34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Improving seizure detection performance reporting: analysing the duration needed for a detection

Lojini Logesparan, Alexander J. Casson and Esther Rodriguez-Villegas

Abstract— **Improving seizure detection performance relies not only on novel signal processing approaches but also on new accurate, reliable and comparable performance reporting to give researchers better and fairer tools for understanding the true algorithm operation. This paper investigates the sensitivity of current performance metrics to the duration of data that must be marked as candidate seizure activity before a seizure detection is made. The results demonstrate that not all metrics are insensitive to this high level choice in the algorithm design, and provide new approaches for comparing between reported algorithm performances in a robust and reliable manner.**

I. INTRODUCTION

Epilepsy is a serious neurological disorder, characterised by recurrent debilitating seizures, that affects 50 million people worldwide [1]. To aid in diagnosis and treatment there has long been an interest in the development of automated seizure detection algorithms [2]–[4]. However, obtaining completely accurate detection is very challenging and is still an active research topic. Based upon the EEG (electroencephalogram), algorithms use signal processing to *emphasise* periods of seizure activity and then *classification* to make an automated detection. Historically, the focus for algorithm improvement has been the investigation of different techniques for the emphasis and classification stages. Many algorithms have been reported, however the wide variety of performance metrics and methodologies used has made fair comparisons between different techniques extremely problematic.

There is therefore now an increasing interest in re-visiting how algorithm performance is assessed to give researchers better and fairer tools for understanding the true algorithm operation [5]–[7]. Recent studies have investigated a number of test methodology factors and their effect on the reported performance:

- The inter-patient variation [3].
- The record duration and number of events [6].
- The time collar around each detection [5].
- The imbalance in the test data (where there is significantly more non-seizure data than seizure data) [7].

Without correction, factors such as these can have a substantial impact on the reported performance level even though the same algorithm may be used in all cases.

In this paper, the impact of another critical factor on the reported seizure detection performance: how long a candidate detection needs to last before an actual detection is made, is investigated. Metrics that can be used to accurately compare algorithms must be insensitive to this high level choice in the algorithm design, but our results demonstrate that not all metrics satisfy this. Our results provide a starting point for allowing comparisons in performance between algorithms that make different choices for this factor.

Section II describes the performance metrics used, giving particular attention to their applicability for *specific* seizure detection problems. *Seizure detection* has many variants including: seizure occurrence detection [2], seizure onset detection [3], seizure termination detection [8], or seizure recording/data selection [9], and all future robust metrics must be closely linked to the specific variant. Section III then uses the fixed output of a seizure detection algorithm, changing how the metrics are calculated, to investigate the impact of the required detection duration. Finally the results are discussed and conclusions drawn in Section IV.

II. PERFORMANCE METRICS

The metrics used in this work are defined below and are split into two categories: *performance* and *cost*. These metrics are defined assuming that the seizure detection algorithm analyses non-overlapping epochs of EEG data such that each epoch can be labelled as either: a true positive detection (TP); a false positive detection (FP); a true negative decision (TN); or a false negative decision (FN).

A. *Performance: Epoch-sensitivity*

The percentage of seizure epochs correctly detected:

$$\text{Epoch-sensitivity} = \frac{1}{M}\sum_{i=1}^{M}\frac{TP}{TP+FN}\times 100\% \qquad (1)$$

where M is the number of EEG records containing seizures and i is the record number. This quantifies the percentage of the total seizure duration that has been correctly detected but it does not indicate *how many* different seizures have been detected. It has also been called recall [5], integral-overlap [10], or more generally sensitivity [9].

Applicability: This is an essential metric for seizure recording/data selection where only short sections of *interesting* EEG are recorded. High epoch-sensitivities show that the interesting data sections have been successfully selected. For seizure occurrence detection to assist in the offline review of data, the metric is less pertinent. If 10 s of EEG are displayed at a time during review, data before and after the seizure

The authors are with the Department of Electrical and Electronic Engineering, Imperial College London, UK. Email: {lojini.logesparan04, acasson, e.rodriguez} @imperial.ac.uk.

The research leading to these results has received funding from the European Research Council under the European Community's 7th Framework Programme (FP7/2007-2013) / ERC grant agreement no. 239749.

marker is naturally displayed to the neurologist, regardless of whether the algorithm identifies all of the data as seizure.

B. *Performance: Event-sensitivity*

This is the percentage of seizure events that are correctly detected, and has also been called average percentage seizures detected [2], good detection rate [5], any-overlap [10], and sensitivity. To achieve 100% event-sensitivity only a single detection in every seizure is required, whereas for 100% epoch-sensitivity all epochs in all seizures must be detected. It is thus generally possible to have better appearing results when considering only the event-sensitivity.

