KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Ansari A.H., Cherian P. J., Caicedo A., Jansen K., Dereymaeker A., De Wispelaere L., Dielman C., Vervisch J., Govaert P., De Vos M., Naulaers G., Van Huffel S. **Weighted performance metrics for automatic neonatal seizure detection using multi-scored EEG data** IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 4, Jul. 2018, pp. 1114-1123.
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	https://ieeexplore.ieee.org/abstract/document/8030998
Journal homepage	https://jbhi.embs.org/
Author contact	your email amirhossein.ansari@kuleuven.be your phone number *+ 32 (0)16 326148*
IR	https://lirias2.kuleuven.be/viewobject.html?cid=1&id=1658351

(article begins on next page)

Weighted performance metrics for automatic neonatal seizure detection using multi-scored EEG data

A. H. Ansari [1], P. J. Cherian [2,3], A. Caicedo [1], K. Jansen [4,5], A. Dereymaeker [4], L. De Wispelaere [6], C. Dielman [7], J. Vervisch [4], P. Govaert [6,7], M. De Vos [8], G. Naulaers [4], S. Van Huffel [1]

***Abstract*—In neonatal intensive care units, there is a need for around the clock monitoring of EEG, especially for recognizing seizures. An automated seizure detector with an acceptable performance can partly fill this need. In order to develop a detector, an extensive dataset labeled by experts is needed. However, accurately defining neonatal seizures on EEG is a challenge, especially when seizure discharges do not meet exact definitions of repetitiveness or evolution in amplitude and frequency. When several readers score seizures independently, disagreement can be high. Commonly used metrics such as good detection rate (GDR) and false alarm rate (FAR) derived from data scored by multiple raters have their limitations. Therefore, new metrics are needed to measure the performance with respect to the different labels. In this paper, instead of defining the labels by consensus or majority voting, popular metrics including GDR, FAR, positive predictive value, sensitivity, specificity, and selectivity are modified such that they can take different scores into account. To this end, 353 hours of EEG data containing seizures from 81 neonates were visually scored by a clinical neurophysiologist and then processed by an automated seizure detector. The scored seizures were mixed with false detections of an automated seizure detector and were relabeled by 3 independent EEG readers. Then, all labels were used in the proposed performance metrics and the result was compared with the majority voting technique and showed higher accuracy and robustness for the proposed metrics. Results were confirmed using a bootstrapping test.**

***Index Terms*— Automated Neonatal Seizure Detection, Multi-scored EEG database, Performance Measurement Metrics**

I. Introduction

Detection of seizures has a high priority in the neonatal intensive care unit (NICU) because of high susceptibility of newborn brain to seizure expression [1]. Occurrence of neonatal seizures usually indicates a serious underlying neurological dysfunction. Unlike epileptic seizures in adults, most neonatal seizures have subtle clinical manifestations or are subclinical and are not readily recognized by clinical observation alone [2], [3]. Although monitoring of the electroencephalogram (EEG) accompanied by video is considered to be the gold standard in diagnosing neonatal seizures [4], reading EEG signals is labor-intensive, expensive, and importantly needs special expertise which may not be available around the clock in some NICUs. Therefore, a reliable, automated neonatal seizure detector is extremely useful and is required in the NICUs. However, measuring the reliability of such methods is not always straightforward, particularly when use of anti-epileptic drugs (AEDs) suppresses the EEG signal and modifies typical seizure patterns. Moreover, the varying degree of expertise among raters and the nature of the database, frequency of occurrence of seizures, presence of artifacts, etc., affect the visual scoring and to account for these in the performance assessment is not easy.

In recent years, there has been an increasing amount of literature on automated neonatal seizure detection using different techniques of pattern recognition and classification [5]–[12]. However, very little research has been devoted to different approaches of measuring the performance of seizure detectors. Sensitivity, specificity, and selectivity were used to measure the performance in [13], one of the first studies in this field. False alarm rate per hour, which is more tangible than specificity for clinicians, is applied by Gotman *et al.* [14]. Navakatikyan clarified the difference between event-based and time-based metrics and explained different approaches for performance measurement [15]. The mean

[1] Department of Electrical Engineering (ESAT), STADIUS, KU Leuven, Belgium, and imec, Leuven, Belgium

