration. The corresponding label in the reference is compared to the hypothesis. Similar to DPALIGN, substitutions, deletions and insertion errors are tabulated with an equal weight of 1 for each type of error. This process is depicted in Figure 6. Epoch-based scoring requires that the entire signal be annotated (every second of the signal must be accounted for in the reference and hypothesis annotations), which is normally the case for sequential decoding evaluations. It attempts to account for the amount of time the two annotations overlap, so it directly addresses the inconsistencies demonstrated in Figures 3 and 4.

One important parameter to be tweaked in this algorithm is the frequency with which we sample the two annotations, which we refer to as the scoring epoch duration. It is ideally set to an amount of time smaller than the unit of time used by the classification system to make decisions. For example, the hypothesis in Figure 6 contains decisions made for every 1 sec of data. The scoring epoch duration should be set less than 1 sec. We set this parameter to 0.25 sec for most of our work because our analysis system epoch duration is typically 1 sec. We find in situations like this the results are not overly sensitive to the choice of the epoch duration as long as it is below 1 sec. This parameter simply controls the precision used to assess the accuracy of segment boundaries.

Because EPOCH scoring samples the annotations at fixed time intervals, it is inherently biased to weigh long seizure events more heavily. For example, if a signal contains one extremely long seizure event (e.g., 1000 secs) and two short events (e.g., each 10 secs in duration), the accuracy with which the first event is detected will dominate the overall scoring. Since seizure events can vary dramatically in duration, this is a cause for concern.

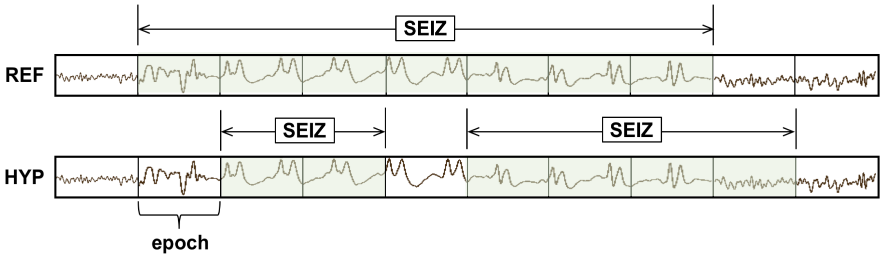

Figure 6. EPOCH scoring directly measures the similarity of the time-aligned annotations. TP, FN and FP are 5, 2 and 1 respectively.

3.4 Any-Overlap Method (OVLP)

In Section 2, we briefly introduced the OVLP metric and indicated it was a popular choice in the neuroengineering community [5][33]. OVLP is a more permissive metric that tends to produce much higher sensitivities. If an event is detected in close proximity to a reference event, the reference event is considered correctly detected. If a long event in the reference annotation is detected as multiple shorter events in the hypothesis, the reference event is also considered correctly detected. Multiple events in the hypothesis annotation corresponding to the same event in the reference annotation are not typically counted as FAs. Since the FA rate is a very important measure of performance in critical care applications, this is another cause for concern.

The OVLP scoring method is demonstrated in Figure 7. It has one significant tunable parameter – a guard band that controls the degree to which a misalignment is still considered as a correct match. In this study, we use a fairly strict setting for this parameter – 1 ms. This has the effect of requiring some overlap between the two events in time – essentially a guard band of zero. The guard band needs to be tuned based on the needs of the application. Sensitivity generally increases as the guard band is increased.

3.5 Time-Aligned Event Scoring (TAES)

Though EPOCH scoring directly measures the amount of overlap between the annotations, there is a possibility that this metric also too heavily weighs single long events. Seizure events can vary in duration from a few seconds to an hour. In some applications, correctly detecting the number of events is as important as their duration. Hence, the TAES metric was designed as a compromise between these competing constraints. The

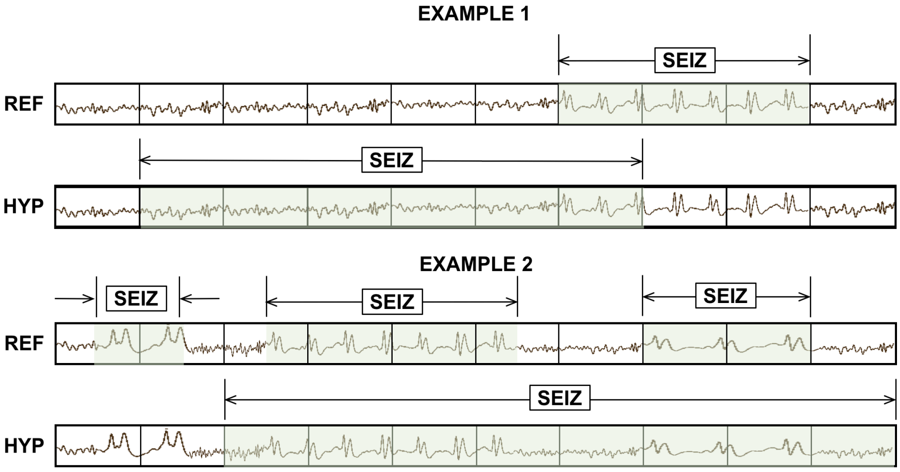

Figure 7. OVLP scoring is very permissive about the degree of overlap between the reference and hypothesis. The TP score for Example 1 is 1 with no false alarms. In Example 2, the system detects 2 out of 3 seizure events, so the TP and FN scores are 2 and 1 respectively.

essential parameters for calculation of sensitivity and specificity such as TP, TN and FP for the TAES scoring metric are defined as follows:

$$TP = \frac{H_{stop}-H_{start}}{Ref_{dur}},\ where\ R_{start} \leq H \leq R_{stop}\,, \tag{11}$$

$$TN = \frac{1-(TH_{stop}-TH_{start})}{Ref_{dur}},\ where\ R_{start} \leq H \leq R_{stop}\,, \tag{12}$$

$$FP = \begin{cases} \frac{H_{stop}-R_{stop}}{Ref_{dur}}, & if\ H_{stop} \geq R_{stop},\ H_{start} \geq R_{start}\ and\ H_{stop} - R_{stop} \leq 1, \\ \frac{R_{start}-H_{start}}{Ref_{dur}}, & if\ R_{start} \geq H_{start},\ R_{stop} \geq H_{stop}\ and\ R_{start} - H_{start} \leq 1, \\ 1, & otherwise. \end{cases} \tag{13}$$

where H and R represent the reference and hypothesis events respectively, and Ref_{dur} represents the duration of the reference events.

