**O****bjective evaluation metrics for automatic classification of EEG events**

**Saeedeh Ziyabari1, Vinit Shah1, Meysam Golmohammadi2,  
Iyad Obeid1 and Joseph Picone1**

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

2 BioSignal Analytics, Inc., 3624 Market Street, Suite 5E, Philadelphia, Pennsylvania, 19104, USA.

**Abstract:** The evaluation of machine learning algorithms in biomedical fields for applications involving sequential data lacks standardization. Common quantitative scalar evaluation metrics such as sensitivity and specificity can often be misleading depending on the requirements of the application. Evaluation metrics must ultimately reflect the needs of users yet be sufficiently sensitive to guide algorithm development. 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 a need for a single scalar figure of merit. In this paper, 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 Corpus. We show that two promising metrics are a measure based on a concept borrowed from the spoken term detection literature, Actual Term-Weighted Value, and 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:** electroencephalograms, EEG, machine learning, evaluation metrics

# 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 has been extensively studied in the past decade [2]-[6]. However, 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][8][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 standardized scoring methodologies.

Corresponding Author: Saeedeh Ziyabari, The Neural Engineering Data Consortium, ENGR 702, Temple University, 1947 North 12th Street, Philadelphia, Pennsylvania, 19122, USA, Tel: 347-610-1465, Fax: 215-204-5960, Email: saeedeh@temple.edu.

There are in general two types of 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 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][15][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.

Sequential pattern recognition applications, such as speech recognition, keyword search or EEG analysis, require additional considerations. Data, typically organized in files on a computer, 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. [19] advocates using a term-based metric involving of sensitivity and specificity. Each term is created by connecting consecutive decisions of the same class. 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. [20] 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. [21] 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 (or after the event has occurred) we can predict an outcome. Accuracy of prediction varies with latency, so this type of performance evaluation adds some complexity to the process. Winterhalder et al. [22] have studied this problem extensively and argue for a scoring based on long-term considerations. In this paper, we are not concerned with these types of prediction problems. We are focused mainly on assessing the accuracy of classification and assessing the proximity of these classifications to the actual event.

Therefore, in this paper, 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 [23][24] 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 overall user acceptance.

# Method

Researchers in biomedical fields typically report performance in terms of sensitivity and specificity [25]. 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’

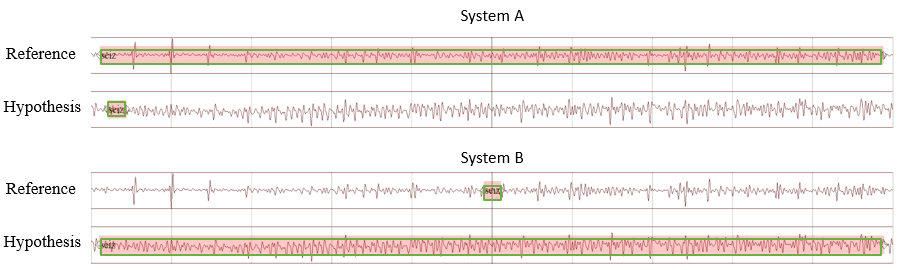
Sensitivity (TP/(TP+FN)) and specificity (TN/(TN+FP)) are derived from these quantities. There are a large number of auxiliary measures that can be calculated from these four basic quantities and are used extensively in the literature. These are summarized concisely in [26]. For example, in information retrieval problems, systems are often evaluated using precision (TP/(TP+FP)), recall (another term for sensitivity) and F1 score (2•TP/(2•TP+FP+FN)). However, none of these measures address the time scale on which the scoring must occur, which is critical in the interpretation of these measures.

In some applications, it is preferable to score every unit of time. With multichannel signals, such as EEGs, scoring for each channel each unit of time might be appropriate. However, it is more common in the literature to simply score a summary decision per unit of time that is based on the per-channel inputs (e.g., a majority vote). We refer to this type of scoring as epoch-based [27][28]. An alternative, that is more common in speech and image recognition applications, is term-based [24][29], 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 duration of events more heavily.

Time-aligned scoring is essential to sequential decoding problems. 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, 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 errors. Term-based scoring would potentially count this as a correct recognition depending on the way overlaps are scored.

**Figure 1.** A typical situation where a hypothesis has a 50% overlap with the reference.

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 [30][31]. TPs are counted when the hypothesis overlaps with reference annotation. FPs correspond to situations in which the 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 Any-Overlap metric fails. In each case, *90%* of the event is incorrectly scored. In the first case, system A does not detect approximately *9* seconds of a seizure event, while in the second case, system B detects incorrectly an additional *9* seconds of time in which the event hypothetically occurred. Any-Overlap is considered a very permissive way of scoring, resulting in artificially high sensitivities.