[2] Section of Clinical Neurophysiology, Department of Neurology, Erasmus MC, University Medical Center, Rotterdam, The Netherlands

[3] Division of Neurology, Department of Medicine, McMaster University, Hamilton, Canada

[4] Department of Development and Regeneration, University Hospitals Leuven, Neonatal Intensive Care Unit, KU Leuven, Leuven, Belgium

[5] Department of Development and Regeneration, University Hospitals Leuven, Child Neurology, KU Leuven, Leuven, Belgium

[6] Section of Neonatology, Department of Pediatrics, Sophia Children's Hospital, Erasmus MC, University Medical Center Rotterdam, The Netherlands

[7] ZNA Koningin Paola Kinderziekenhuis, Antwerp, Belgium

[8] Institute of Biomedical Engineering, Department of Engineering, University of Oxford, Oxford, UK

A.H.A. and S.V. are supported by: Bijzonder Onderzoeksfonds KU Leuven (BOF): Center of Excellence (CoE) PFV/10/002 (OPTEC), SPARKLE – Sensor-based Platform for the Accurate and Remote monitoring of Kinematics Linked to E-health #: IDO-13-0358, The effect of perinatal stress on the later outcome in preterm babies #: C24/15/036, TARGID - Development of a novel diagnostic medical device to assess gastric motility #: C32-16-00364 ; Fonds voor Wetenschappelijk Onderzoek Vlaanderen (FWO) projects: G.0A5513N (Deep brain stimulation); Agentschap Innoveren & Ondernemen (VLAIO) projects: STW 150466 - OSA+, O&O HBC 2016 0184 eWatch; imec: Strategic Funding 2017, ICON- HBC.2016.0167 SeizeIT; Belgian Federal Science Policy Office: IUAP P7/19/ (DYSCO, 'Dynamical systems, control and optimization', 2012-2017); Belgian Foreign Affairs-Development Cooperation: VLIR UOS programs (2013-2019); European Union's Seventh Framework Programme (FP7/2007-2013): EU MC ITN TRANSACT 2012, #316679, The HIP Trial: #260777; ERASMUS +: NGDIVS 2016-1-SE01-KA203-022114; European Research Council (ERC) Advanced Grant, #339804 BIOTENSORS This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information. A.D. is supported by IWT PHD grant: TBM 110697-NeoGuard. A.C. is a postdoctoral fellow from the Fonds voor Wetenschappelijk Onderzoek Vlaanderen (FWO).

false detection duration was proposed by Temko *et al.* [16]. Mathieson *et al.* used Cohen's kappa between expert's scores and the output of an automated seizure detector [17]. In these reviewed studies, the ground truth was clearly defined by an expert or by consensus of different experts. However, Smit *et al.* employed 3 experts and showed that there is only moderate agreement between the raters, with a mean inter-observer agreement of 0.57. Subsequently, the final labels were reached by consensus of the raters. In that study, instead of considering two labels, 'seizure' and 'non-seizure', which is very typical, a third label called 'possible seizure', was used [18]. Cherian *et al.* also reported a fair agreement between used raters (kappa = 0.4). They called the third label 'dubious seizure' and showed a remarkable difference between the results when keeping or removing the dubious seizures in the test database [19]. This difference was also reported by Ansari *et al.* for several methods [12]. Since the only possible way to measure the reliability of a seizure detector is comparing its outputs with the labels scored by an expert EEG reader, disagreement of the experts leads to uncertainty of the ground truth and makes the measurement labeling ambiguous. Although building a consensus solves this problem, it is not always possible in practice and is not the best solution, because of possible effect of social group dynamics among the expert EEG readers [20]. Therefore, there is a need to introduce a new set of metrics taking the dubious seizures as well as agreement of the raters into account. To this end, in this paper, some important typically used metrics for performance measurement of neonatal seizure detection are extended.

This paper is organized as follows: Section II describes the database and scoring strategy. In section III, the details of the classical and proposed metrics are provided. Furthermore, a boot strapping test for measuring the robustness and accuracy of the metrics are described in this section. The results are reported in section IV. Discussion is given in section V and conclusions are drawn in section VI.