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

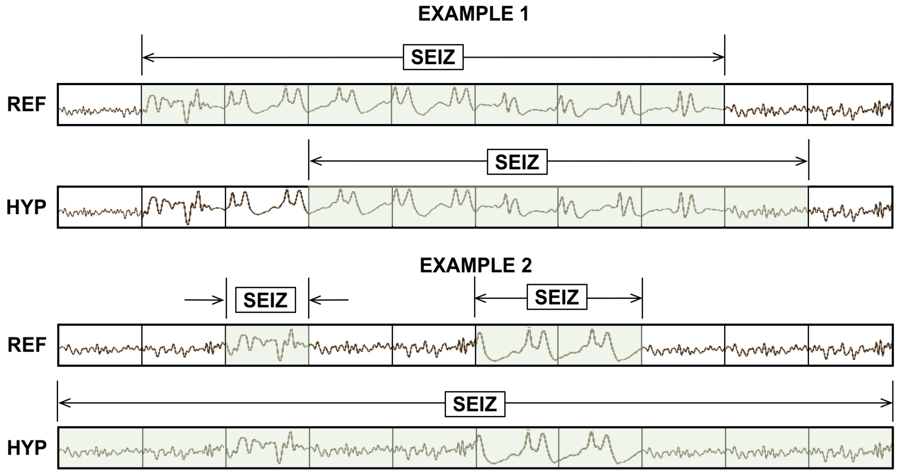

Figure 8. TAES scoring accounts for the amount of overlap between the reference and hypothesis. TAES scores Example 1 as 0.71 TP, 0.29 FN and 0.14 FP. Example 2 is scored as 1 TP, 1 FN and 1 FP.

3.6 Inter-Rater Agreement (IRA)

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

$$\kappa = \frac{p_0 - p_e}{1 - p_e}, \tag{14}$$

where p_o is the relative observed agreement among raters and p_e is the hypothetical probability of chance agreement.

The range of the Kappa coefficient is $[-1,\ 1]$ where $\kappa = 1$ corresponds to complete agreement and $\kappa = -1$ which corresponds to no agreement. It has been used extensively to assess inter-rater agreement for experts manually annotating seizures in EEG signals. Values in the range of $0.5 \le \kappa \le 0.8$ are common for these types of assessments [37]. The variability amongst experts mainly involves fine details in the annotations, such as the exact onset of a seizure. These kinds of details are extremely important for machine learning and hence we need a metric that is sensitive to small variations in the annotations. For completeness, we use this measure as a way of evaluating the amount of agreement between two annotations.

3.7 A Brief Comparison of Metrics

A simple example of how these metrics compare on a specific segment of a signal is shown in Figure 9. A 10-sec section of an EEG signal is shown subdivided into 1-sec segments. The reference has three isolated events. The system being evaluated outputs one hypothesis that starts in the middle of the first event and continues through the remaining two events.

ATWV scores the system as 1 TP and 2 FNs since it assigns the extended hypothesis event to the center reference event and leaves the other two undetected. The ATWV score is 0.33 for seizure events, 0.25 for background events, resulting in an average ATWV of 0.29. The sensitivity and FA rates for seizure events for this metric are 33% and 0 per 24 hrs. respectively.

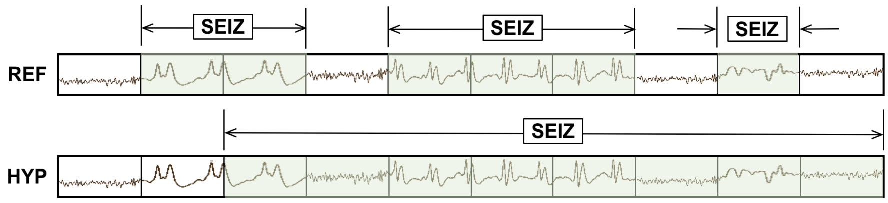

Figure 9. An example that summarizes the differences between scoring metrics

DPALIGN scores the system the same way since time alignments are ignored and the first event in each annotation are matched together, leaving the other two events undetected.

The EPOCH method scores the alignment 5 TP, 3 FP and 1 FN using a 1-sec epoch duration because there are 4 epochs for which the annotations do not agree and 5 epochs where they agree. The sensitivity is 83.33% and the FA rate per 24 hrs. is very high because of the 3 FPs.

The OVLP method scores the segment as 3 TP and 0 FP because detected events have partial to full overlap with all the reference events, giving a sensitivity of 100% with an FA rate of 0. TAES scores this segment as 0.5 TP and 2.5 FN because the first event is only 50% correct and there are FN errors for the 5th to 7th and 9th epochs (an example of multiple overlapping reference events), giving a sensitivity of 16.66% and a corresponding high FA rate.

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

It is difficult to conclude from this example which of these measures are most appropriate for EEG analysis. However, we see that ATWV and DPALIGN generally produce similar results. The EPOCH metric produces larger counts because it samples time rather than events. OVLP produces a high sensitivity while TAES produces a low sensitivity but a relatively higher FA rate. In the next section we conduct a more rigorous evaluation of these metrics using the output of several automatic seizure detection systems.

4 Evaluation

In order to evaluate the behavior of our scoring metrics, we analyzed the performance of several machine learning systems on a seizure detection task. We briefly introduce the TUH Seizure Detection Corpus. Next we introduce 5 different hybrid machine learning architectures based on deep learning principles. We then conduct a very detailed statistical analysis of the performance of these systems using the scoring metrics introduced in Section 3.

4.1 The TUH EEG Seizure Corpus

To demonstrate the differences between these metrics on a realistic task, we have evaluated a range of machine learning systems on a seizure detection task based on the TUH EEG Seizure (TUSZ) Corpus [38]. This is a subset of the TUH EEG Corpus developed at Temple University [39] that has been manually annotated. An overview of the corpus is given in Table 1. This is the largest open source corpus of its type. It consists of clinical data collected at Temple University Hospital. TUSZ represents a very challenging machine learning task because it contains a rich variety of common real-world problems (e.g. patient movements and artifacts) found in clinical data as well as various types of seizures (e.g., absence, tonic-clonic).

Table 1. The TUSZ 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

The version of the seizure database used for this study was v1.1.1 which contains 196 patients in the training set and 50 patients in the evaluation set, making it adequate to accurately assess fine differences in algorithm performance for machine learning algorithms. Although this database provides event-based as well as term-based annotations, for our study, we only include term–based annotations: a single decision is made at each point in time based on examination of all channels. Though annotations are channel-based (each channel is annotated independently), these annotations are aggregated to produce a single decision at each point in time. More information about the annotation process is available in Ochal et al. [40].

4.2 Machine Learning Architectures

For EEG signals, it is appropriate to use algorithms which can learn spatial as well as temporal context efficiently. Sequential algorithms such as hidden Markov models (HMMs), recurrent neural networks (RNNs) and convolutional neural networks (CNNs) are perfect candidates as the building blocks of the recognition system. We develop five different hybrid networks which use these algorithms in their system design. A general architecture for the five machine learning systems evaluated is shown in Figure 10.