Applicability: This is an essential metric for all variants of seizure detection, showing how many seizures are detected.

C. *Cost: Specificity*

The percentage of non-seizure epochs correctly identified as non-seizure:

$$\text{Specificity} = \frac{1}{M}\sum_{i=1}^{M}\frac{TN}{TN+FP} \times 100\%. \qquad (2)$$

Specificity is a common cost metric although it can be weighted by imbalanced datasets: if most of the analysed data is non-seizure, TN can be large giving a high specificity, even if the false positive rate is impractically high.

Applicability: Specificity is an appropriate metric for both data selection and seizure occurrence detection as it is a measure of the percentage of non-seizure data (100%−specificity) that the neurologist will unnecessarily review.

D. *Cost: False positive rate & Duration under false positive*

False positive rate is the number of epochs incorrectly detected as seizure epochs, normalized by non-seizure duration. [5] suggested modifying this to be the total time duration of false positive epochs per hour to account for different methods of grouping closely spaced false positives. If no grouping of false detections is done, with non-overlapping epochs the two metrics are directly proportional.

Applicability: False positive rate is most applicable to algorithms that raise an alarm for intervention. For non-overlapping epochs, duration under false positive is mathematically equal to (100%−specificity) × 3600 (s).

E. *Cost: Precision*

The fraction of all detections that are correct:

$$\text{Precision} = \frac{1}{M}\sum_{i=1}^{M}\frac{TP}{TP+FP} \times 100\% \qquad (3)$$

also known as selectivity [11]. Precision overcomes the imbalance issue of specificity, but can be weighted if records with no seizure events are analysed. In such records, TP (and precision) will be zero, reducing the reported average precision across M records. Additionally, algorithms tested with little non-seizure data will inevitably have high precision as the number of false positives possible will be low.

Applicability: The precision is particularly pertinent for applications that require a sampled EEG review by the neurologist, such as a review of seizure frequency.

III. DETECTION DURATION IMPACT

Given the well defined and comprehensive analysis framework from Section II, it is now possible to investigate the impact of how much of a seizure needs to be identified as seizure activity before an automated detection is made. Previous work on seizure detection has used differing sizes of non-overlapping analysis epochs, typically from 1 s to 20 s [4], [9] and metrics that can accurately compare algorithms must be insensitive to this high level choice in the algorithm design.

To investigate this, the seizure detection algorithm described in the Appendix has been evaluated on 18-channel EEG data using short 1 s epochs. The output of this is then post-processed so that 1 s, 2 s, 5 s, 10 s and 20 s sections of data must be continuously marked as seizure activity before a detection is made. The impact of this duration on the performance metrics from Section II is then plotted. The algorithm has been tested on publicly available scalp EEG data [12], [13] obtained from 22 paediatric patients. This imbalanced data set contains 635 records (>916 hours) with 102 records containing a single expert marked seizure (total duration 7817 s). The results from the analysis of this data are given below and discussed in Section IV.

A. *Impact on performance metrics*

The detection algorithm is simulated at a particular decision threshold (see Appendix) to generate a pair of epoch-sensitivity and event-sensitivity values. Fig. 1 plots these pairs as 1 s, 2 s, 5 s, 10 s and 20 s of data needs to be identified as seizure in order to make an overall detection. Different decision thresholds are also used to illustrate the full range of sensitivity values. As expected, event-sensitivity consistently reports better appearing results than epoch-sensitivity. The difference between the two is small when both sensitivities are very high or very low, and the effect is reduced as larger durations are required for detection. Inspection of the changes shows that epoch-sensitivity is approximately constant with detection duration; the differences arise due to substantial changes in the event-sensitivity. Thus epoch-

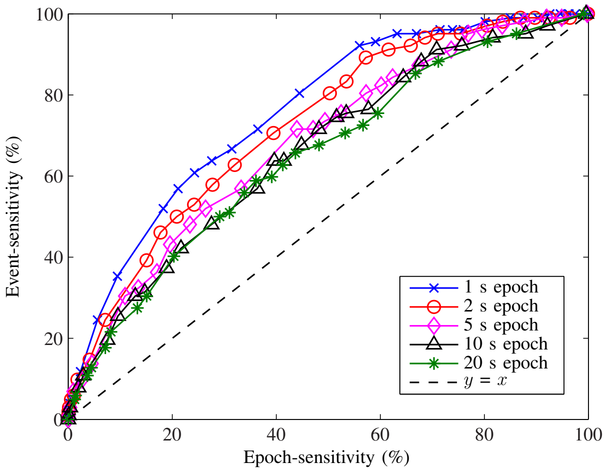

Fig. 1. Epoch-sensitivity and event-sensitivity, for the algorithm in the Appendix at different thresholds, as the required detection duration is varied.