**Figure 2.** Detection rates for the Any-Overlap metric are 100% even though large portions of the event are missed.

It is very difficult to compare the performance of various systems when only two values are reported (e.g. sensitivity and specificity) and when the prior probabilities vary significantly (in seizure detection, the a priori probability of a seizure is very low). Therefore, often a more holistic view is preferred. In this case, a Receiver Operating Characteristic (ROC) [15] or a Detection Error Trade-off (DET) curve [16] are preferred. An ROC curve displays the TP rate as a function of the FP rate while a DET curve displays the FN rate as a function of the TP rate. When a single metric is preferred, the area under an ROC curve (AUC) [32][33] is also an effective way of comparing the performance. A random guessing approach to classification will give an AUC of *0.5* while a perfect classifier will give and AUC of *1.0*.

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 [23][24][34]. These evaluations resulted in the introduction of a single metric, Actual Term-Weighted Value (ATWV) [24], that attempted to address concerns about tradeoffs for the different types of errors that occur. Despite being popular in the voice processing community, ATWV has not been extensively used in the bioengineering community.

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

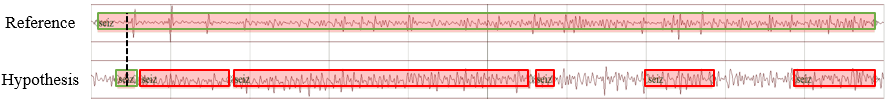
1. *NIST Actual Term-Weighted Value (ATWV):* based on NIST’s popular scoring package (F4DE v3.3.1), this metric, originally developed for the NIST 2006 Spoken Term Detection evaluation, uses an objective function that accounts for both temporal overlap between the reference and hypothesis using the detection scores assigned by the system.
2. *Dynamic Programming Alignment* (DPALIGN): similar to the NIST package known as SCLite [35], this algorithm uses a dynamic programming algorithm to time-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): assess 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.

We now briefly describe each of these approaches and provide several examples that illustrate their strengths and weaknesses.

## NIST Actual Term-Weighted Value (ATWV)

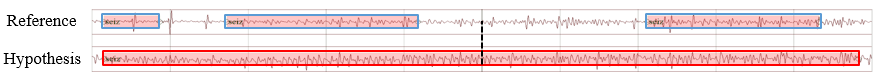
ATWV is a measure that balances sensitivity and FA rate. ATWV essentially assigns an application-dependent reward to each correct detection and a penalty to each incorrect detection. An ATWV greater than 0.5 typically indicates a system is usable. The metric accepts as input a list of *N*-tuples, each of which consists of a start time, end time and system detection score. These entries are matched to the reference occurrences using an objective function that accounts for both temporal overlap between the reference and posting list occurrences and the detection scores assigned by the system. The probabilities of miss and false alarm errors at a detection threshold *θ* are computed using a standard calculation [23]. A term-weighted value is then computed that specifies a trade-off between misses and false alarms. ATWV is defined as the value of TWV at the system’s chosen detection threshold. A standard implementation of this approach is available at *https://github.com/usnistgov/F4DE*.

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 approach will consider the detection rate (TP) to be *100%*. This is somewhat generous given that *50%* of the event was not detected. ATWV will, however, count *5* false positives. The ATWV metric is relatively insensitive to the duration of the reference event. However, the important issue here is that the hypothesis correctly detected about *80%* of the seizure event, and yet because of the large number of false positives, it will be penalized heavily.



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

In Figure 4, we demonstrate a similar case in which the metric penalizes the hypothesis heavily for missing events. The reference consists of *3* distinct short events in a *10-sec* span while hypothesis is one long event. As the mid-point of event in hypothesis does not fall within any events in the reference, ATWV considers all the reference events as miss. The *3* events are scored as FN errors even though the system has correctly identified all the seizure events. Additionally, the detected event in hypothesis is considered as an FA. While this type of scoring might make sense in an application like voice keyword search where each event carries a unique function in the overall understanding of the signal, for an application such as seizure detection, such scoring is penalizing too heavily the failure of the hypothesis to detect the boundaries of these isolated events.