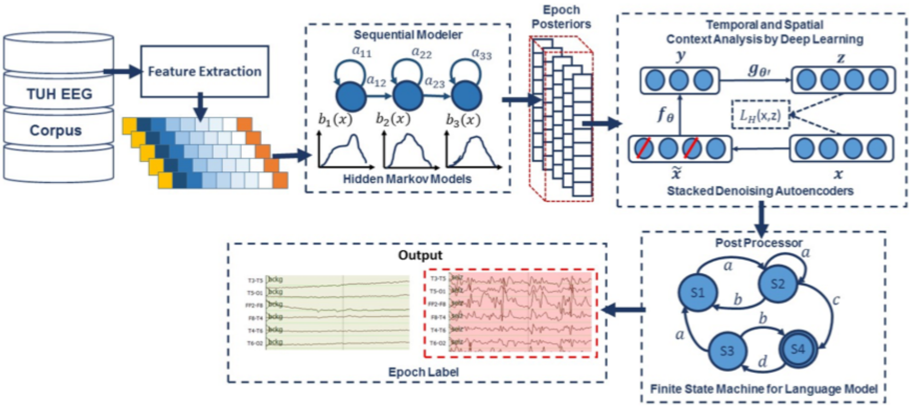

Figure 10. A hybrid deep learning architecture that integrates temporal and spatial context

The first step in this architecture is to convert an EEG signal, typically stored in a European Data Format (EDF) file [41], to a sequence of feature vectors. Linear Frequency Cepstral Coefficients features [42] are created using a 0.1 sec frame duration and a 0.2 second analysis window for each channel. We use the first 7 cepstral coefficients along with their first and second derivatives. We add several energy terms which bring the total feature vector dimension to 26. Attempts to circumvent the feature extraction process by using a deep learning-based approach have not produced significantly better results than these model-based features.

A group of frames are classified into an event on a per-channel basis using a combination of deep learning networks. The deep learning system essentially looks across multiple epochs, which we refer to as the temporal context, and multiple channels, which we refer to as the spatial context since each channel is associated with a location of an electrode on a patient's head. There are a wide variety of algorithms that can be used to produce a decision from these inputs. Even though seizures occur on a subset of the channels input to such a system, we focus on a single decision made across all channels at each point in time.

The five systems we included in this study were carefully selected because they represent a range of performance that is representative of state of the art on this task and because these systems exhibit different error modalities. The performance of these systems is sufficiently close so that the impact of these different scoring metrics becomes apparent. The systems selected are briefly described below.

(1) *HMM/SdA[43]:* a hybrid system consisting of a hidden Markov model (HMM) decoder and a postprocessor that uses a Stacked Denoising Autoencoder (SdA). An N-channel EEG was transformed into N independent feature streams using a standard sliding window-based approach. The hypotheses generated by the HMMs were postprocessed using a second stage of processing that examines the temporal and spatial context. We apply a third pass of postprocessing that uses a stochastic language model to smooth hypotheses involving sequences of events so that we can suppress spurious outputs. This third stage of postprocessing provides a moderate reduction in the false alarm rate.

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

The output of the epoch-based decisions was postprocessed by a deep learning system. The SdA network has three hidden layers with corruption levels of 0.3 for each layer. There are 800 nodes in the first layer, 500 nodes in the second layer, and 300 nodes in the third layer. The parameters for pre-training are: learning rate = 0.5, number of epochs = 150, batch size = 300. The parameters for fine-tuning are: learning rate = 0.1, number of epochs = 300, batch size = 100.

The overall result of the second stage is a probability vector of dimension two containing a likelihood that each label could have occurred in the epoch. A soft decision paradigm is used rather than a hard decision paradigm because this output is smoothed in the third stage of processing.

(2) *HMM/LSTM [43]:* an HMM decoder postprocessed by a Long Short-Term Memory (LSTM) network. Like the HMM/SdA hybrid approach previously described, the output of the HMM system is a vector of dimension: number of classes (2) × number of channels (22) × the window length (7) = 308. Therefore, we also use Principal Components Analysis (PCA) before LSTM in this approach to reduce the dimensionality of the data to 20. For this study, we used a window length of 41 for LSTM. This layer is composed of one hidden layer with 32 nodes. The output layer nodes in this LSTM level use a sigmoid activation function. The parameters of the models are optimized to minimize the error using a cross-entropy loss function. Adaptive Moment Estimation (Adam) is used in the optimization process.

(3) *IPCA/LSTM [43]:* a preprocessor based on Incremental Principal Component Analysis (IPCA) followed by an LSTM decoder. The EEG features are delivered to an IPCA layer for spatial context analysis and dimensionality reduction. A batch size of 50 is used in IPCA and the output dimension is 25. The output of IPCA is delivered to an LSTM for classification. We used a one-layer LSTM with a hidden layer size of 128. A batch size of 128 was used along with Adam optimization and a cross–entropy loss function.

(4) *CNN/MLP[45]:* a pure deep learning-based approach that uses a Convolutional Neural Network (CNN) decoder and a Multi-Layer Perceptron (MLP) postprocessor. The network contains six convolutional layers, three max pooling layers and two fully-connected layers. A rectified linear unit (ReLU) non-linearity is applied to the output of every convolutional and fully-connected layer.

(5) *CNN/LSTM[45]:* a pure deep learning-based architecture that uses a combination of CNN and LSTM networks. In this architecture, we integrate 2D CNNs, 1D CNNs and LSTM networks to better exploit long-term dependencies. Exponential Linear Units (ELU) are used as the activation functions for the hidden layers. Adam is used in the optimization process along with a mean squared error loss function.

The details of these systems are not critical to this study. We selected these systems because we needed a range of typical system performance that would expose the differences in the scoring metrics. What is more important is how the range of performance is reflected in these metrics.

A comparison of the performance is presented in Table 2. For each scoring metric, we provide the measured sensitivity, specificity and FA rate. For the ATWV metric, we also provide the ATWV score. Though the rankings of these systems vary as a function of the metric, the overall trends are accurately represented in Table 2. HMM/SdA generally performs the poorest of these systems, delivering a respectable sensitivity at

Table 2. Performance vs. scoring metric

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%
	FA/24 hr.	98.65	75.59	94.41	94.25	12.78
	ATWV	-0.8392	-0.8469	-0.4628	-0.7971	0.1737
DPALIGN	Sensitivity	44.11%	33.77%	35.77%	43.35%	32.46%
	Specificity	66.87%	72.99%	69.59%	71.49%	95.17%
	FA/24 hr.	86.15	66.98	81.17	77.67	10.19
EPOCH	Sensitivity	20.71%	50.46%	51.02%	65.03%	9.784%
	Specificity	98.22%	94.82%	94.09	91.55%	99.84%
	FA/24 hr.	1418.02	4133.34	4711.58	6738.82	125.79
OVLP	Sensitivity	35.35%	30.05%	32.97%	39.09%	30.83%
	Specificity	73.35%	80.53%	77.57%	76.84%	96.86%
	FA/24 hr.	77.39	60.92	73.52	77.19	6.75
TAES	Sensitivity	17.29%	22.84%	22.12%	31.58%	12.48%
	Specificity	66.04%	70.41%	66.64%	64.75%	95.24%
	FA/24 hr.	82.26	68.31	83.01	91.53	7.54

a high FA rate. CNN/LSTM typically delivers the highest overall performance because it has a low FA rate, which is very important in this type of application.

