Validation of Temporal Scoring Metrics for Automatic Seizure Detection

V. Shah[1], I. Obeid[1], J. Picone[1], G. Ekladious[2], R. Iskander[2] and Y. Roy[3]

1. Neural Engineering Data Consortium, Temple University, Philadelphia, Pennsylvania, USA
2. Novela Neurotech, Alameda, California, USA
3. University of Montreal and NeuroTechX, Montreal, Quebec, Canada
{vinitshah, obeid, picone}@temple.edu,
geskander2014@gmail.com, ray@novelaneuro.com, yannick@neurotechx.com

Abstract— There has been a lack of standardization of the evaluation of sequential decoding systems in the bioengineering community. Assessment of the accuracy of a candidate system's segmentations and measurement of a false alarm rate are examples of two performance metrics that are very critical to the operational acceptance of a technology. However, measurement of such quantities in a consistent manner require many scoring software implementation details to be resolved. Results can be highly sensitive to these implementation details.

In this paper, we revisit and evaluate a set of metrics introduced in our open source scoring software for sequential decoding of multichannel signals. This software was used to rank sixteen automatic seizure detection systems recently developed for the 2020 Neureka® Epilepsy Challenge. The systems produced by the participants provided us with a broad range of design variations that allowed assessment of the consistency of the proposed metrics. We present a comprehensive assessment of four of these new metrics and validate our findings with our previous studies. We also validate a proposed new metric, time-aligned event scoring, that focuses on the segmentation behavior of an algorithm. We demonstrate how we can gain insight into the performance of a system using these metrics.

I. Introduction

An electroencephalogram (EEG) is a popular non-invasive tool for recording signals and diagnosing brain-related illnesses [1]. We have recently developed a number of resources to promote next generation research into automatic seizure detection [2]-[4]. With recent advances in machine learning technology, there is increasing interest in developing an automated seizure detection technology for clinical use [5]. Open source challenges such as those conducted by the speech and image recognition research communities [6]-[8], have been proven to enrich the research, expedite progress, share knowledge and unify the efforts of academic researchers and industrial partners.

In 2020, the first Neureka® Epilepsy Challenge [9], sponsored by Novela Neurotech, was conducted. A total of 19 teams participated, 16 of which provided submissions that could be scored using the approaches described in this paper. The competition used v1.5.1 of the TUH EEG Seizure Detection Corpus [10] and v3.3.3 of the NEDC open source scoring software [11]. The latter was a simplified version of our open source evaluation software designed to make it easy for participants to interface their systems to the scoring tools. The time-aligned event scoring metric was used as a basis for ranking the competition submissions. A heuristic penalty was added for the number of channels used in data processing in an effort to encourage novel solutions using a small number of channels. The final leaderboard is available from the challenge web site [12], as are the details of the competition. In this paper, we use the results from this competition to further validate our scoring metrics since we now have access to independent evaluation data.

II. Evaluation metrics

Any algorithm used to evaluate sequential decoding in machine learning must compute one or more of these four fundamental quantities: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). However, as shown in Figure 1, there are many ways to calculate these quantities depending on what criteria are used to assess overlap. We have implemented four standard scoring metrics in our scoring tools [4][11]: NIST actual term-weighted value (ATWV), dynamic programming alignment (DPAL), epoch-based sampling (EPCH) and the any-overlap method (OVLP). The latter is popular in the EEG research community. The first two have been widely used in other research communities.

We also have introduced a new metric, time-aligned event scoring (TAES), which is summarized in Figure 2. Though EPCH 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 hours. In some applications, correctly detecting the number of events is as important as their duration. Hence, the TAES metric was designed as a compromise to these competing constraints. TAES gives equal weight to each event, but it calculates a partial score for each event based on the amount of overlap. TAES also penalizes multiple overlapping reference events with a single hypothesis event as shown in example 2 of Figure 2. TAES was chosen to be the primary evaluation metric for the Neureka® challenge because it heavily weights errors in the temporal alignment of a hypothesis.

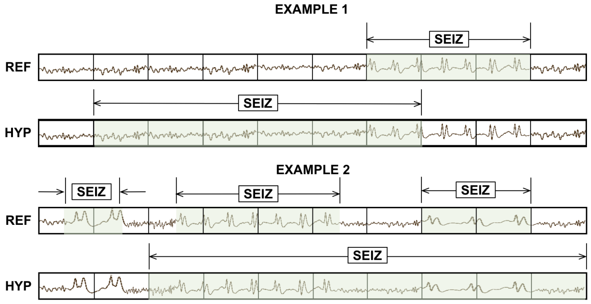

Figure 1. OVLP scoring is very permissive about the degree of overlap between the reference and hypothesis. A lack of standard rules for assessing overlap makes it difficult to directly compare research results.

