**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 evaluation metrics such as sensitivity, specificity, false alarm rate and receiver operating characteristic 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 numeric 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 [x]-[x]. 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 [x]. 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 this type of technology: user acceptance testing [x][x] and objective performance evaluation using annotated reference data. User acceptance testing is expensive and time-consuming to conduct and has never been a practical way for guiding technology development. Hence evaluations using objective performance metrics, such as sensitivity and specificity, are common in the machine learning field [x][x]. 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 [x], 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., “seizure” or “no seizure”). 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 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, Smith et al. (2XXX) advocates using a metric involving.... Jones et al. (2XXX) advocates using a metric involving... Doe et al. (2XXX) advocates using a ... The ranking of various machine learning approaches, and the overall level of performance, is highly dependent on the choice of a metric. As we will show, the same system can range in performance from a sensitivitivy of 30% to 90% depending on the scoring 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 level of complication to the process. Smith et al. (2XXX) have studied this problem extensively and argue for a particular type of 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, the goal of this paper is to ...

# Method

 sensititivity depe

 The performance of a system is assessed by the scoring procedures that detect and classify errors by comparing the reference and the hypothesis transcriptions. Each evaluation system reveals different aspects of systems’ behaviors. Choosing an appropriate evaluation metric for an application is a challenge within itself and is the focus of this study. The result of the classifier is presented in a confusion matrix, which gives a very useful overview of each scoring metric’s performance. For example, for a two-class problem such as seizure/non-seizure event detection, a confusion matrix has the following categories:

True Positives (TP) refer to the number of seizure events detected correctly.

True Negatives (TN) refer to the number of non-seizure events detected correctly.

False Positives (FP) refer to the number of non-seizure events incorrectly detected as seizure.

False Negatives (FN) refer to the number of seizure events incorrectly detected as non-seizure.

Sensitivity or detection rate (TP/TP+FN), specificity (TN/TN+FP) and miss rate (FN/FN+TP) are derived from these quantities.

Researchers typically report performance in terms of sensitivity and specificity [15] of epochs [16,17] or terms [18] in biomedical research applications. An epoch is defined as a partition of EEG that lasts for a while . The duration of epoch can be different based on the need of application. For the seizure detection application in our study, we define epoch size as one second. The epoch-based scoring metric considers each epoch as a separate testing example even though EEG events can span multiple epochs. Figure 1 shows performance of an electrographic seizure detection system based on epochs. Because 5 epochs in hypothesis are in alignment with 10 epochs in reference, the performance of the system based on Epoch-based evaluation metric is 50% detection rate and 50% miss rate.



**Figure 1.** A system detects about 5 seconds out of 10 seconds seizure, the rates of hits and misses of the system based on epoch-based evaluation metric are 50%.

The term-based metrics connect the subsequent events of the same class to create a term. The typical approach of calculating term-based evaluation metric is Any-Overlap. True Positives (TPs) are counted when the hypothesis overlaps with reference annotation. False Positives(FPs) correspond to the events in which the hypothesis does not overlap with the reference [18,19]. The metric ignores the duration of the correct term. Figure 2 shows a more offhand way of the functioning of the method where the detection rate of the system is 100% while system A does not detect about 9 seconds of a seizure event and system B detects about 9 seconds of event additional to the reference event.



**Figure 2.** Detection rates of both systems based on Any-Overlap metric are 100%.

Additionally, discussing one more popular metric in biomedical field called Dynamic Programming (DP) alignment metric. This metric is commonly used in DNA sequence matching to efficiently align two strings with different lengths and compute distance between them. The DP alignment method is based on minimizing the Levenshtein distance. To find the alignment with minimum score using DP, the error matrix of two sequences is computed. The error matrix E (i, j) is distance between element ith element in first sequence and jth element in second sequence [20]. The DP alignment scoring metric emphasizes on the sequence of event’s appearance with respect to the reference event. Both Any-Overlap and DP alignment metric suffer from lack of focus on temporal domain. Apart from holding few similar features, these two metrics are significantly different among themselves. One obvious difference to notice is: Any-Overlap metric considers partial overlap as a hit score whereas DP alignment only relies on the reference and hypothesis event’s sequences.

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].

The balance between sensitivity and FA rate is desired for developing a stable system and has been studied extensively in other communities focused on event-spotting technology such as Spoken Term Detection (STD) in voice signals [23]. In this article we introduce a measure that we borrow from this research community is the Actual Term-Weighted Value (ATWV) [24] which is based on the notion of a Detection Error Tradeoff (DET) curve [25]. A DET curve is very similar to a Receiver Operating Characteristic (ROC) originally developed to assess the performance of a communications system [26]. Also, the article argues that the commonly-accepted evaluation metrics does not fully meet the need of machine learning problems in the EEG research community. The deficiencies of the current evaluation measures lead us to suggest Spoken Term Detection (STD) which is used widely in speech recognition field and Time-Aligned Event Scoring **(**TAES) which counts fractional hits, miss and false alarms per event as a novel metric for biomedical applications.