5 Derived Measures

Most supervised machine learning algorithms are designed to classify labels with some type of bounded or unbounded confidence measure such as a posterior probability or a log-likelihood. Possible exceptions are nonparametric techniques such K-nearest neighbors and decision trees. These confidence measures allow algorithm designers to sweep through threshold values for the confidence measures and observe performance at different operating points. In this section, we analyze the performance of these systems using DET curves and derived measures such as AUC and F scores.

5.1 Detection Error Trade-off Analysis

Evaluating systems from a single operating point is always a bit tenuous. It is very difficult to compare the performance of various systems when only two values are reported (e.g. sensitivity and specificity) because these systems might simply be designed to balance the four basic error categories differently (e.g., using a different threshold to reject FPs). For example, in seizure detection, the a priori probability of a seizure is very low, which means assessment of background events dominate the error calculations. The degree to which a system is capable of producing a seizure hypotheses will greatly impact its specificity. Further, sensitivity varies significantly when the FA rate is very low. Therefore, comparing systems that differ significantly in FA rate can be misleading. Often, we prefer a more holistic view of performance that is provided by a Receiver Operating Characteristic (ROC) curve or a Detection Error Trade-off (DET) curve. A ROC curve displays TP as a function of FP while a DET curve displays FN as a function of FP.

In Figure 11, we provide DET curves for the systems presented in Table 2. We refer to this analysis as a derived measure because these curves require calculations of the four error categories, which in turn requires the selection of a scoring metric. The DET curves in Figure 11 were derived from output generated using the OVLP scoring metric. The shapes of the DET curves do not change significantly with the scoring metric though the absolute numbers vary similarly to what we see in in Table 2.

From this data it is clear that CNN/LSTM performance is significantly different from the other systems. This is primarily because of its low FA rate. For this particular application, sensitivity drops rapidly as the FA rate is lowered. Therefore, comparing a single data point for each system is dangerous because the systems are most likely operating at different points on a DET curve if the sensitivities are significantly different. We find tuning these systems to have a comparable FA rate is important when comparing two systems only based on sensitivity.

The sensitivity for each metric is given in Table 2. For example, for HMM/SdA, we see the lowest sensitivities are produced by the TAES and EPOCH metrics, while the

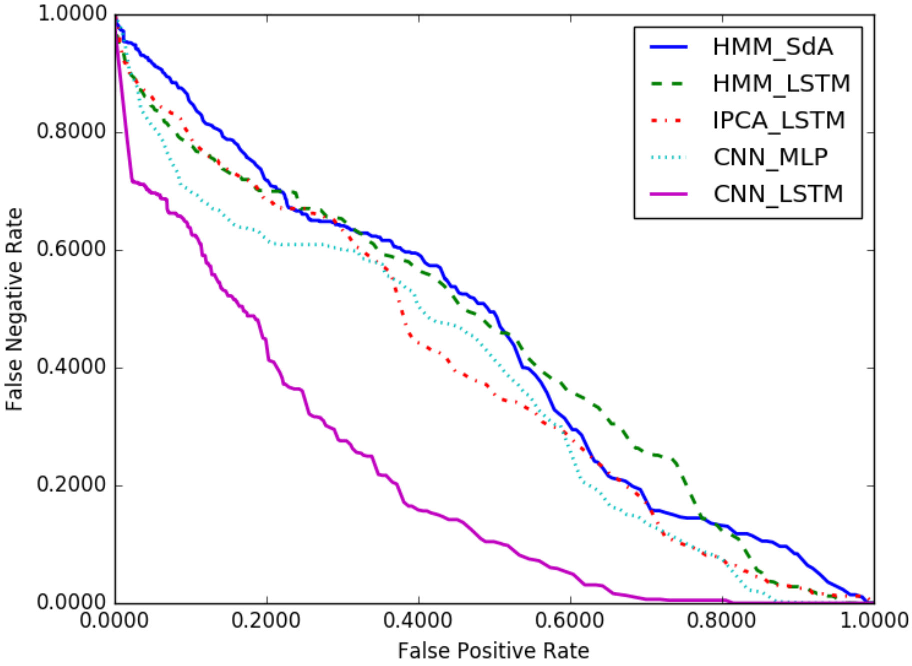

Figure 11. A comparison of DET curves

highest sensitivities are produced by OVLP and DPALIGN. This makes sense because OVLP and DPALIGN are very forgiving of time alignment errors, while TAES and EPOCH penalize time alignment errors heavily. We see similar trends for CNN/LSTM though the range of differences between the three highest scoring metrics is smaller. We also see that the five algorithms are ranked similarly by each scoring metric even though the scale of the numbers varies by metric. HMM/SdA consistently scores the lowest and CNN/LSTM consistently scores the highest. The other three systems are very similar in their performance.

The ATWV scores for all algorithms are extremely low. The ATWV scores are below 0.5 which indicates that overall performance is poor. However, the ATWV score for CNN/LSTM is significantly higher than the other four systems. ATWV attempts to reduce the information contained in a DET curve to a single number, and does a good job reflecting the results shown in Figure 11. The DET curves for HMM/LSTM and HMM/SdA overlap considerably for an FP rate between 0.25 and 1.0, and this is a primary reason why their ATWV scores are similar. However, for seizure detection we are primarily interested in the low FP rate region, and in that range, HMM/LSTM and IPCA/LSTM perform similarly.

When a single metric is preferred, the area under a DET or ROC curve (AUC) is also an effective way of comparing the performance. A random guessing approach to classification, assuming equal priors for each class, will give an AUC of 0.5 while a perfect classifier will give an AUC of 1.0. In Table 3 we provide AUCs for these DET curves calculated using OVLP and TAES for comparison. AUC values in Table 3 also follow a similar trend but the differences are less pronounced than in Figure 11 or in Table 2.

Table 3. AUC comparison

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

Note that the AUC value for the presumptive best system, CNN/LSTM, is significantly lower than the other four systems. If we examined the AUC in the FPR range of [0.0, 0.2], which corresponds to a low FA rate, and is the region of greatest interest, CNN/LSTM is still significantly better than the other algorithms, but the margin of difference shrinks slightly. The difference in the FPR range of [0.2, 0.8] is more pronounced. This is something we often see when evaluating new machine learning algorithms. They tend to deliver their best performance in the upper ranges of FPR but are not as impressive when the FPR rate is very low. This suggests the major issues an algorithm needs to address in the low FPR region are more related to auxiliary issues such as segmentation and noise rejection rather than optimal modeling of a complex decision surface. It is not uncommon that in machine learning applications involving real-world applications, such as clinical data, low-level issues such as segmentation of the data and robustness to spurious noises ultimately limit performance.