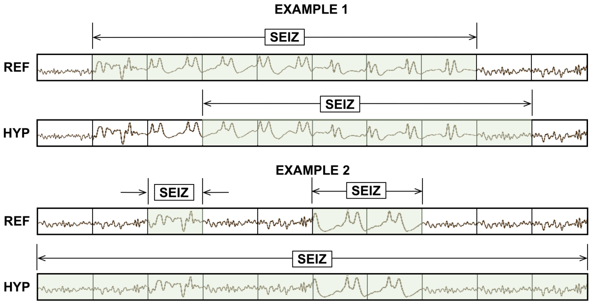

Figure 2. TAES scoring accounts for the amount of overlap between the reference and hypothesis. TAES scores example 1 as 0.71TP, 0.29 FN and 0.14 FP. Example 2 is scored as 1 TP, 1 FN and 1 FP.

We also added a penalty for the number of channels used by a system. The overall metric was:

$$P = sens - 2.5 * (fa) - 7.5 * (nc/19) \ , \tag{1}$$

where $sens$ is the sensitivity, fa is the false alarm rate, both as measured by TAES, and *nc* is the number of channels used to process the EEG signal. False alarm rate is penalized very heavily because clinicians are emphatic about keeping the false alarms to as low as possible (e.g. less than 1 per 24 hours). The original data had a maximum of 19 signal channels. The weights were set experimentally based on our expectations of performance for this data set [13].

TUSZ v1.5.1 was used for this competition. Three sets of files were provided to the participants: *train*, *dev* and *eval*. Reference annotations were provided for *train* and *dev*, but obviously not for the blind evaluation set. Statistics of this database is provided in Table 1. The statistics of the *eval* set are very similar to the *dev* set. In fact, performance on the *eval* set correlates well with the *dev* set, so improvements on the *dev* set, for which reference annotations were provided, generally hold up on the *eval* set if a system is not overtrained. All three databases of TUSZ have no overlap between patients. To evaluate their model performance, participants were given the scoring software with four evaluation metrics (the NIST ATWV metric was excluded to make software installation easy and minimize portability issues).

Table 1. TUSZ v1.5.1 statistics

Description	**Train**	**Dev**
Patients	592	50
Sessions	1185	238
Files	4599	1013
#Seizure events	2377	673
Seizure Dur (sec.)	169,794	58,445
Total Dur (sec.)	2,710,483	613,232

III. Competition Results

To understand the performance differences between various metrics, we requested a few participants to provide their results on both the dev and eval sets. Three of these models (*lzk* [14], *pnc98* [15], and *yff* [16]) along with an internally developed baseline system (*nedc* [17]) are shown below in Table 2 and Table 3. The system labeled *nedc* was one of the systems used to evaluate the scoring software as it was being developed. The other

Table 2. Dev set performance

Scores		nedc	1zk	pnc98	yff
DPAL	Sens	37.23	27.64	6.98	20.51
	Spec	96.88	85.27	98.33	91.98
	FPs	5.63	29.16	2.54	14.09
EPCH	Sens	36.28	13.78	1.56	31.18
	Spec	97.30	97.89	99.99	94.06
	FPs	2,101.13	1,647.00	8.45	4,644.04
OVLP	Sens	40.29	23.92	6.39	26.15
	Spec	97.56	90.29	99.65	94.19
	FPs	5.77	25.36	0.85	14.23
TAES	Sens	32.60	14.36	2.04	14.03
	Spec	90.72	83.53	99.42	87.44
	FPs	17.03	31.32	0.87	21.42

Table 3. Eval set performance

Scores		nedc	1zk	pnc98	yff
DPAL	Sens	42.96	31.90	8.61	24.66
	Spec	94.39	92.58	99.44	95.91
	FPs	11.77	15.11	0.95	7.79
EPCH	Sens	51.58	22.36	5.09	32.33
	Spec	98.38	99.14	100.00	96.58
	FPs	1,301.09	692.58	2.07	2,750.99
OVLP	Sens	42.96	30.33	8.41	24.66
	Spec	95.54	94.75	99.93	96.02
	FPs	11.45	13.52	0.16	10.02
TAES	Sens	35.55	19.99	2.04	16.00
	Spec	91.80	92.21	99.90	91.36
	FPs	17.23	15.59	0.17	16.54

three systems came from the Neureka® competition [9] and had never been previously evaluated with this software. Though it would have been ideal to compare receiver operating characteristic (ROC) or detection error tradeoff (DET) curves for these systems, it was not possible to get this type of data from many of the competitors. Although each model is operating at a different point on a ROC curve, we see that performance of each metric is consistent between the *dev* and *eval* sets. Further, each model performs slightly better on the *eval* set than the *dev* set. This suggests that *eval* set is slightly easier than the *dev* set.