II. Database

Three hundred and fifty three hours of EEG-polygraphic recordings from 81 neonates (71 with seizure and 10 without seizure) recorded in the NICUs of Sophia Children's Hospital (part of the Erasmus University Medical Center Rotterdam, The Netherlands) (EMCR) and the NICU of the University Hospital of Leuven, Belgium (UZL) are used in this research. The polygraphic signals are composed of the electrocardiogram, electro-oculogram, chin or limb surface electromyogram, and abdominal respiratory movement signal. All the neonates of EMCR (48 neonates with seizures and 10 without seizures) had hypoxic ischemic encephalopathy (HIE) while the neonates of UZL (23 neonates with seizures) had different etiologies including 6 HIE, 5 metabolic, 5 stroke, 2 genetic, and 5 others. Except for excluding neonates with heart malformations, no other preselection was made on the data. No effort was made to remove low-quality signals, artifacts, or dubious seizures. For each neonate, a board-certified clinical neurophysiologist, experienced in neonatal EEG, selected a window of the recording in which at least one seizure was observed. The window length varied from 2h to 8h for different neonates and it resulted in 281h of recordings with seizures. In addition, 72h of recordings from 10 neonates without seizure, but with severe or moderately severe HIE, were also added to the database (in total 353h).The exact start-time and duration of each seizure were marked by the expert. Then, the automated neonatal seizure detector described in [21] was applied on the data and the false alarms were defined. All seizures scored by the clinical neurophysiologist, as well as the false alarms of the detector, were shuffled and relabeled by 3 independent EEG readers. As a result, 4 labels, 1 from the primary and 3 from the secondary raters, were assigned to each event, where event refers to a labeled seizure or a false/true detection. Note that the secondary raters could not change the start time or duration of the events. They only scored each event as "definite seizure", "dubious seizure", or "definite non-seizure". More details can be found in [22]. All private information of the neonates was removed prior to relabeling in each medical center and the study was approved by the ethical committee of the respective hospitals.

III. Methods

A. Classical metrics in seizure detection problems having clear ground truth

In order to measure the performance of a seizure detector, different metrics are used in the literature. In general, these metrics can be classified into two groups: event-based and epoch-based (time-based) metrics. In the former, each event equally influences the performance metrics regardless of its duration. Therefore, a very short detected seizure has the same effect on the performance as a very long one. However, in the latter, each event affects the performance metrics proportionally to its duration. Subsequently, if the epoch-based metrics are used to optimize a seizure detector, many short seizures may be sacrificed for a very long one, which may not be satisfactory in some applications. Accordingly, different researchers may use one of these groups, based on their need and problem definition. In the following, the three most typical metrics of each group (Event/Epoch-based) are described. Furthermore, to illustrate the terms and definitions, Fig. 1 shows an imaginary example of 3 scored seizures as well as 4 automated detections over 3 hours.

Good Detection Rate (event-based): Good detection rate (GDR), also known as event-based sensitivity, expresses the overall percentage of the detected seizures and is calculated by

$$GDR = \frac{S_{over}}{S_{tot}} \times 100 \quad (1)$$

where S_{over} is the number of labeled seizures having overlap, even very tiny, with at least one of the detections and S_{tot} is the total number of labeled seizures. For instance, in Fig. 1, 2 out of 3 seizures are truly detected and the GDR equals 67%.

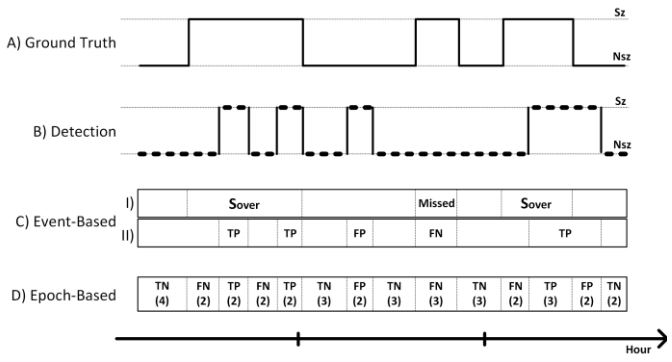

Fig. 1. An imaginary example of an event and epoch-based assessment. A) The ground truth including 3 scored seizures in 3 hours. B) The output of an automated detector when each line segment corresponds to an epoch. C) Event-based segments checking the overlap of the ground truth with the detections (I) and vice versa (II). D) Epoch-based segments and the number of involved epochs.