5.2 Accuracy and Other Derived Scores

A commonly used metric in the machine learning community that is somewhat intuitive is accuracy. The accuracies of the five systems are shown in Table 4. Accuracy places an equal weight on each type of error (though it is possible to apply heuristic weights in practice). This is acceptable if the dataset is balanced. However, for many bioengineering applications, such as seizure detection, the target class, or class of interest, occurs infrequently. According to the accuracies presented in Table 4, we see that CNN/LSTM is still significantly more accurate than the other four systems and the differences between the remaining four systems are minimal.

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. F1 scores for the five systems are shown in Table 5. We see there are significant variations between the systems and the results don't completely correlate with Table 4. For example, for the TAES and EPOCH metrics, which stress time alignments, the best performing system is not CNN/LSTM. F1 scores in our experience do not adequately emphasize the FA rate for applications such as seizure detection where the classes are imbalanced.

The Matthews Correlation Coefficient (MCC) [28] is an effective solution when a significant class imbalance exists. MCC is a contingency matrix method of calculating the Pearson product-moment correlation coefficient [46] between actual and predicted values. Recall (sensitivity) is the fraction of relevant samples that are correctly retrieved. Its dual metric, precision, is the fraction of retrieved samples that are relevant. Meaningfully combining precision and recall generates alternative performance evaluation measures such as the F1 ratio, which combines these scores using a geometric mean. MCC takes into account all four values in the confusion matrix. A value close to 1 means that both classes are predicted well, even if one class is disproportionately represented. Since MCC is a correlation coefficient, it ranges from $[-1, 1]$. Perfect misclassification corresponds to a value of -1, perfect classification corresponds to a value of 1, and random guessing with equal priors corresponds to a value of 0. Since no class is more important than the other, MCC is symmetric.

In Table 6, we present MCC results for the five systems and the five metrics. It is interesting to note that for the overall best system CNN/LSTM, MCC produces higher correlations for the first three metrics (ATWV, DPALIGN and OVLP). These metrics are based less on time alignments of the hypotheses. The latter two metrics (EPOCH and TAES) weigh the time alignments more heavily and generally produce lower scores because their matching criteria are more stringent.

IRA is an extremely useful measure for the development of reference annotations. It is not uncommon that a team of annotators will be involved in the annotation of a large corpus. Individual annotators are evaluated and compared using IRA [47]. Though there are numerous ways to measure IRA, Cohen's Kappa statistic, as shown in Eq. (14), is one of the most popular ways to compute IRA. In Table 7, we show IRA values for the five systems. Again, we observe that CNN/LSTM has higher IRA values

Table 4. Accuracy vs. metric

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

Table 5. F1 vs. metric

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

Table 6. MCC vs. metric

Metric	HMM/ SdA	HMM/ LSTM	IPCA/ LSTM	CNN/ MLP	CNN/ LSTM
ATWV	-0.07	-0.04	-0.11	-0.05	0.30
DPALIGN	0.01	0.07	0.05	0.15	0.35
EPOCH	0.28	0.43	0.41	0.45	0.23
OVLP	0.08	0.11	0.11	0.16	0.41
TAES	-0.16	-0.07	-0.12	-0.04	0.13

Table 7. Cohen’s Kappa (κ) vs. metric

Metric	HMM/ SdA	HMM/ LSTM	IPCA/ LSTM	CNN/ MLP	CNN/ LSTM
ATWV	-0.07	-0.04	-0.11	-0.05	0.26
DPALIGN	0.09	0.07	0.05	0.15	0.31
EPOCH	0.26	0.43	0.41	0.43	0.12
OVLP	0.08	0.11	0.11	0.16	0.35
TAES	-0.16	-0.07	-0.11	-0.04	0.09

than the other systems, except for the EPOCH metric. Both MCC and IRA report similar trends for CNN/LSTM versus the other four systems for the EPOCH metric.

5.3 Additional Insight

We generally prefer operating points where performance in terms of sensitivity, specificity and FAs is balanced. The ATWV metric explicitly attempts to encourage balancing of 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.

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. 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.

6 Statistical Analysis

To understand the pairwise statistical difference between these evaluation metrics and deep architectures, we have performed three tests: Kolmogorov-Smirnov (KS), Pearson's R (correlation coefficient) and z-test [48]. 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.

6.1 Kolmogorov-Smirnov and Pearson's R Tests

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 range from 0.61 – 0.71 for sensitivity and 0.99 – 1.00 for specificity with the *p*-values equal to zero. We then evaluated the correlation coefficient (Pearson's R) between pairs of metrics.

Correlations for each pair of scoring metrics are shown in Table 8 (for sensitivity) and Table 9 (for specificity). It can be seen that the pairwise correlations between OVLP, ATWV and DPALIGN are highest, while the pairs ATWV-EPOCH and DPALIGN-EPOCH have the lowest correlation (~0.5). 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

Table 8. Correlation of the scoring metrics based on sensitivity (p < 0.001)

Metric	ATWV	DPALIGN	EPOCH	OVLP	TAES
ATWV	---	0.87	0.50	0.92	0.71
DPALIGN		---	0.48	0.90	0.69
EPOCH			---	0.62	0.87
OVLP				---	0.78
TAES					---

Table 9. Correlation of the scoring metrics based on specificity (p < 0.001)

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

6.2 *Z*-tests

To understand the statistical significance of each system, we perform two-tailed z-tests for sensitivity as shown in Table 10 and for specificity as shown in Table 11. Cells in these tables contain entries that consist of the sensitivity/specificity differences between the systems and a binary classification value (Yes/No) based on extracted p-values from the z-test with 95% confidence. (Due to space constraints, the five classification systems are represented using the abbreviations M1 to M5.) The data was prepared by scoring systems on individual patients. Prior to performing z-tests, the Gaussianity of each sample was evaluated using a KS test. All the samples were confirmed as normal with $p < 0.001$.

From Table 10, it can be observed that, aside from the EPOCH and TAES scoring metrics, the differences between the CNN-LSTM system and all the other systems are statistically significant (rejecting the null hypothesis with $p < 0.05$. On the other hand, the EPOCH and TAES metrics fail to reject the null hypothesis for CNN-LSTM. According to these metrics, the performance of HMM-SDA is statistically different

Table 10. Significance calculated using z-tests for α = 0.05 (for sensitivity)