In this section, we explore the Spoken Term Detection (STD) and Time-Aligned Event Scoring **(**TAES) algorithms. The Spoken Term Detection (STD) is a speech processing task in which the goal is to find all the occurrences of a textual “keyword”, a sequence of one or more words, in a large corpus of speech data. In 2006, the U.S. National Institute of Standards and Technology (NIST) created the STD evaluation initiative to facilitate research and development of technology for retrieving information from archives of speech data. STD detects all the occurrences of each given term in the reference files. In this evaluation methodology, an estimate is required for the number of trails in the reference. If there are no discrete trials in a continuous reference, a constant will be specified as number of trials. Then, alignment between detected occurrence and reference is needed. This step is done by applying the Hungarian solution to the Bipartite Graph [27] matching problem. It uses the kernel function that numerically compares the mapping of system and reference occurrences, as well as the missed detections and false alarms. The kernel function first determines if the reference/system occurrences are mappable by requiring the system occurrence to be within a temporal tolerance collar (ΔT) of the reference occurrence. Specifically, the midpoint of the system occurrence must be within the interval from ΔT before the beginning to ΔT after the end of the reference occurrence. The performance of a system is evaluated base on detection error tradeoff (DET) curves and an Actual Term-Weighted Value (ATWV) for a specific operating point. ATWV essentially assigns an application-dependent reward to each correct detection and a penalty to each incorrect detection. The ATWV is a measure for balancing between sensitivity and FA rate. The commonly acceptable system has ATWV greater than 0.5.



**Figure 2.** The system has one True Positive (TP) and five False Positives (FPs) based on STD evaluation metric.

The STD method, just like any other metrics, has its own set of challenges including approach of calculating detection rate and number of FAs. The deficiency of the method in computing detection rate is similar to Any-Overlapped approach where both metrices ignore duration of correct term. For example, detection rates of systems which are illustrated in figure 1 and figure 2 based on STD method are 100% without any misses and FAs, so that yields a system with perfect sensitivity. Regarding the method of counting FA, consider a case which is illustrated in figure 3, there is a 10 second seizure in reference; the system detects 6 distinct short seizures. The STD considers the first event as a hit and all the remaining ones as FAs even when their midpoints fall within the duration of reference event. This implies that only one hypothesis event is accepted as a true detection. Consequently, the system has 1 TP and 5 FPs.

The other challenge of a similar case is shown in figure 4, where there are 5 distinct short seizures in reference and system detects one long seizure. The midpoint of the hypothesis falls within the range of third seizure event in reference. Therefore, the performance of the system is 1 TP, 5 FNs.



**Figure 3.** There are one True Positive(TP) and five False Negatives(FNs) and five False Positives (FPs) based on STD evaluation metric.

The STD algorithm is based on an overall assessment of metric parameters. Therefore, some information regarding the characteristics of the identified events, such as event duration, may be lost. Therefore, in this study, we propose an additional algorithm that calculates the fraction of duration of (1) detected event, (2) missed event and (3) False Alarm (FA). The performance of the system which is depicted in figure 1 based on the TAES is 0.5 TPs, 0.5 FNs and zero FP. Also, the performance of system in figure 3 presents 0.72 TPs and 0.28 FNs and 0 FP. In TAES algorithm, 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 missed term divided by total duration of reference term. The TAES algorithm is formulated as follow:

$TP\_{TAES}= \frac{ THyp\_{stop}-THyp\_{start}}{TRef\_{stop}- TRef\_{start}}, where TRef\_{start} \leq THyp \leq TRef\_{stop}$ , (1)

 $TN\_{TAES}= \frac{ 1-(THyp\_{stop}-THyp\_{start})}{TRef\_{stop}- TRef\_{start}}, where TRef\_{start} \leq THyp \leq TRef\_{stop}$ , (2)

$FP\_{TAES}=\left\{\begin{array}{c}\frac{THyp\_{stop}- TRef\_{stop}}{TRef\_{stop}- TRef\_{start}}, if THyp\_{stop}\geq TRef\_{stop}, THyp\_{start}\geq TRef\_{start} and THyp\_{stop}- TRef\_{stop}\leq 1;\\ where THyp \in [ TRef\_{start}, TRef\_{stop} ] \\\frac{TRef\_{start}- THyp\_{start}}{TRef\_{stop}- TRef\_{start}}, if TRef\_{start}\geq THyp\_{start}, TRef\_{stop}\geq THyp\_{stop} and TRef\_{start}- THyp\_{start}\leq 1;\\where THyp \in [ TRef\_{start}, TRef\_{stop} ]\\\frac{(THyp\_{stop}- TRef\_{stop}) + (TRef\_{start}- THyp\_{start})}{TRef\_{stop}- TRef\_{start}}, if THyp\_{stop}\geq TRef\_{stop}, TRef\_{start}\geq THyp\_{start} \\and THyp\_{stop}- TRef\_{stop}\leq 1; where THyp \in [ TRef\_{start}, TRef\_{stop} ]\\otherwise 1\end{array}\right. $ (3)