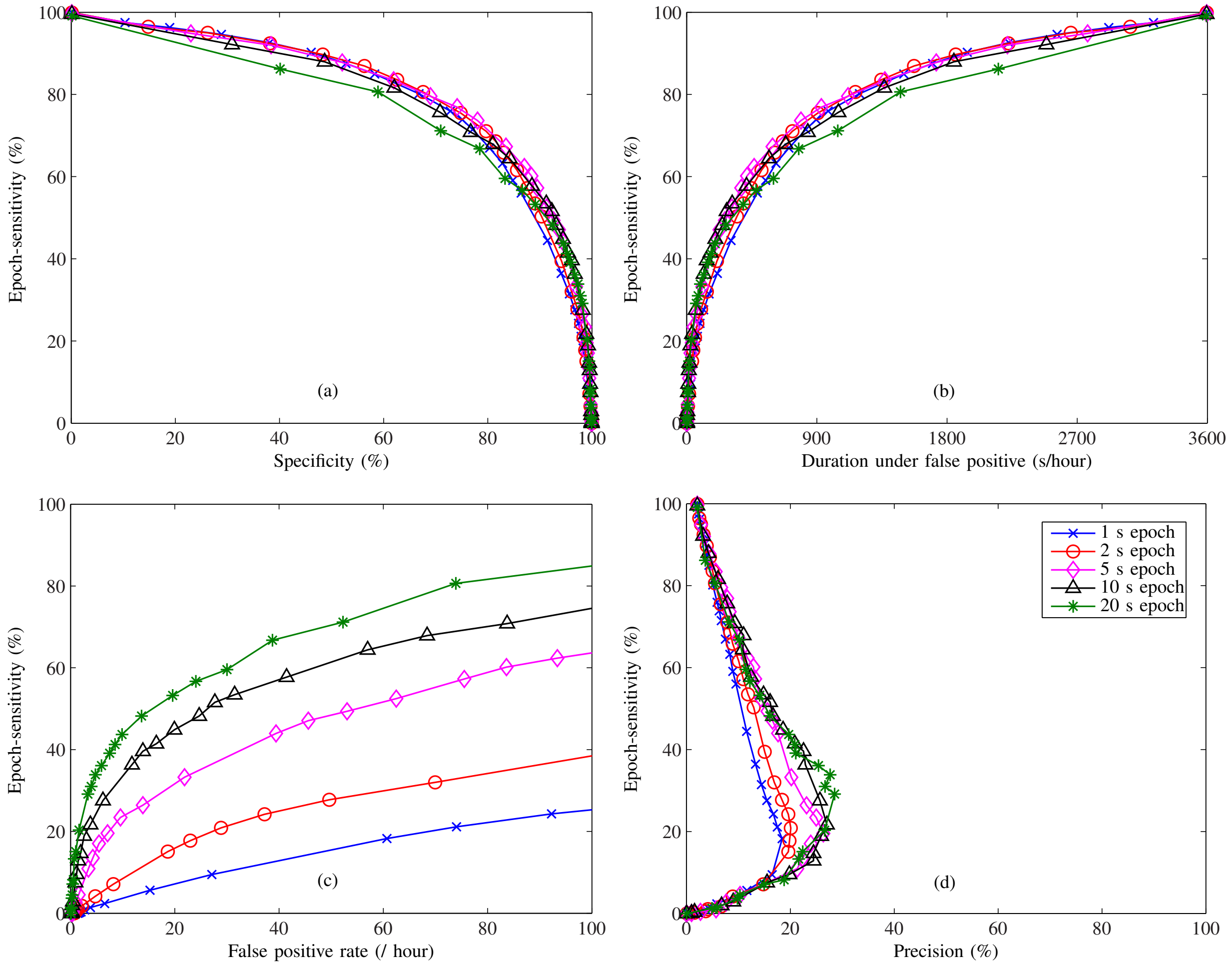

Fig. 2. Variation in *cost* metrics with required detection duration. (a) Specificity. (b) Duration under false positive. (c) False positive rate. (d) Precision.

sensitivity is selected to investigate the change in results reported by each of the four cost metrics.