ATWV (Abs. Sensitivity Difference (%), Significant/Non-significant)					
ML Systems (Sens.)	**CNN-LSTM (M1)**	**CNN-MLP (M2)**	**HMM-LSTM (M3)**	**HMM-SDA (M4)**	**IPCA-LSTM (M5)**
M1 (30.34%)	---	(00.82%) Y	(03.61%) Y	(00.01%) Y	(05.61%) Y
M2 (29.52%)		---	(02.79%) N	(00.83%) N	(04.79%) N
M3 (26.73%)			---	(03.62%) N	(02.00%) N
M4 (30.35%)				---	(05.62%) N
M5 (24.73%)					---
DPALIGN (Abs. Sensitivity Difference, Significant/Non-significant)					
ML Systems (Sens.)	**CNN-LSTM (M1)**	**CNN-MLP (M2)**	**HMM-LSTM (M3)**	**HMM-SDA (M4)**	**IPCA-LSTM (M5)**
M1 (32.46%)	---	(10.89%) Y	(01.31%) Y	(11.65%) Y	(03.31%) Y
M2 (43.35%)		---	(09.58%) N	(00.76%) N	(07.58%) N
M3 (33.77%)			---	(10.34%) N	(02.00%) N
M4 (44.11%)				---	(08.34%) N
M5 (35.77%)					---
EPOCH (Abs. Sensitivity Difference, Significant/Non-significant)					
ML Systems (Sens.)	**CNN-LSTM (M1)**	**CNN-MLP (M2)**	**HMM-LSTM (M3)**	**HMM-SDA (M4)**	**IPCA-LSTM (M5)**
M1 (09.78%)	---	(55.25%) N	(40.68%) N	(10.93%) Y	(41.24%) N
M2 (65.03%)		---	(14.57%) Y	(44.32%) Y	(14.01%) N
M3 (50.46%)			---	(29.75%) Y	(00.56%) N
M4 (20.71%)				---	(30.31%) Y
M5 (51.02%)					---
OVLP (Abs. Sensitivity Difference, Significant/Non-significant)					
ML Systems (Sens.)	**CNN-LSTM (M1)**	**CNN-MLP (M2)**	**HMM-LSTM (M3)**	**HMM-SDA (M4)**	**IPCA-LSTM (M5)**
M1 (30.83%)	---	(08.26%) Y	(02.14%) Y	(04.52%) Y	(02.14%) Y
M2 (39.09%)		---	(09.04%) N	(03.74%) N	(06.12%) N
M3 (30.05%)			---	(05.30%) N	(02.92%) N
M4 (35.35%)				---	(02.38%) N
M5 (32.97%)					---
TAES (Abs. Sensitivity Difference, Significant/Non-significant)					
ML Systems (Sens.)	**CNN-LSTM (M1)**	**CNN-MLP (M2)**	**HMM-LSTM (M3)**	**HMM-SDA (M4)**	**IPCA-LSTM (M5)**
M1 (12.48%)	---	(19.10%) N	(10.36%) N	(04.81%) Y	(09.64%) N
M2 (31.58%)		---	(08.74%) N	(14.29%) Y	(09.46%) N
M3 (22.84%)			---	(05.55%) Y	(00.72%) N
M4 (17.29%)				---	(04.83%) Y
M5 (22.12%)					---

from the other systems, confirming its poor performance. This can also be observed from EPOCH/TAES results shown in Table 2.

Table 11 shows a different trend than Table 10. The EPOCH metric fails to reject null hypothesis for all the systems. Since specificity is calculated from TN and FP

Table 11. Significance calculated using *z*-tests for α = 0.05 (for specificity)

ATWV (Abs. Specificity Difference (%), Significant/Non-significant)					
ML Systems (Spec.)	**CNN-LSTM (M1)**	**CNN-MLP (M2)**	**HMM-LSTM (M3)**	**HMM-SDA (M4)**	**IPCA-LSTM (M5)**
M1 (93.15%)	---	(27.28%) Y	(24.22%) Y	(31.77%) Y	(28.64%) Y
M2 (65.87%)		---	(03.06%) N	(04.49%) N	(01.36%) N
M3 (68.93%)			---	(07.55%) Y	(04.42%) N
M4 (61.38%)				---	(03.13%) N
M5 (64.51%)					---
DPALIGN (Abs. Specificity Difference (%), Significant/Non-significant)					
ML Systems (Spec.)	**CNN-LSTM (M1)**	**CNN-MLP (M2)**	**HMM-LSTM (M3)**	**HMM-SDA (M4)**	**IPCA-LSTM (M5)**
M1 (95.17%)	---	(23.68%) Y	(22.18%) Y	(28.30%) Y	(25.58%) Y
M2 (71.49%)		---	(01.50%) N	(04.62%) Y	(01.90%) N
M3 (72.99%)			---	(06.12%) Y	(03.40%) N
M4 (66.87%)				---	(02.72%) Y
M5 (69.59%)					---
EPOCH (Abs. Specificity Difference (%), Significant/Non-significant)					
ML Systems (Spec.)	**CNN-LSTM (M1)**	**CNN-MLP (M2)**	**HMM-LSTM (M3)**	**HMM-SDA (M4)**	**IPCA-LSTM (M5)**
M1 (99.84%)	---	(08.29%) N	(05.02%) N	(01.62%) N	(05.75%) N
M2 (91.55%)		---	(03.27%) N	(06.67%) N	(02.54%) N
M3 (94.82%)			---	(03.40%) N	(00.73%) N
M4 (98.22%)				---	(04.13%) N
M5 (94.09%)					---
OVLP (Abs. Specificity Difference (%), Significant/Non-significant)					
ML Systems (Spec.)	**CNN-LSTM (M1)**	**CNN-MLP (M2)**	**HMM-LSTM (M3)**	**HMM-SDA (M4)**	**IPCA-LSTM (M5)**
M1 (96.86%)	---	(20.02%) Y	(16.33%) Y	(23.51%) Y	(19.29%) Y
M2 (76.84%)		---	(03.69%) N	(03.49%) Y	(00.73%) N
M3 (80.53%)			---	(07.18%) Y	(02.96%) N
M4 (73.35%)				---	(04.22%) Y
M5 (77.57%)					---
TAES (Abs. Specificity Difference (%), Significant/Non-significant)					
ML Systems (Spec.)	**CNN-LSTM (M1)**	**CNN-MLP (M2)**	**HMM-LSTM (M3)**	**HMM-SDA (M4)**	**IPCA-LSTM (M5)**
M1 (95.24%)	---	(31.21%) Y	(24.83%) Y	(29.20%) Y	(28.60%) Y
M2 (64.03%)		---	(06.38%) N	(02.01%) Y	(02.61%) N
M3 (70.41%)			---	(04.37%) Y	(03.77%) N
M4 (66.04%)				---	(00.60%) Y
M5 (66.64%)					---

values, for an evaluation set 167 hours in duration and an epoch size 0.25, a few thousand seconds of FPs do not make any significant difference in terms of specificity. This can also be directly observed in Table 2, where the specificity of all systems according to the EPOCH metric is always greater than 90%. The huge difference between the duration of background and seizure events is the primary reason for such high specificities. However, the OVLP and TAES metrics completely agree with each other's *z*-test results for specificity.

7 Conclusions

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

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

Finally, the Python implementation of these metrics is available at the project web site: *https://www.isip.piconepress.com/projects/tuh_eeg/downloads/nedc_eval_eeg*. This scoring software described here has been publicly available since late 2018. It has been used for two open source evaluations [49][50]. Readers are encouraged to refer to the software for detailed questions about the specific implementations of these algorithms and the tunable parameters available.