**Figure 4.** ATWV scores this segment as *0* TP and *3* FNs.

## Dynamic Programming Alignment (DPALIGN)

The DPALIGN algorithm essentially performs a minimization of a Levenshtein distance function [12]. The quantities being measured here are often referred to as substitution, insertions and deletion errors. These are assigned costs and a dynamic programming algorithm decides the best alignment of the reference and hypothesis based on these weights. Though there are versions of this algorithm that perform time-aligned scoring in which both the reference and hypothesis must include start and end times, this algorithm is most commonly used without time-alignment information.

**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 \*\*\*\* \*\*\*\***

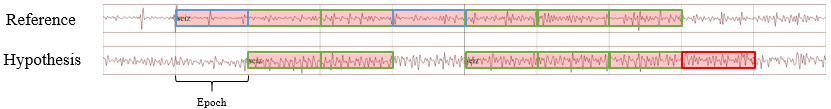
**(Hits: 5 Sub: 0 Ins: 0 Del: 2 Total Errors: 2)**

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

The algorithm is best demonstrated using the two examples shown in Figure 5. In the first example, the reference signal had three seizure events but the hypothesis only detected two seizure events, so there were two insertion errors. In the second example the hypothesis missed the third seizure event, so there were two deletion errors. For convenience, lowercase symbols indicate correct detections while uppercase symbols indicate errors. The asterisk symbol is used to denote deletion and insertion errors. Note that there is ambiguity in these alignments. For example, it is not really clear which of the three seizure events in the second example corresponded to each of the seizure events in the hypothesis. Nevertheless, this imprecision doesn’t really influence the overall scoring. 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].

## Epoch-Based Sampling (EPOCH)

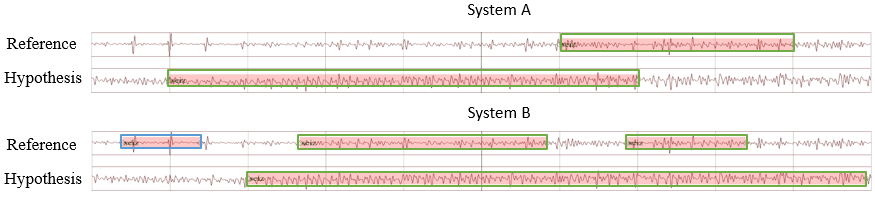
Epoch-based scoring uses a metric that treats the reference and hypothesis as signals. These signals are sampled at a fixed epoch duration. The corresponding label in the reference is compared to the hypothesis. This process is depicted in Figure 6. Epoch-based scoring requires that the entire signal be annotated, which is normally the case for sequential decoding problems. It attempts to account for the amount of time the two annotations overlap, so it directly addresses the inconsistencies Figure 3 and Figure 4.



**Figure 6.** EPOCH scoring directly measures the similarity of the time-aligned annotations. The detection rate, miss rate and false alarm rate of system are *71.4%*, *28.5%* and *33.33%* respectively.

## Any-Overlap Method (OVLP)

The neuroengineering community has favored a more permissive method of scoring known as the any‑overlap method [19] that tends to produce much higher sensitivities and lower false alarm rates. 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 still considered correctly detected. Multiple events in the hypothesis annotation corresponding to the same event in the reference annotation are not typically counted as false alarms. Since FA rate is a very critical measure of performance in critical care applications, this is a cause for concern. The OVLP scoring method is summarized in Figure 7.

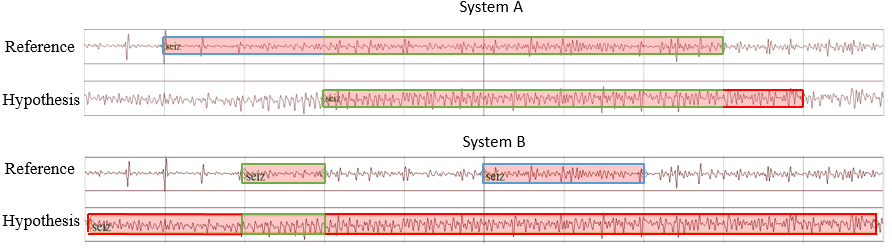


**Figure 7.** OVLP scoring is very permissive about the degree of overlap between the reference and hypothesis. Detection rate of system A is *100%* with no false alarm. System B detects *2* out of 3seizure events. The detection rate and miss rate of system B are *66.6%* and *33.3%* respectively.

## Time-Aligned Event Scoring (TAES)

Though EPOCH scoring directly measures the amount of overlap between the annotations, there is a possibility that this too heavily weights single long events. Seizure events can vary in duration from a few seconds to many minutes. In some applications, correctly detecting the number of events is as important as their duration. Hence, the TAES algorithm gives equal weight to each event, but it calculates a partial score for each event based on the amount of overlap.