B. Impact on cost metrics

Fig. 2 demonstrates the variance of the four *cost* metrics with the required detection duration. The metrics are plotted against epoch-sensitivity; results against event-sensitivity can be found by comparing with Fig. 1.

Both specificity (Fig. 2(a)) and duration under false positive (Fig. 2(b)) are very robust, showing little variation as the required detection duration changes. Whilst the duration under false positive remains insensitive, the same is not true for false positive rate (Fig. 2(c)). As the short 1 s analysis epochs are grouped into larger epochs for decision making, the absolute number of false positives must inevitably reduce as the number of time instances at which a false positive can be detected is reduced. This effect also manifests in the precision (Fig. 2(d)). At low sensitivities the number of true positives (TP) is low, and the absolute number of false positives (FP) has an appreciable effect on the calculation ($TP/(TP + FP)$). At higher sensitivities the precision is more independent of the detection duration.

IV. DISCUSSION AND CONCLUSIONS

Comparison of the epoch-sensitivity and event-sensitivity (Fig. 1) shows that the event-sensitivity consistently reports higher values for the same algorithm output. It is also more affected by changes in the required detection duration. This is reinforced by the results of Fig. 2 where plotting cost against epoch-sensitivity shows little vertical change in epoch-sensitivity, all shifts are in the horizontal cost direction. Epoch-sensitivity can therefore be robustly compared between different algorithms. Algorithms that report only event-sensitivity may not be directly compared if they use different detection durations.

Similarly, from Fig. 2 specificity and duration under false positive are insensitive to changes in detection duration. As both the number of true negatives (TN) and false positives (FP) are inversely proportional to detection duration, the specificity calculation ($TN/(TN + FP)$) is independent. Some small deviations in the calculated values are present in Fig. 2, and these are due to the change in the absolute number of TN and FP present in each case, affecting the mathematical accuracy of the specificity calculation. As a result, algorithms that use specificity or duration under false positive as their cost metric can be accurately directly compared. Ones using the precision metric cannot be.

However, we agree with [5] that precision is an important metric to report for future algorithms. Hence, weighting of the precision by the detection duration and amount of non-seizure data analysed should be accounted for and controlled if possible. Likewise, we believe that reporting the duration under false positive should be preferred to reporting the raw false positive rate. Furthermore, comparing the specificity performance curve (Fig. 2(a)) with the precision performance curve (Fig. 2(d)) reveals very different pictures of underlying algorithm performance. Depending on the pertinence of the metrics for the specific seizure detection application, as discussed in Section II, this can have a substantial impact on the applicable algorithm performance.

Overall, for ease of comparison between the performance of different seizure detection algorithms it is necessary to use metrics that are independent of high-level algorithm design choices which would be tailored to the needs of the specific detection application. Section II overviewed six metrics, giving particular attention to their utility in different specific applications. These have then been assessed for their independence to the amount of data needed to be detected in order to make a seizure detection. Section III demonstrated the impact of this, and our results allow more accurate and reliable comparisons between reported algorithm performances. Inevitably there will be other important factors affecting truly fair comparisons, such as record duration, the number of events in the database, and patient-dependence. An effort to combine these and develop a conclusive framework for performance evaluation of seizure detection algorithms will offer significant improvements to future seizure detection algorithms that goes beyond only attempting to develop new signal processing techniques.

Appendix

The seizure detection algorithm used in this study is shown in Fig. 3. Initially the single channel input EEG data $y(k)$ where k is the sample number, is filtered using a first order high-pass filter with a cut-off frequency of 0.16 Hz. Then $y(k)$ is split into non-overlapping epochs of 1 s duration to calculate the line length feature [3]:

$$L(x) = \sum_{k=2}^{N} |y(k-1) - y(k)| \tag{4}$$

where N is the total number of samples within epoch x. To estimate line length during non-seizure data, a median decaying memory (referred to as $BG(x)$) is calculated over the past 240 s of the feature:

$$\begin{aligned} BG(x) =& (1-\lambda)\text{median}(L(x-1)\cdots L(x-240)) \\ &+ \lambda BG(x-1) \end{aligned} \tag{5}$$

where the decay constant $\lambda = 0.99923$ [14]. In the first 240 s of data the background is calculated by taking the median of $(L(x-1)\cdots L(1))$. For the first epoch, $x = 1$, $BG(1) = L(1)$. The normalized line length $N(x) = L(x)/BG(x)$ is then compared to a pre-selected fixed threshold β. If $N(x)$ exceeds β the current epoch is classified as a candidate seizure epoch, otherwise it is marked as non-seizure data. If an epoch is classified as a seizure in one channel, then epochs from all channels at that time are marked as seizure.