False alarm rate (event-based): False alarm rate (FAR) demonstrates the average number of false alarms that one can expect to see every hour and is defined as follows:

$$FAR = \frac{FP^{Event}}{T_{total}} \tag{2}$$

where FP^{Event} is the number of falsely detected events and T_{total} is the total duration of the recordings per hour. In Fig. 1, one false alarm in 3 hours are falsely detected and the FAR is 0.3 h^{-1}.

Positive Predictive Value (event-based): Positive predictive value (PPV) shows how reliable automated detections are and is measured using (3).

$$PPV = \frac{TP^{Event}}{TP^{Event} + FP^{Event}} \times 100 \tag{3}$$

where TP^{Event} shows the number of truly detected events which means the number of events having overlap with at least one of the labeled seizures. In Fig. 1, PPV equals 75% (3 out of 4). Note that TP^{Event} can be different from S_{over} used in GDR. The former shows the number of detections having overlap with labeled seizures, whereas the latter is the number of labeled seizures having overlap with the detections. In Fig. 1, TP^{Event} and S_{over} equal 3 and 2 respectively.

Sensitivity (epoch-based): Sensitivity, also called true positive rate, corresponds to GDR while the number of epochs is counted instead of the number of events. It is defined as

$$SEN = \frac{TP^{Epoch}}{TP^{Epoch} + FN^{Epoch}} \times 100 \tag{4}$$

where TP^{Epoch} and TN^{Epoch} are the number of truly detected (True Positive) and falsely not-detected (False Negative) epochs respectively. In Fig. 1, sensitivity is 44% (7 out of 16 epochs).

Specificity (epoch-based): Specificity, also known as true negative rate, measures the percentage of non-seizure epochs which are truly not identified by the seizure detector as follows:

$$SPE = \frac{TN^{Epoch}}{TN^{Epoch} + FP^{Epoch}} \times 100 \tag{5}$$

where TN^{Epoch} shows the number of truly not-detected epochs (True Negative) and FP^{Epoch} denotes the number of falsely detected epochs (False Positive). In Fig. 1, specificity is 79% (15 out of 19).

Selectivity (epoch-based): Selectivity, also called epoch-based PPV or precision, shows the reliability of the detections based on their duration and is measured by (6). In Fig. 1, selectivity equals 64% (7 out of 11).

$$SEL = \frac{TP^{Epoch}}{TP^{Epoch} + FP^{Epoch}} \times 100 \tag{6}$$

B. Multi-label problems

The aforementioned metrics are applicable when the ground truth is clearly predefined. In other words, the events should be clearly labeled as seizure or non-seizure by an expert beforehand. However, in neonatal seizure detection, it is possible that some repetitive patterns suspicious for seizures, called 'dubious seizures', appear without fulfilling all characteristics that define a seizure particularly after treatment with high dose of anti-epileptic drugs (AEDs) or in neonates with very severe encephalopathy. These seizures usually have suppressed amplitude, and lack of typical seizure evolution patterns. Consequently, it is possible that different experts have different opinions, resulting in several labels for each event. To illustrate, Fig. 2 shows a seizure having clear evolution of frequency and morphology on 'C4' and 'T4'. This seizure was scored as 'seizure' not only by the primary neurophysiologist, but also by all secondary raters. Fig .3 displays another event that lasted for 20s and scored by the primary and 2 of secondary raters as 'dubious seizure', while one of the rater believed that it is a 'seizure'. Finally, Fig. 4 shows an event that was primarily score as 'seizure', but the secondary raters completely disagreed with each other and scored it as 'seizure', 'dubious', and 'non-seizure'. In such cases, there are different ways to measure the performance, as explained below.

C. Averaging (ideal reference)

For measuring the performance of an algorithm in a multi-scored database, the simplest approach is measuring each performance metric multiple times based on each expert's labels independently and then average them. For example, the output of the algorithm is compared with each rater's scores and sensitivity is measured for each rater. Then, the average is reported as total sensitivity. Given the assumption that all used raters have enough experience and so their labels are reliable, this averaged metric can be the best approach since it exactly measures the average satisfaction if each of these experts uses such an algorithm in clinical practice. In this paper, the