The detection rate of a term is total duration of detected term divided by the total duration of reference term. Also, the miss rate is fraction of missing term over total duration of reference term. The FA is total duration of over-detection term divided by total duration of reference term. One more functionality borrowed from DPALIGN metric in TAES is sequence based calculation of miss rate. If the first event in hypothesis overlaps with more than one reference events, all remaining events but the first fall in to miss rate. Also, maximum false alarm rate obtained per reference event cannot exceed 100%. TAES scoring with its properties is depicted in Figure 8.



**Figure 8.** TAES scoring accounts for the amount of overlap between the reference and hypothesis. TAES scores system A as 71.42% detection rate (% overlap w.r.t. reference), 28.57% miss rate and 14.28% false alarm rate. System B is scored as 100% detection rate (first event), 100% miss rate (second event) and 100% FA.

## Inter-Rater Agreement (IRA)

Inter-rater agreement (IRA) is a popular measure when comparing the relative similarity of two annotations. This is most often measured using Cohen’s Kappa coefficient [36], which measures compares the observed accuracy with the expected accuracy. It is computed using:

 , (1)

where is the relative observed agreement among raters and  is the hypothetical probability of chance agreement. The Kappa coefficient ranges between (complete agreement) and (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 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.

## 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. The reference consists of *3* distinct short events while the hypothesis consists of one long detected event. The OVLP scores the segment as *3* TP and *0* FP because detected events have partial to full overlap with all the reference events. ATWV and DPALIGN scores the system as *1* TP and *2* FN but due to different reasons. ATWV considers the second event in the reference as correctly detected, while DPALIGN considers the first event in the reference as correctly detected and within the range of the same hypothesis event, it considers all the remaining events as FN. Also, the segment is scored as *5* TP, *3* FP and *1* FN by EPOCH. TAES scores this segment as 0.5 TP, 2.5 FN (50% from the first event and 2 following reference events) and 1 FP.



**Figure 9.** Performance of the section based on different metrics.

**A example where we compare the results of each algorithm**

**Figure 9.** ... a comparison example ...

# Results

In this study, we investigate the characteristics five different deep learning based seizure detection systems based on different evaluation metrics. An EEG signal is input to the systems, typically in the European Data Format (EDF) file. The signal is converted to a sequence of feature vectors. A group of frames are classified into an event on a per-channel basis using combination of deep learning networks. The deep learning 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. We explore five different deep learning systems, the comprehensive details about the architectures is available in [28], consisting of (1) hybrid of an HMM and a Stacked Denoising Autoencoder (SdA); (2) an HMM and a Long Short‑Term Memory (LSTM) network combination (3) compound of an Incremental Principal Component Analysis (IPCA) and a LSTM network; (4) a Convolutional Neural Networks (CNN) and a MultiLayer Perceptron (MLP) mixture; (5) hybrid of a CNN and LSTM networks. The reported results are based on the TUH EEG Seizure Corpus [29]. This dataset, which is publicly available, is a subset of the TUH EEG Corpus that focuses on the problem of seizure detection. A summary of the corpus (v1.1.1) is shown in table 1. A comparison of the performance of the different architectures is presented in table 2.

**Table 1.** An overview of the TUH EEG Seizure Corpus (V1.1.1).

|  |  |  |
| --- | --- | --- |
| **Description** | **Train** | **Eval** |
| Patients | 196 | 50 |
| Sessions | 456 | 230 |
| Files | 1,505 | 984 |
| Seizure (secs) | 51,140 | 53,930 |
| Non-Seizure (secs) | 877,821 | 547,728 |
| Total (secs) | 928,962 | 601,659 |

1. **Discussion**

In machine learning, finding suitable measures that show the characteristics of the system remains challenging. The commonly used and intuitive measure is accuracy which is formulated as:

 , (4)

The accuracy of the deep learning structures based on different evaluation metrics are depicted in table 3. The measure is not focusing on the specific class. The accuracy does not distinguish