Acknowledgements

Research reported in this publication was most recently supported by the National Human Genome Research Institute of the National Institutes of Health under award number U01HG008468. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This material is also based in part upon work supported by the National Science Foundation under Grant No. IIP-1622765. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We are also grateful to our colleagues at IBM Research: Stefan Harrer, Jiri Navratil and Subhrajit Roy. Their careful reading of early versions of this manuscript and valuable insights about the work were greatly appreciated. We also would like to acknowledge Seyedeh Saeedeh Khoshgoftar Ziyabari for her contributions to the initial portions of this work.

Conflict of Interest Statement

Author Meysam Golmohammadi is employed by Internet Brands, El Segundo, California, USA. This work was completed at the Neural Engineering Data Consortium at Temple University prior to his employment at Internet Brands. All other authors declare no conflict of interest.

References

[1] T. Yamada and E. Meng, *Practical guide for clinical neurophysiologic testing: EEG*. Philadelphia, Pennsylvania, USA: Lippincott Williams & Wilkins, 2017. *https://doi.org/10.1111/j.1468-1331.2009.02936.x*.

[2] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *J. Neural Eng.*, vol. 16, no. 5, p. 37, 2019. *https://doi.org/ 10.1088/1741-2552/ab260c*.

[3] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: a review," *J. Neural Eng.*, vol. 16, no. 3, p. 031001, 2019. *https://doi.org/10.1088/1741-2552/ab0ab5*.

[4] S. B. Wilson and R. Emerson, "Spike detection: a review and comparison of algorithms," *Clin. Neurophysiol.*, vol. 113, no. 12, pp. 1873–1881, Dec. 2002. *https://doi.org/10.1016/S1388-2457(02)00297-3*.

[5] J. Gotman, D. Flanagan, J. Zhang, and B. Rosenblatt, "Automatic seizure detection in the newborn: Methods and initial evaluation," *Electroencephalogr. Clin. Neurophysiol.*, vol. 103, no. 3, pp. 356–362, 1997. *https://doi.org/10.1016/S0013-4694(97)00003-9*.

[6] J. Gotman, "Automatic recognition of epileptic seizures in the EEG," *Electroencephalogr. Clin. Neurophysiol.*, vol. 54, no. 5, pp. 530–540, Nov. 1982. *http://www.sciencedirect.com/science/article/pii/0013469482900384*.

[7] G. D. Clifford et al., "False alarm reduction in critical care," P*hysiol. Meas.*, vol. 37, no. 8, pp. E5–E23, 2016. *https://doi.org/10.1088/0967-3334/37/8/E5*.

[8] M. Cvach Maria, "Managing hospital alarms," *Nurs. Crit. Care*, vol. 9, no. 3, pp. 13–27, 2014. *https://doi.org/10.1097/01.CCN.0000446255.81392.b0*.

[9] A. C. Bridi, T. Q. Louro, and R. C. L. Da Silva, "Clinical Alarms in intensive care: implications of alarm fatigue for the safety of patients," *Rev. Lat. Am. Enfermagem*, vol. 22, no. 6, p. 1034, 2014. *https://doi.org/10.1590/0104-1169.3488.2513*.

[10] P. von Goethem and B. Hambling, *User Acceptance Testing: A step-by-step guide*. Swindon, United Kingdom: BCS Learning & Development Limited, 2013. *https://www.oreilly.com/library/view/user-acceptance-testing/9781780171678/*.

[11] R. Banchs, A. Bonafonte, and J. Perez, "Acceptance Testing of a Spoken Language Translation System," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2006, p. 106. *http://www.lrec-conf.org/proceedings/lrec2006/pdf/60_pdf.pdf.*

[12] J. Picone, G. Doddington, and D. Pallett, "Phone-mediated word alignment for speech recognition evaluation," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 38, no. 3, pp. 559–562, Mar. 1990. *https://doi.org/10.1109/29.106877*.

[13] M. Michel, D. Joy, J. G. Fiscus, V. Manohar, J. Ajot, and B. Barr, "Framework for Detection Evaluation (F4DE)," National Institute of Standards and Technology, 2017. [Online]. [Accessed: 16-May-2017]. *https://github.com/usnistgov/F4DE*.

[14] D. G. Altman and J. M. Bland, "Diagnostic Tests 1: Sensitivity And Specificity," *Br. Med. J.*, vol. 308, no. 6943, p. 1552, 1994. *https://doi.org/10.1136/bmj.308.6943.1552*.

[15] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*. New York City, New York, USA: Wiley, 1965. *https://books.google.com/books/about/Principles_of_communication_engineering.html?id=4ORSAAAAMAAJ.*

[16] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, 1997, pp. 1895–1898. *http://www.isca-speech.org/archive/eurospeech_1997/e97_1895.html*.

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

[18] D. Mostefa, O. Hamin, and K. Choukri, “Evaluation of Automatic Speech Recognition and Speech Language Translation within TC-STAR: Results from the first evaluation campaign,” in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2006, pp. 149-154. *http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.323.5822&rank=4*.

[19] K. M. Kelly et al., “Assessment of a scalp EEG-based automated seizure detection system,” *Clin. Neurophysiol.*, vol. 121, no. 11, pp. 1832–1843, 2010. *https://doi.org/10.1016/j.clinph.2010.04.016*.

[20] S. Baldassano et al., “A novel seizure detection algorithm informed by hidden Markov model event states,” *J. Neural Eng.*, vol. 13, no. 3, p. 036011, 2016. *https://doi.org/10.1016/j.clinph.2010.04.016*.

[21] M. Winterhalder, T. Maiwald, H. U. Voss, R. Aschenbrenner-Scheibe, J. Timmer, and A. Schulze-Bonhage, “The seizure prediction characteristic: a general framework to assess and compare seizure prediction methods,” *Epilepsy Behav.*, vol. 4, no. 3, pp. 318–325, 2003. *https://doi.org/10.1016/S1525-5050(03)00105-7*.

[22] S. Wegmann, A. Faria, A. Janin, K. Riedhammer, and N. Morgan, “The TAO of ATWV: Probing the mysteries of keyword search performance,” in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 192–197.

[23] J. Fiscus, J. Ajot, J. Garofolo, and G. Doddingtion, “Results of the 2006 Spoken Term Detection Evaluation,” in *Proceedings of the ACM Special Interest Gruoup on Information Retrieval (SIGIR) Workshop "Searching Spontaneous Conversational Speech"*, 2007, pp. 45–50. *https://www.nist.gov/publications/results-2006-spoken-term-detection-evaluation*.

[24] S. J. Mason and N. E. Graham, “Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation,” *Q. J. R. Meteorol. Soc.*, vol. 128, no. 584, pp. 2145–2166, Jul. 2002. *https://doi.org/10.1256/003590002320603584*.

[25] K. Hajian-Tilaki, “Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation,” *Casp. J. Intern. Med.*, vol. 4, no. 2, pp. 627–635, Dec. 2013. *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/*.