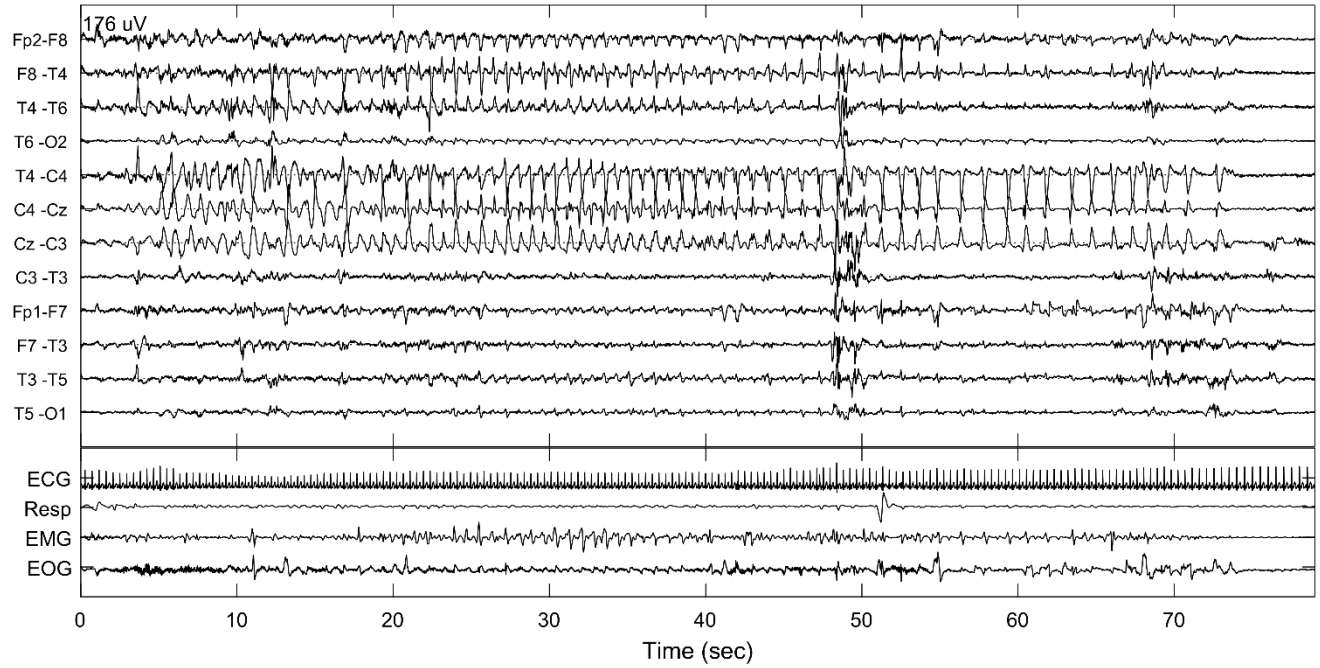

Fig. 2. An agreed seizure starting from sec 10 and lasting for 70s. This event was primarily scored as 'seizure' by the primary neurophysiologist with rhythmic oscillation and sharp spikes (mixed) at 'C4' and 'T4' (peak to peak amplitude of spikes is 200μv). All secondary raters also rescored it as 'seizure'.