**Table 2.** Performance of the deep learning systems based on different evaluation metrics on the TUH EEG Seizure Corpus.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Metric** | **HMM/ SdA** | **HMM/ LSTM** | **IPCA/ LSTM** | **CNN/ MLP** | **CNN/ LSTM** |
| **NIST STD** | **Sensitivity** | 30.35% | 26.73% | 24.73% | 29.52% | 30.18% |
| **Specificity** | 61.38% | 68.93% | 64.51% | 65.87% | 92.28% |
| **FAs/24 hrs** | 98 | 75 | 94 | 94 | 12 |
| **ATWV** | -0.8392 | -0.8469 | -0.4628 | -0.7971 | 0.1537 |
| **Any-overlap** | **Sensitivity** | 35.35% | 30.05% | 32.98% | 39.09% | 30.83% |
| **Specificity** | 73.35% | 80.53% | 77.58% | 76.84% | 97.10% |
| **FAs/24 hrs** | 77 | 60 | 73 | 77 | 6 |
| **DP Alignment** | **Sensitivity** | 44.11% | 33.77% | 35.77% | 43.35% | 32.13% |
| **Specificity** | 66.87% | 72.99% | 69.59% | 71.49% | 94.24% |
| **FAs/24 hrs** | 86 | 66 | 81 | 77 | 10 |
| **TAES** | **Sensitivity** | 17.67% | 22.94% | 23.08% | 32.12% | 11.33% |
| **Specificity** | 68.59% | 73.56% | 69.67% | 67.99% | 96.12% |
| **FAs/24 hrs** | 81 | 67 | 82 | 88 | 7 |
| **Epoch** | **Sensitivity** | 20.71% | 50.46% | 51.02% | 65.03% | 7.47% |
| **Specificity** | 98.22% | 94.82% | 94.09 | 91.55% | 99.84% |
| **FAs/24 hrs** | 1418 | 4133 | 4711 | 6738 | 125 |

between types of errors the system makes. This would be acceptable if the dataset is balanced. In the case of seizure and non-seizure detection system, the ignoring issue may lead to catastrophic accepting failure of the system. The issue is addressed by precision and recall and F-score measures. The precision and recall are computed as:

(5)

, (6)

The F- score convoy balance between precision and recall. The evenly balanced F-score, F1-score, is twice of the multiplication of precision and recall divided by the sum of precision and recall. The result of the deep learning systems based on F1-score is illustrated in table 4. They distinguish correct label classification within different classes. However, they do not indicate the true negative and specificity of the system.

The commonly accepted measures of reporting a performance of a system in clinical care application is sensitivity (TP/TP+FN), specificity (TN/TN+FP) and number of false alarms (FAs). These terms can be calculated based on different evaluation metrics. Each scoring metric performs optimally based on specific application revealing certain strength and weak points of a system. As an example, the overall behavior of each deep learning structures described can be inferred from table 2. Here, the operating points for each system has been selected to make results as comparable as possible, especially in terms of sensitivity and FA. Low sensitivity from Epoch-based scoring and moderate sensitivity from Any-Overlap method for HMM-SdA experiment suggests that the system is pruning to miss longer seizures.

**Table 3.** Accuracy of the deep learning structures based on different evaluation metrics.

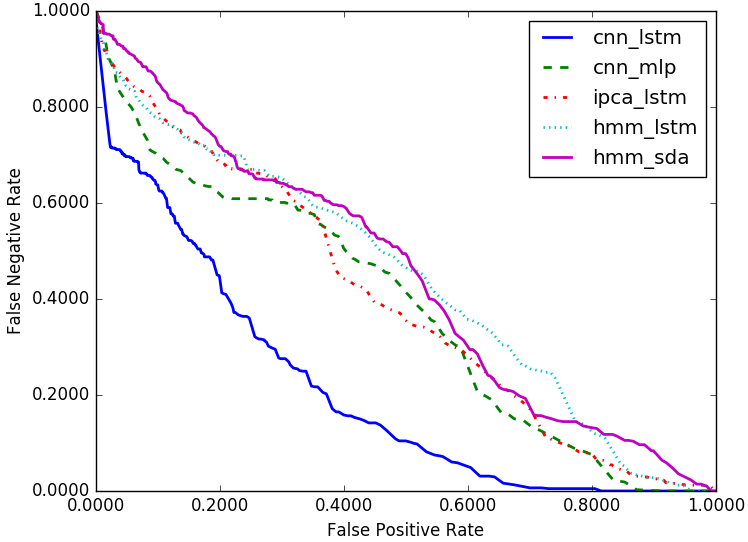
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **HMM/ SdA** | **HMM/LSTM** | **IPCA/LSTM** | **CNN/ MLP** | **CNN/LSTM** |
| **NIST STD** | 53.95% | 54.00% | 52.14 % | 54.87% | 70.74% |
| **Any-overlap** | 65.11% | 66.54% | 65.63% | 66.85% | 78.94% |
| **DP Alignment** | 61.49% | 60.20% | 59.16% | 62.88% | 73.61% |
| **TAES** | 56.55% | 57.31% | 55.38% | 57.17% | 69.71% |
| **Epoch** | 92.34 % | 91.45% | 90.82 % | 89.54% | 91.46% |