[26] N. Japkowicz and M. Shah, “Evaluating Learning Algorithms: a classification perspective.” Cambridge University Press, New York City, New York, USA, p. 424, 2014. *https://doi.org/10.1017/CBO9780511921803*.

[27] “Confusion matrix.” [Online]. Available: [Accessed: 31-Oct-2017]. *https://en.wikipedia.org/wiki/Confusion_matrix*.

[28] D. Chicco and G. Jurman, “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation,” *BMC Genomics*, vol. 21, no. 1, p. 6, Jan. 2020. *https://doi.org/10.1186/s12864-019-6413-7*.

[29] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A Method for Automatic Evaluation of Machine Translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318. *https://doi.org/10.3115/1073083.1073135*.

[30] A. Liu, J. S. Hahn, G. P. Heldt, and R. W. Coen, “Detection of neonatal seizures through computerized EEG analysis,” *Electroencephalogr. Clin. Neurophysiol.*, vol. 82, no. 2, pp. 32–37, 1992. *https://doi.org/ 10.1016/0013-4694(92)90179-L*.

[31] M. A. Navakatikyan, P. B. Colditz, C. J. Burke, T. E. Inder, J. Richmond, and C. E. Williams, “Seizure detection algorithm for neonates based on wave-sequence analysis,” *Clin. Neurophysiol.*, vol. 117, no. 6, pp. 1190–1203, Jun. 2006. *https://doi.org/10.1016/j.clinph.2006.02.016*.

[32] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, “The Microsoft 2017 Conversational Speech Recognition System,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017, pp. 5255–5259. *https://doi.org/ 10.1109/ICASSP.2018.8461870*.

[33] S. B. Wilson, M. L. Scheuer, C. Plummer, B. Young, and S. Pacia, “Seizure detection: Correlation of human experts,” *Clin. Neurophysiol.*, vol. 114(11), pp. 2156–2164, 2003. *https://doi.org/10.1016/S1388-2457(03)00212-8*.

[34] J. G. Fiscus, “Overview of the NIST Open Keyword Search 2013 Evaluation Workshop,” in *IEEE Signal Processing Society - SLTC Newsletter*, 2013. *https://www.nist.gov/publications/overview-nist-open-keyword-search-2013-evaluation-worksho*.

[35] J. G. Fiscus, “Speech Recognition Scoring Toolkit,” National Instutue of Standards and Technology, 2017. [Online]. [Accessed: 17-Oct-2017]. *https://github.com/usnistgov/SCTK*.

[36] M. L. McHugh, “Interrater reliability: the kappa statistic,” *Biochem. Medica*, vol. 22, no. 3, pp. 276–282, Oct. 2012. *https://doi.org/10.11613/BM.2012.031*.

[37] J. J. Halford et al., "Inter-Rater Agreement on Identification of Electrographic Seizures and Periodic Discharges in ICU EEG Recordings.," *Clin. Neurophysiol.*, vol. 126, no. 9, pp. 1661–1669, 2015. *https://doi.org/10.1016/j.clinph.2014.11.008*.

[38] V. Shah et al., "The Temple University Hospital Seizure Detection Corpus," *Front. Neuroinform.*, vol. 12, pp. 1–6, 2018. *https://doi.org/10.3389/fninf.2018.00083*.

[39] I. Obeid and J. Picone, "The Temple University Hospital EEG Data Corpus," in *Augmentation of Brain Function: Facts, Fiction and Controversy*. Volume I: Brain-Machine Interfaces, 1st ed., vol. 10, M. A. Lebedev, Ed. Lausanne, Switzerland: Frontiers Media S.A., 2016, pp. 394–398. *https://doi.org/10.3389/fnins.2016.00196*.

[40] D. Ochal, S. Rahman, S. Ferrell, T. Elseify, I. Obeid, and J. Picone, "The Temple University Hospital EEG Corpus: Annotation Guidelines," Philadelphia, Pennsylvania, USA, 2020. *https://www.isip.piconepress.com/publications/reports/2020/tuh_eeg/annotations/*.

[41] R. Kemp, "European Data Format," Department of Neurology, Leiden University Medical Centre, The Netherlands, 2013. [Online]. [Accessed: 06-Jan-2013]. *http://www.edfplus.info*.

[42] A. Harati, M. Golmohammadi, S. Lopez, I. Obeid, and J. Picone, "Improved EEG Event Classification Using Differential Energy," in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, 2015, pp. 1–4. *https://doi.org/10.1109/SPMB.2015.7405421*.

[43] M. Golmohammadi, A. Harati, S. de Diego, I. Obeid, and J. Picone, "Automatic Analysis of EEGs Using Big Data and Hybrid Deep Learning Architectures," *Front. Hum. Neurosci.*, vol. 13, p. 76, 2019. *https://doi.org/ 10.3389/fnhum.2019.00076*.

[44] M. Golmohammadi, S. Ziyabari, V. Shah, I. Obeid, and J. Picone, "Deep Architectures for Spatio-Temporal Modeling: Automated Seizure Detection in Scalp EEGs," in *Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018, pp. 745–750. *https://doi.org/10.1109/ICMLA.2018.00118*.

[45] M. Golmohammadi, V. Shah, I. Obeid, and J. Picone, "Deep Learning Approaches for Automatic Seizure Detection from Scalp Electroencephalograms," in *Signal Processing in Medicine and Biology: Emerging Trends in Research and Applications*, 1st ed., I. Obeid, I. Selesnick, and J. Picone, Eds. New York, New York, USA: Springer, 2020, pp. 233–274. *https://doi.org/10.1007/978-3-030-36844-9*.

[46] D. M. W. Powers, “Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation,” *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011. *https://bioinfopublication.org/files/articles/2_1_1_JMLT.pdf.*

[47] V. Shah, E. von Weltin, T. Ahsan, I. Obeid, and J. Picone, “On the Use of Non-Experts for Generation of High-Quality Annotations of Seizure Events,” *J. Clin. Neurophysiol*. (under review), 2020. *https://www.isip.piconepress.com/publications/unpublished/journals/2019/elsevier_cn/ira/.*

[48] F. Hammond, J. Malec, R. Buschbacher, and T. Nick, *Handbook for Clinical Research : Design, Statistics, and Implementation*. New York City, New York, USA: Demos Medical Publishing, 2015. *https://www.springerpub.com/handbook-for-clinical-research-9781936287543.html.*

[49] I. Kiral et al., “The Deep Learning Epilepsy Detection Challenge: Design, Implementation, and Test of a New Crowd-Sourced AI Challenge Ecosystem,” presented at the Neural Information Processing Systems (NeurIPS) Workshop on Challenges in Machine Learning Competitions for All (CiML), 2019. *https://isip.piconepress.com/publications/conference_presentations/2019/neurips_ciml/epilepsy_challenge/*

[50] Y. Roy, R. Iskander, and J. Picone, “The Neureka(^{TM}) 2020 Epilepsy Challenge,” NeuroTechX, 2020. [Online]. [Accessed: 16-Apr-2020]. *https://neureka-challenge.com/.*