1. **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:

$Acuuracy=\frac{tp+tn}{tp+fn + tn+fp}$ , (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:

$Precision=\frac{tp}{tp+fp} ,$ (5)

$Recall=\frac{tp}{tp+fn}$, (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 5.** 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**

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 Tand Meng E 2009
2. Obeid I and Picone J 2017 *Machine Learning Approaches to Automatic Interpretation of EEGs* Sejdik E, Falk T Biomedical Signal Processing in Big Data (Boca Raton, Florida, USA) CRC Press (www.isip.piconepress.com/publications/book\_sections/2017/crc\_press/auto\_eeg/)
3. Gotman J 1982 Automatic recognition of epileptic seizures in the EEG *Electroencephalogr Clinical Neurophysiology* **54** 530–40
4. Gotman J 1999 Automatic detection of seizures and spikes. *J Clinical Neurophysiology* **16** 130–40
5. Osorio I, Frei MG and Wilkinson SB 1998 Real-time automated detection and quantitative analysis of seizures and short-term prediction of clinical onset *Epilepsia* **39** 615–27
6. Sartoretto F and Ermani M 1999 Automatic detection of epileptiform activity by single-level wavelet analysis *Clinical Neurophysiology* **110** 239–49
7. Schad A, Schindler K, Schelter B, Maiwald T, Brandt A, Timmer J and Schulze-Bonhage A 2008 Application of a multivariate seizure detection and prediction method to non-invasive and intracranial long-term EEG recordings *Clinical Neurophysiology* **119** 197–211
8. Schindler K, Wiest R, Kollar M and Donati F 2001 Using simulated neuronal cell models for detection of epileptic seizures in foramen ovale and scalp EEG *Clinical Neurophysiology* **112** 1006–17
9. Stam CJ 2005 Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field *Clinical Neurophysiology* ***116 2266-2301***
10. Deburchgraeve W, Cherian PJ, De Vos M, Swarte RM, Blok JH, Visser GH, Govaert P and Van Huffel S 2008 Automated neonatal seizure detection mimicking a human observer reading EEG *Clinical Neurophysiology* **119** 2447–54
11. Khamis H, Mohamed A and Simpson S 2009 Seizure state detection of temporal lobe seizures by autoregressive spectral analysis of scalp EEG *Clinical Neurophysiology* **120** 1479–88
12. Ramgopal S 20014 Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy *Epilepsy& Behavior* **37** 291–307.
13. Alotaiby T, Alshebeili S, Alshawi T, Ahmad I and Abd El-Samie F 2014 EEG seizure detection and prediction algorithms: a survey *EURASIP J. Adv. Signal Process* **2014** 1–21
14. Temko A, Korotchikova, Marnane W, Lightbody G and Boylan G 2010 EEG-based neonatal seizure detection with Support Vector Machines *Proc. 3rd Int. Conf. Bio-inpsired Systems Signal Processing Proc.(Valencia)* PP 312-317
15. Japkowicz N and Shah M 2011 *Evaluating Learning Algorithms:* *A Classification Perspective* (New York, USA) Cambridge University Press. p 423
16. Liu A, Hahn JS, Heldt GP and Coen RW 1992 Detection of neonatal seizures through computerized EEG analysis. *Electroencephalogr Clinical Neurophysiology* **82** 30-37
17. Navakatikyan MA, Colditz PB, Burke CJ, Inder TE, Richmond J and Williams CE 2006 Seizure detection algorithm for neonates based on wave-sequence analysis *Clinical Neurophysiology* **117** 1190–203
18. Gotman J, Flanagan D, Zhang J and Rosenblatt B 1997 Automatic seizure detection in the newborn: Methods and initial evaluation *Electroencephalography Clinical Neurophysiol*ogy **103** 356–62
19. Wilson SB, Scheuer ML, Plummer C, Young B and Pacia S 2003 Seizure detection: correlation of human experts *Clinical Neurophysiology* **114** 2156–64
20. Schedl M, Gómez E and Urbano J 2014 Music information retrieval: Recent developments and applications *Foundations and Trends in Information Retrieval* **8** 127–261
21. 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>
22. 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.
23. 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
24. Doddington GR, Przybocki MA, Martin AFand Reynolds DA 2000 The NIST speaker recognition evaluation – Overview, methodology, systems, results, perspective *Speech Community* **31** 225–54
25. 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
26. Jacobs IM and Wozencraft JM 1965 *Principles of communication engineering (*Long Grove, Illinois USA) Waveland Pr Inc p 720
27. 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
28. Golmohammadi M, Ziyabari S, Shah V, Obeid I and Joseph P 2017 Deep Architectures for Automated Seizure Detection in Scalp EEGs *AAAI*
29. 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
30. 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