Comparing performance of two systems across a range of metrics can often provide insight into an algorithm's deficiencies. To understand this, let's compare the performance of *pnc98* and *yff* on the dev set in Table 2. *Pnc98*'s performance is consistent between the pairs (DPAL, OVLP) and (EPCH, TAES). The performance of *yff* according to EPCH and TAES is vastly different. EPCH sensitivity is almost double the sensitivity calculated by TAES (31.18% versus 14.03%). This suggests that *yff* is biased towards longer seizure events. This model shows consistent performance on the *dev* set based on the DPAL and OVLP metrics (~20-26% sensitivity with ~14 FAs / 24 hours) but a very high false alarm rate based on EPCH metric. This significant difference of sensitivity and FA rate according to the EPCH metric compared to other metrics suggests that this model suffers from over-detection (detection outside the reference event boundaries) of seizure events. DPAL and OVLP show similar levels of performance because both metrics are very lenient in penalizing errors.

IV. Analysis of the Metrics

To understand the differences between all five metrics, we compare the best performing system (*sia* [18]) with the baseline system (*nedc*) in Table 4. We also provide a DET curve analysis in Figure 3. We show only the range from [0, 0.2] because this is the range of greatest interest for this particular task. Both models perform similarly according to the ATWV, DPAL and OVLP metrics. This suggests that both models are comparable in terms of alignments, sequence of event strings, and event detection. These metrics do not incorporate the estimate of the seizure event's duration.

According to EPCH, however, the sensitivity of *nedc* improves (51.57%) whereas the sensitivity worsens for *sia* (12.83%). TAES calculates a slightly lower sensitivity (35.54%) for *nedc* but a higher FA (17.22). Sensitivity for *sia* drops to 11.37% at ~1 FA/24 hours. These differences between EPOCH and TAES suggests that the *nedc* system tends to detect seizures over a wide range of durations (brief and prolonged seizures). *Sia* tends to detect seizure events of moderate duration.

Table 4. Performance comparison using all 5 metrics

Scores		*sia*	*nedc*
ATWV	Sens	22.70	41.08
	Spec	99.02	93.20
	FPs	1.61	13.36
DPAL	Sens	23.45	42.96
	Spec	99.47	94.39
	FPs	0.96	11.77
EPCH	Sens	12.84	51.58
	Spec	99.97	98.38
	FPs	25.85	1,301.09
OVLP	Sens	23.26	42.96
	Spec	99.74	95.54
	FPs	0.64	11.45
TAES	Sens	11.37	35.55
	Spec	99.46	91.80
	FPs	0.99	17.23

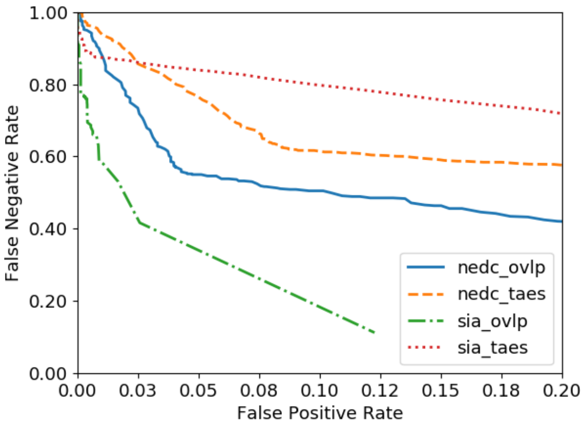

Figure 3. DET curves using the OVLP and TAES metrics

This analysis becomes very clear in Figure 4 where the duration of all the detected events are plotted within the range [0, 300] seconds. The majority of the detections of *nedc* occur for durations of 4 seconds or greater. *Sia* tends to detect most of its seizure events within a duration of [15, 40] seconds. This observation is supported by the TAES metric which tends to penalize errors on very short and very long events heavily.