Conversely, HMM-LSTM and IPCA-LSTM experiments with higher sensitivities according to Epoch-based scoring and lower sensitivities according to TAES, STD and Any-Overlap method implies that these systems tend to detect the longer seizures. Results of the CNN-MLP experiment, with relatively low sensitivity according to STD than any other scoring metric implies that midpoint of mostly detected seizures do not fall within reference transcription’s range. This suggests temporal misalignments implying either delay or early detection. Finally, CNN-LSTM with lower sensitivity according to TAES and Epoch-based scoring suggests that the system is detecting seizures with smaller durations only.

However, if system can have higher sensitivity if one is willing tolerance a poor specificity or a high false alarm. The sensitivity, specificity and false alarm should be balanced. The STD balances these measures by introducing an ATWV which assigns a reward to each correct detection and a cost to each incorrect detection. A perfect system has the maximum ATWV which is one. ATWV of system with no output is zero [30]. Negative TWV is feasible. This measure is useful when it is preferred to compare two systems based on a single number, though it is always better to compare DET curves over a range of operating characteristics. ATWV and DET curves are our recommended way to evaluate EEG interpretation systems. Based on results in table 2 and DET curve (figure 6) and results in table 2, CNN/LSTM system has the best ATWV because of highest sensitivity and lowest false alarms. The ATWV score is extremely poor for these systems largely due to the large emphasis this metric place on false alarms. The fundamental cause of having high false alarm is poorly detection of long events.

**Table 4.** F1-score of the deep learning structures based on different evaluation metrics.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **HMM/ SdA** | **HMM/LSTM** | **IPCA/LSTM** | **CNN/ MLP** | **CNN/LSTM** |
| **NIST STD** | 0.2399 | 0.2817 | 0.2431 | 0.2837 | 0.4171 |
| **Any-overlap** | 0.3053 | 0.3324 | 0.3395 | 0.3843 | 0.4452 |
| **DP Alignment** | 0.3511 | 0.3562 | 0.3516 | 0.4166 | 0.4472 |
| **TAES** | 0.1613 | 0.2566 | 0.2410 | 0.3115 | 0.1890 |
| **Epoch** | 0.2910 | 0.4729 | 0.4580 | 0.4858 | 0.1370 |



**Figure 7.** DET curves of all systems.

None of the conventional metric described here considers the fraction of the detected event; which is the inspiration behind the development of TAES scoring. This approach of TAES scoring makes scoring rules stricter than any other metric. Consequently, the performance scored using this scoring method tend to have lower sensitivity as shown in table 2. All mentioned evaluation metrics are perfectly correlated. The detail of how the sensitivities of metrics are correlated with each other is depicted in table 5.

**Table 5.** Correlation between sensitivity of the evolution metrics.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **NIST STD** | **DP Alignment** | **Any-overlap** | **TAES** | **Epoch** |
| **NIST STAD** | 1 | 0.832997 | 0.80882 | 0.602636 | 0.265321 |
| **DP Alignment** | 0.832997 | 1 | 0.976855 | 0.888646 | 0.668998 |
| **Any-overlap** | 0.80882 | 0.976855 | 1 | 0.911713 | 0.699448 |
| **TAES** | 0.602636 | 0.888646 | 0.911713 | 1 | 0.924567 |
| **Epoch** | 0.265321 | 0.668998 | 0.699448 | 0.924567 | 1 |

1. **Conclusion**
2. A common reason for clinical practices not relying on commercially available tools is due to their high FA rate. This has been confirmed by conducting series of interviews with certified neurologists around United States [21]. This is perhaps the single most important consideration today in guiding machine learning research applications in critical care. Critical care units are overwhelmed with the number of FA that the automated tools generate. To put this in perspective, one FA per bed within 1 hour in a 12-bed Intensive Care Unit (ICU) causes 12 interrupts per hour that must be serviced. This can easily overwhelm healthcare providers. Because there are many types of automated monitoring equipment used in an ICU setting, each yielding significantly high FAs, the number of FAs that must be serviced by healthcare providers is overwhelming [22]. As a result, clinicians who report that in practice simply ignore these systems [2].