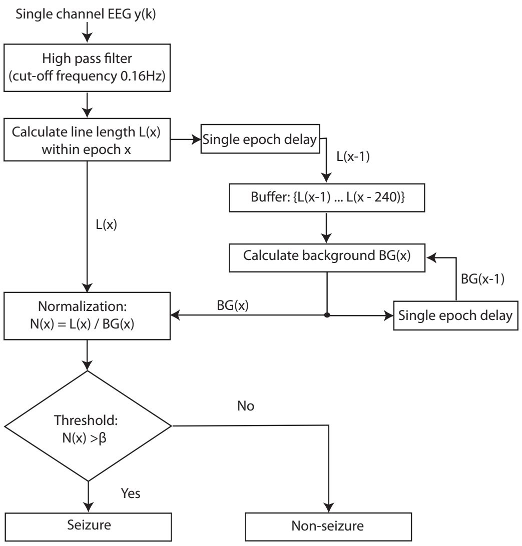
Single channel EEG y(k)
High pass filter (cut-off frequency 0.16Hz)
Calculate line length L(x) within epoch x
Single epoch delay
L(x-1)
Buffer: {L(x-1) ... L(x - 240)}
L(x)
Calculate background BG(x)
BG(x-1)
Normalization: N(x) = L(x) / BG(x)
BG(x)
Single epoch delay
No
Threshold: N(x) >β
Yes
Seizure
Non-seizure

Fig. 3. Flowchart of the seizure detection algorithm.

References

[1] S. Sisodiya, "Etiology and management of refractory epilepsies," *Nat. Clin. Pract. Neurol.*, vol. 3, no. 6, pp. 320–330, 2007.

[2] J. Gotman, D. Flanagan, and B. Rosenblatt, "Automatic seizure detection in the newborn: methods and initial evaluation," *Electroen. Clin. Neurophysiol.*, vol. 103, no. 3, pp. 356–362, 1997.

[3] R. Esteller, J. Echauz, T. Tcheng, B. Litt, and B. Pless, "Line length: an efficient feature for seizure onset detection," in *IEEE EMBC*, 2001.

[4] P. E. McSharry, T. He, L. A. Smith, and L. Tarassenko, "Linear and non-linear methods for automatic seizure detection in scalp electro-encephalogram recordings," *Med. Biol. Eng. Comput.*, vol. 40, no. 4, pp. 447–461, 2002.

[5] A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. Boylan, "Performance assessment for EEG-based neonatal seizure detectors," *Clin. Neurophysiol.*, vol. 122, no. 3, pp. 474–482, 2011.

[6] A. J. Casson, E. Luna, and E. Rodriguez-Villegas, "Performance metrics for the accurate characterisation of interictal spike detection algorithms," *J. Neurosci. Methods*, vol. 177, no. 2, pp. 479–487, 2009.

[7] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.

[8] A. Shoeb, A. Kharbouch, J. Soegaard, S. Schachter, and J. Guttag, "A machine-learning algorithm for detecting seizure termination in scalp EEG," *Epilepsy Behav.*, vol. 22, Suppl. 1, pp. S36–S43, 2011.

[9] B. R. Greene, S. Faul, W. P. Marnane, G. Lightbody, I. Korotchikova, and G. B. Boylan, "A comparison of quantitative EEG features for neonatal seizure detection," *Clin. Neurophysiol.*, vol. 119, no. 6, pp. 1248–1261, 2008.

[10] S. B. Wilson, "A neural network method for automatic and incremental learning applied to patient-dependant seizure detection," *Clin. Neurophysiol.*, vol. 116, no. 8, pp. 1785–1795, 2005.

[11] H. S. Park, Y. H. Lee, N. G. Kim, D. S. Lee, and S. I. Kim, "Detection of epileptiform activities in the EEG using neural network and expert system," *Stud. Health Technol. Inform.*, vol. 52, no. 2, pp. 1255–1259, 1998.

[12] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[13] A. Shoeb, "Application of machine learning to epileptic seizure onset detection and treatment," Ph.D. dissertation, MIT, September 2009.

[14] L. Kuhlmann, A. Burkitt, M. Cook, K. Fuller, D. Grayden, L. Seiderer, and I. Mareels, "Seizure detection using seizure probability estimation: Comparison of features used to detect seizures," *Ann. Biomed. Eng.*, vol. 37, pp. 2129–2145, 2009.