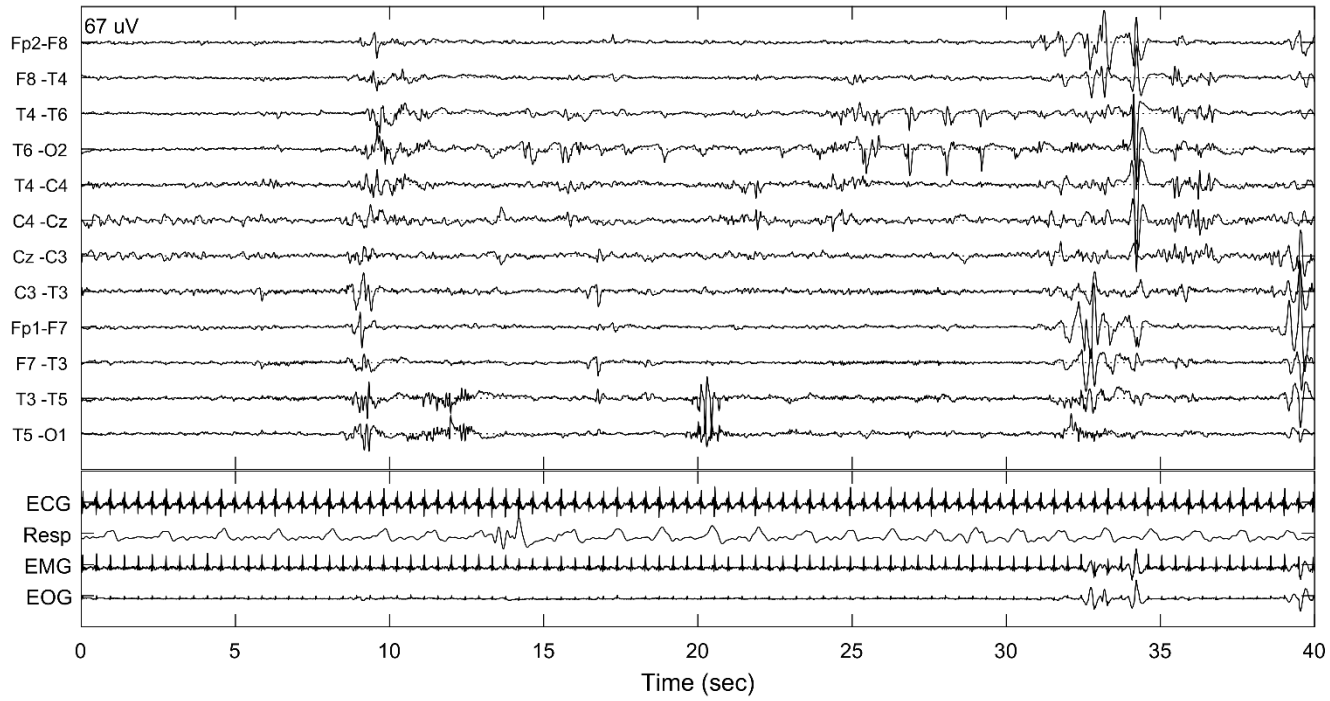

Fig. 3. A moderately agreed dubious seizure starting from sec 10 and lasting for 20s. This event was primarily scored as 'dubious seizure' by the primary neurophysiologist with arrhythmic oscillation and sharp spikes (mixed) at 'T6-O2' (peak to peak amplitude of spikes is 40-50μv). It was rescored as (Sz, Dsz, Dsz) by the secondary raters.