Finding the proper evaluation metrics that shows the deficiency and strength points of a system is the main concern of researchers. In this paper, we began by introducing the current evaluation metrics in biomedical filed. Various evaluation scoring metrics possess their own strengths and weaknesses that shadow the analyze of recognition system’s behavior. Usually, researchers report the performance base on only one metric which cannot provide all the subtle details about the behavior of a system and can be misleading during development. The preferred way of measuring the scores is to select multiple scoring metrics and weigh the results based on specific application. We design the TAES scoring which simply scores the fraction hits, miss and FA on event basis. Also, we propose Average Term Weighted Value (ATWV) as a unique measure to balance the sensitivity, specificity and false alarm.

**Acknowledgments**

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. The TUH EEG Corpus work was funded by (1) the Defense Advanced Research Projects Agency (DARPA) MTO under the auspices of Dr. Doug Weber through the Contract No. D13AP00065, (2) Temple University’s College of Engineering and (3) Temple University’s Office of the Senior Vice-Provost for Research.

**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. Retrieved from *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. Retrieved from *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. Retrieved from *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. Retrieved from *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. Retrieved from *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. Retrieved from *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. Retrieved from *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. Retrieved from *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. Retrieved from *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. In *Proceedings of LREC* (p. 106). Genoa, Italy. Retrieved from *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. Retrieved from *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). Retrieved from *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. Retrieved from *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. Retrieved from *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. Retrieved from *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. Retreived from *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. Retrieved from *http://lrec-conf.org/proceedings/lrec2006/pdf/813\_pdf.pdf*.
19. Wilson, S.B., Scheuer, M.L., Plummer, C., Young, B., & Pacia, S. (2003) Seizure detection: correlation of human experts. *Clinical Neurophysiology, 114*(11), 2156–64. Retrieved from *https://doi.org/10.1016/S1388-2457(03)00212-8*.
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. J., & Sackellares, J. C. (2010). Assessment of a scalp EEG-based automated seizure detection system. *Clinical Neurophysiology*, *121*(11), 1832–43. Retrieved from *https://doi.org/10.1016/j.clinph.2010.04.016*.
21. Baldassano, S., Wulsin, D., Ung, H., Blevins, T., Brown, M-G., Fox, E., & Litt, B. (2016). A novel seizure detection algorithm informed by hidden Markov model event states. *Journal of Neural Engineering,13*(3), 36011. Retrieved from *https://doi.org/10.1088/1741-2560/13/3/036011.*
22. Winterhalder, M., Maiwald, T., Voss, H. U., Aschenbrenner-Scheibe, R., Timmer, J., & Schulze-Bonhage, A. (2003). The seizure prediction characteristic: a general framework to assess and compare seizure prediction methods. *Epilepsy and Behavior*, *4*(3), 318–325. Retrieved from *https://doi.org/10.1016/S1525-5050(03)00105-7*.
23. Wegmann, S., Faria, A., Janin, A., Riedhammer, K., & Morgan, N. (2013). The TAO of ATWV: Probing the mysteries of keyword search performance. *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding* (ASRU). Olomouc, Czech Republic: IEEE. Retrieved from *https://doi.org/10.1109/ASRU.2013.6707728*.
24. Fiscus, J., Ajot, J., Garofolo, J., & Doddingtion, G. (2007). Results of the 2006 Spoken Term Detection Evaluation. In *Proceedings of the SIGIR 2007 Workshop*: *Searching Spontaneous Conversational Speech* (pp. 45–50). Amsterdam, Netherlands. Retrieved from *https://www.nist.gov/publications/results-2006-spoken-term-detection-evaluation*.
25. Japkowicz, N., & Shah, M. (2014). *Evaluating Learning Algorithms: a classification perspective*. New York City, New York, USA: Cambridge University Press. Retrieved from *https://doi.org/10.1017/CBO9780511921803*.
26. Confusion matrix. (n.d.). Retrieved October 31, 2017, from *https://en.wikipedia.org/wiki/ Confusion\_matrix*.
27. Liu, A., Hahn, J. S., Heldt, G. P., & Coen, R. W. (1992). Detection of neonatal seizures through computerized EEG analysis. *Electroencephalography and Clinical Neurophysiology*, *82*(2), 32–37. Retrieved from *https://doi.org/10.1016/0013-4694(92)90179-L*.
28. Navakatikyan, M. A., Colditz, P. B., Burke, C. J., Inder, T. E., Richmond, J., & Williams, C. E. (2006). Seizure detection algorithm for neonates based on wave-sequence analysis. *Clinical Neurophysiology*, *117*(6), 1190–1203. Retrieved from *http://dx.doi.org/10.1016/j.clinph.2006.02. 016*.
29. Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2017). The Microsoft 2017 Conversational Speech Recognition System. *https://arxiv.org/abs/1708.06073*.
30. 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. Retrieved from *https://doi.org/10.1016/S0013-4694(97)00003-9*.
31. 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. Retrieved from *https://doi.org/10.1016/S1388-2457(03)00212-8*.
32. Mason, S. J., & Graham, N. E. (2002). Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, *128*(584), 2145–2166. Retreived from *https://doi.org/10.1256/003590002320603584*.
33. Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian Journal of Internal Medicine*, *4*(2), 627–635. Retrieved from *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3755824/*.
34. Fiscus, J. G., & Chen, N. (2013). *Overview of the NIST Open Keyword Search 2013 Evaluation Workshop*. Bethesda, Maryland, USA. Retrieved from *http://ws680.nist.gov/publication/get\_pdf.cfm?pub\_id=914517*.
35. National Institute of Standards and Technology (NIST) (2017). *Speech Recognition Scoring Toolkit*. Retrieved from *https://github.com/usnistgov*.
36. McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3), 276–282. Retrieved from *http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/*.
37. Halford, J. J., Shiau, D., Desrochers, J. A., Kolls, B. J., Dean, B. C., Waters, C. G., … LaRoche, S. M. (2015). Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG recordings. *Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology*, *126*(9), 1661–9. Retrieved from *https://doi.org/10.1016/j.clinph.2014.11.008*.