Overall, *nedc* performs similarly in terms of sensitivity for all metrics. Seizure sensitivities remain within the range of [35%, 50%] though the FA rate fluctuates. In contrast, *sia's* performance is very stable in terms of FAs since the FA rate remains within a range of [0.64, 1.61].

V. Statistical Analysis

We performed a statistical analysis by calculating Pearson's correlation coefficient [19] for all four metrics. The samples for this experiment are results collected from all the participants. We estimate pairwise

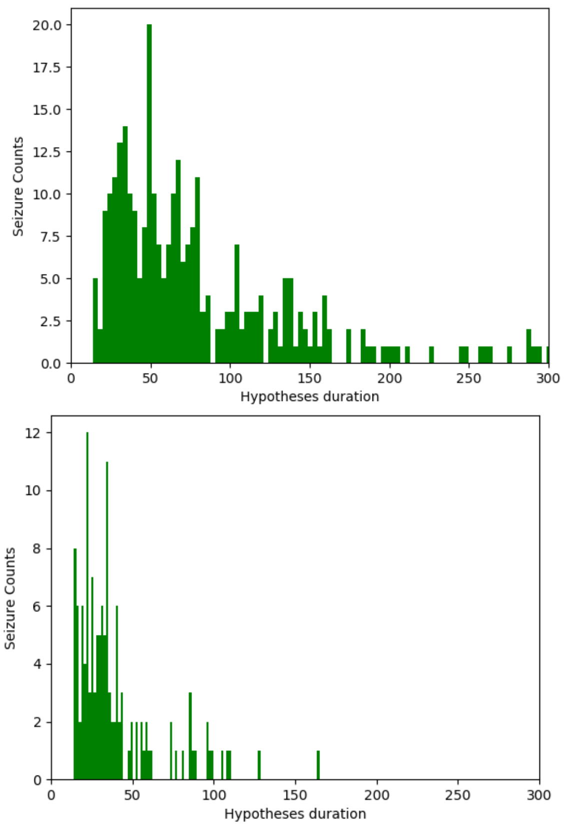

Figure 4. Duration distributions of hypothesized events: *nedc* system (top) versus *sia* system (bottom)

correlation between the metrics and provide the p-values to show the significance (reliability) of the results. Typically, p-values below 0.05 suggest statistically significant results. Values below 0.001 suggest a high confidence in the significance of the results.

Table 5 and Table 6 shows the correlation between metrics for sensitivity and specificity respectively. OVLP and DPAL are highly correlated in terms of both sensitivity and specificity. The only statistically nonsignificant correlation observed is between DPAL and EPCH (0.4029 with p-value = 0.121) for sensitivity. EPCH tends to correlate poorly with all other metrics (~0.6) except TAES (~0.9). This makes sense since both of these metrics compare partial overlaps and treat the sequence of events as a time series. Since EPCH does not score on an event basis, it is inherently biased to favor longer duration events. EPCH specificities do not correlate well with other metrics due to the high imbalance between seizure and background. Seizure frames are only 7% of the data, makes the specificity very high since it is dominated by the background class.

One caveat in this analysis is that in this data, DPAL specificity seems to be highly correlated with TAES specificity. This can be attributed to how models were designed for this competition. Because an extremely low FA rate was encouraged, the specificities for DPAL and TAES are similar even though both metrics have completely different objectives.

Table 5. Correlation between metrics for sensitivity

	DPAL	EPCH	OVLP	TAES
DPAL	1.00	0.4029 (p=0.121)	0.9535 (p<0.001)	0.5746 (p=0.019)
EPCH	—	1.00	0.6141 (p=0.011)	0.9144 (p<0.001)
OVLP	—	—	1.00	0.7641 (p<0.001)
TAES	—	—	—	1.00

Table 6. Correlation between metrics for specificity

	DPAL	EPCH	OVLP	TAES
DPAL	1.00	0.6676 (p=0.004)	0.9985 (p<0.001)	0.9980 (p<0.001)
EPCH	—	1.00	0.6777 (p=0.003)	0.6948 (p=0.002)
OVLP	—	—	1.00	0.9971 (p<0.001)
TAES	—	—	—	1.00

TAES scores events considering the partial overlaps and weighs all the seizure events equally. This provides the best balance between epoch-based and event-based scoring. From this analysis, we can observe that TAES metric incorporates both event-specific scores and partial overlap scores between the reference and hypothesis events. This is why we favor this metric for evaluation of sequential decoding systems.