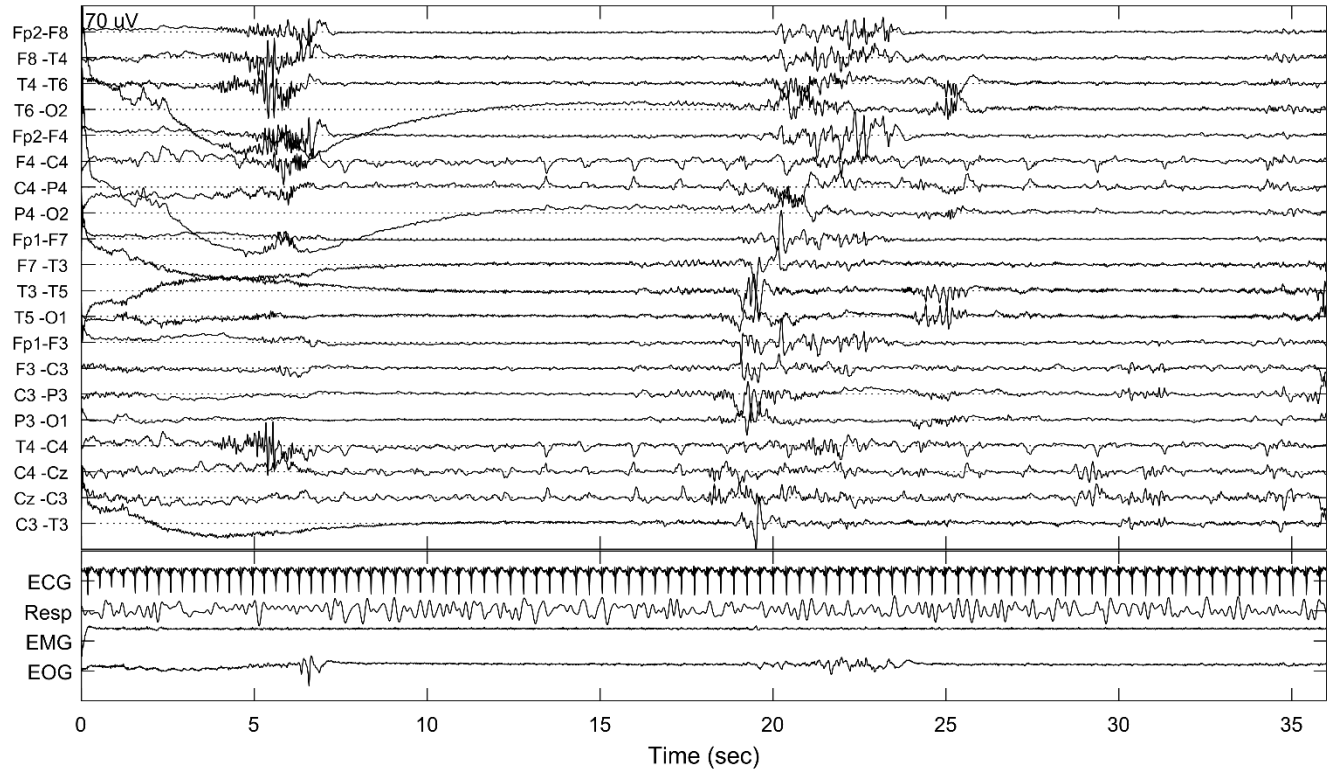

Fig. 4. Disagreed seizure starting from sec 10 and lasting for 25s. This event was primarily scored as 'seizure' by the primary neurophysiologist with rhythmic oscillation and sharp spikes (mixed) at 'C4' (peak to peak amplitude of spikes is 10-20μv). It was then rescored as (Sz, Dsz, Nsz) by the secondary raters.

averaged metrics are used as 'ideal reference' and other approaches are compared with them. However, in practice, recording EEG data and scoring seizures are very time-consuming and big datasets have gradually been made over years. Therefore, it is possible that several experts label different parts of the database. Similarly, multi-center databases may also have different scorers. It means that such averaged performance is not practically available and there is a need to introduce new performance metrics for multi-labeled databases.

Note that for measuring the performance based on each rater's scorings, the dubious events should be skipped. It means that if a segment is scored as dubious, neither it is counted as false alarm if a classifier detects it, nor it is calculated as missed seizure if a classifier cannot detect it [12], [19].

D. Majority Voting

In this approach, the most frequent label is defined as the aggregated label for each event individually. In this way, unlike the averaging method, different events may primarily have different number of labels. The main issue of this method is that the minority votes are not taken into account in any case. For instance, a seizure with 100% of votes have the same importance as a seizure with (50%+1) votes. However, it seems evident that seizures with a high degree of agreement are more important than the controversial ones and are expected to have more effect on the performance measurement.

E. Proposed method –weighted metrics

The key aspect is taking the minority votes into the measurement in order to quantify the overall satisfaction of the raters. Since there is not yet a clear "gold standard" for the interpretation of the EEG signals for defining seizures in the literature, using Fuzzy logic may help compensate for the inherent uncertainty of such problem.

Assuming the events are labeled as 1) definite seizure (Sz), 2) dubious seizure (Dsz), and 3) definite non-seizure (Nsz), three corresponding fuzzy sets are derived and subsequently each event has three membership values: μ_S, μ_D, and μ_N showing the membership of the event to each set. These values are defined based on the percentage of their corresponding labels. Thus, $\mu_S + \mu_D + \mu_N = 1$. For instance, μ_S, μ_D, and μ_N of an event which is labeled as [Sz, Sz, Sz, Dsz, Nsz] by 5 raters are respectively 0.6, 0.2, and 0.2. Since it is rational to say that detecting dubious events is neither true nor false, the defined μ_D is not used from here on and does not appear in any formula. The mentioned classical metrics are extended as follows:

Weighted good detection rate ($WGDR$) is measured by

$$WGDR = \frac{\sum_{n=1}^{N_{Scored}} \mu_S(n) O_{Scored}^{Event}(n)}{\sum_{n=1}^{N_{Scored}} \mu_S(n)} \times 100 \quad (7)$$

where N_{Scored} is the total number of labeled events and $O_{Scored}^{Event}(n)$ equals 1 if there is overlap between the nth labeled event and at least one of the automated detections, and otherwise, it equals 0.