gggggggggggggggggggggggggggggggggggggggggg

1. Provost, F., Fawcett, T., & Kohavi, R. (1997). The Case Against Accuracy Estimation for Comparing Induction Algorithms. *Proc. Fifteenth International Conference on Machine Learning* (pp. 445–453). Retrieved from *http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.8058.pdf.*
2. Wilson SB, Scheuer ML, Plummer C, Young B and Pacia S 2003 Seizure detection: correlation of human experts *Clinical Neurophysiology* **114** 2156–64
3. Schedl M, Gómez E and Urbano J 2014 Music information retrieval: Recent developments and applications *Foundations and Trends in Information Retrieval* **8** 127–261
4. Obeid I and Picone J 2015 *NSF ICORPS Team: AutoEEG. NSF Innovation Corps (I-CORPS)* (National Science Foundation) Available from: <http://www.isip.piconepress.com/proposals/2015/nsf/icorps>
5. Christensen M, Dodds A, Sauer J and Watts N 2014 Alarm setting for the critically ill patient: a descriptive pilot survey of nurses’ perceptions of current practice in an Australian Regional Critical Care Unit *Intensive and Critical Care Nursing* ***30*** 204–10.
6. Mandal A, Prasanna Kumar KR and Mitra P 2014 Recent developments in spoken term detection: a survey *International Journal of Speech Technology* **17** 183–98
7. Doddington GR, Przybocki MA, Martin AFand Reynolds DA 2000 The NIST speaker recognition evaluation – Overview, methodology, systems, results, perspective *Speech Community* **31** 225–54
8. Martin A, Doddington G, Kamm T, Ordowski M and Przybocki M 1997 The DET curve in assessment of detection task performance *Proc. of Eurospeech* *(Greece)* pp. 1895–1898
9. Jacobs IM and Wozencraft JM 1965 *Principles of communication engineering (*Long Grove, Illinois USA) Waveland Pr Inc p 720
10. Kuhn HW 2010 *The Hungarian method for the assignment problem* Jünger Jünger M, Liebling TM, Naddef D, Nemhauser GL, Pulleyblank WR, Reinelt G,Rinaldi, G. Wolsey, L.A. 50 Years of Integer Programming 1958-2008: From the Early Years to the State-of-the-Art Springer (Berlin Heidelberg) pp 29–47
11. Golmohammadi M, Ziyabari S, Shah V, Obeid I and Joseph P 2017 Deep Architectures for Automated Seizure Detection in Scalp EEGs *AAAI*
12. Shah V, Von Weltin E, Lopez S, Golmohammadi M, Ziyabari S, Obeid I and Picone J 2017 The TUH EEG Seizure Corpus *Front Neuroscience* p 6
13. Fiscus J, Ajot J, Garofolo J and Doddingtion G 2007 Results of the 2006 Spoken Term Detection Evaluation *Proc.s of the SIGIR 2007 Workshop: Searching Spontaneous Conversational Speech*  (Amsterdam, Netherlands) pp 45–50