VI. Summary

In this paper, we have validated our findings from a previous study [4] on a new evaluation task involving a wide variety of machine learning systems. We have analyzed the results of the 2020 Neureka® Epilepsy Challenge and demonstrated that the TAES metric is a viable alternative to traditional scoring metrics. Since seizure events can vary significantly in duration, each event should be weighed equally regardless of its duration. However, we also need to assess the accuracy of the time alignments. The TAES metric provides a nice framework for balancing these competing needs. Incorporating multiple metrics gives us better insight into a model's behavior and performance.

Acknowledgments

Research reported in this publication was most recently supported by the National Science Foundation Partnership for Innovation award number IIP-1827565 and the Pennsylvania Commonwealth Universal Research Enhancement Program (PA CURE). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the official views of any of these organizations.

References

[1] T. Yamada and E. Meng, *Practical guide for clinical neurophysiologic testing: EEG*. Philadelphia, Pennsylvania, USA: Lippincott Williams & Wilkins, 2009.

[2] I. Obeid and J. Picone, “The Temple University Hospital EEG Data Corpus,” *Front. Neurosci. Sect. Neural Technol.*, vol. 10, p. 196, 2016.

[3] I. Obeid and J. Picone, “Machine Learning Approaches to Automatic Interpretation of EEGs,” in *Biomedical Signal Processing in Big Data*, 1st ed., E. Sejdik and T. Falk, Eds. Boca Raton, Florida, USA: CRC Press, 2017 (in press).

[4] V. Shah, M. Golmohammadi, I. Obeid, and J. Picone, “Objective Evaluation Metrics for Automatic Classification of EEG Events,” 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 City, New York, USA: Springer, 2021, pp. 1–26.

[5] 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.

[6] Multimodal Information Group, “Open Speech Analytic Technologies (OpenSAT) Evaluation Series,” *Information Access Division*, National Institute of Standards and Technology, 2020. [Online]. Available: *https://www.nist.gov/itl/iad/mig/opensat*. [Accessed: 01-Jul-2020].

[7] “Competitions,” 2020. [Online]. Available: *https://www.kaggle.com/competitions*. [Accessed: 01-Jul-2020].

[8] I. Kiral *et al.*, “The Deep Learning Epilepsy Detection Challenge: Design, Implementation, and Test of a New Crowd-Sourced AI Challenge Ecosystem,” in *Challenges in Machine Learning Competitions for All (CiML)*, 2019, pp. 1–3.

[9] Y. Roy, R. Iskander, and J. Picone, “The Neureka® 2020 Epilepsy Challenge,” *NeuroTechX*, 2020. [Online]. Available: *https://neureka-challenge.com/*. [Accessed: 01-Jul-2020].

[10] V. Shah *et al.*, “The Temple University Hospital Seizure Detection Corpus,” *Front. Neuroinform.*, vol. 12, pp. 1–6, 2018.

[11] V. Shah and J. Picone, “NEDC Eval EEG: A Comprehensive Scoring Package for Sequential Decoding of Multichannel Signals,” *The TUH EEG Project Web Site*, 2019. [Online]. Available: *https://www.isip.piconepress.com/projects/tuh_eeg/downloads/nedc_eval_eeg/*. [Accessed: 01-Jul-2020].

[12] Y. Roy, R. Iskander, and J. Picone, “The Neureka® 2020 Epilepsy Challenge: Results,” *NeuroTechX*, 2020. [Online]. Available: *https://neureka-challenge.com/results*. [Accessed: 01-Jul-2020].

[13] 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.

[14] L. Wei and C. Mooney, “An Automatic Seizure Detection Method for Clinical EEG data,” in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2020, p. 6.

[15] Y. Yang, N. D. Truong, C. Maher, A. Nikpour, and O. Kavehei, “Two-Channel Epileptic Seizure Detection with Blended Multi-Time Segments Electroencephalography (EEG) Spectrogram,” in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2020, p. 6.

[16] T. Anand, M. G. Kumar, M. Sur, R. Aghoram, and H. Murphy, “Seizure Detection Using Time Delay Neural Networks and LSTM,” in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2020, p. 6.

[17] V. Shah, “Improved Segmentation for Automated Seizure Detection Using Channel-Dependent Posteriors,” Temple University, 2020.

[18] C. Chatzichristos *et al.*, “Epileptic Seizure Detection in EEG via Fusion of Multi-View Attention-Gated U-net Deep Neural Networks,” in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 2020, p. 7.

[19] 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.