Weighted false alarm rate ($WFAR$) is calculated by

$$WFAR = \frac{\sum_{n=1}^{N_{Detection}} \mu_N(n) O_{Det}^{Event}(n)}{T_{total}} \quad (8)$$

where $N_{Detection}$ is the total number of detected events and $O_{Det}^{Event}(n)$ equals 1 if there is overlap between the nth detected event and at least one of the labeled events, and otherwise, it is 0.

Weighted positive predictive value (WPPV) is defined as

$$WPPV = \frac{\sum_{n=1}^{N_{Detection}} \mu_S(n) O_{Det}^{Event}(n)}{\sum_{n=1}^{N_{Detection}} (\mu_S(n)+\mu_N(n)) O_{Det}^{Event}(n)} \times 100. \quad (9)$$

In the same way, *weighted sensitivity (WSEN)*, *weighted specificity (WSPE)*, and *weighted selectivity (WSEL)* are defined by the (10)–(12).

$$WSEN = \frac{\sum_{p=1}^{P} \left(\mu_S(p) O^{Epoch}(p)\right)}{\sum_{p=1}^{P} \mu_S(p)} \times 100 \quad (10)$$

$$WSPE = \frac{\sum_{p=1}^{P} \left(\mu_N(p) \left(1 - O^{Epoch}(p)\right)\right)}{\sum_{p=1}^{P} \mu_N(p)} \times 100 \quad (11)$$

$$WSEL = \frac{\sum_{p=1}^{P} \left(\mu_S(p) O^{Epoch}(p)\right)}{\sum_{p=1}^{P} \left((\mu_N(p)+\mu_S(p)) O^{Epoch}(p)\right)} \times 100 \quad (12)$$

P is the total number of epochs and $O^{Epoch}(p)$ equals 1 if the pth epoch is detected by the automated classifier, and otherwise, it equals 0.

Note that if $\mu_{N,S} \in \{0, 1\}$, which means only one rater scored the data or there is 100% agreement of the raters, the weighted and classical metrics are exactly equal. For instance, in the formula of $WSEL$, $\left(\mu_S(p) O^{Epoch}(p)\right)$ equals 0 for all values of p except when the pth epoch is detected by the algorithm and labeled as seizure by the rater. Therefore, its summation over all p equals the total number of true positives, the numerator of SEL, in (4). Thus, these proposed metrics can be seen as a generalization of the classical metrics for multi-rater approaches.

F. Bootstrapping test

In this work, in order to compare the accuracy and robustness of the classical metrics and proposed weighted metrics, a bootstrapping test was applied. Bootstrapping is a powerful analysis for measuring the characteristics of a variable from real observations. This method is useful in complicated situations when the characteristics of the real model generating the data are unknown. The main idea of bootstrap is mimicking the real generator behavior by repetitive selections of 'n-out of-m' samples from the real observations with replacement. It means that after bootstrapping, one vector with m observations will be converted to N vectors of n observations, where N is the number of repetitions. For each vector, the target variable is computed. It results in N variables in total from the N vectors. Then, the mean, median, variance, PDF, or other required statistical characteristics of the target variable are calculable [23]–[25].

In this paper, since the effect of the change in the raters' scorings is to be considered, the bootstrap is applied on the raters' opinions. It means that for each event, 4 labels were selected with replacement from the 4 existing labels. This selection is performed for the whole database and all metrics are computed using the bootstrapped labels and the fixed classifier output. This selection and computation is repeated 10000 times resulting in 10000 values for each metric. Then, a PDF is estimated for each metric by a kernel smoothing technique explained in [26] and the results are plotted and discussed.

G. Tested Seizure Detectors

In order to reduce the effect of the classifier on the comparison of the considered measurement metrics, three different previously published neonatal seizure detectors including a heuristic classifier [21], an improved version of the heuristic classifier [6], and a multi-stage detector [12] were used. These algorithms are described briefly in the following paragraphs.

I) Heuristic classifier: The goal of this classifier is mimicking a human expert EEG reader for detecting seizures. To this end, two parallel algorithms were proposed to detect seizures from each channel of EEG. In the first one, the spikes are firstly detected by splitting the EEG into 5s epochs and comparing the peaks of nonlinear energy with the background activity. Then, if sufficient number of sequential spikes have a high correlation together, the spikes are detected as spike train type seizure. In the second algorithm, first, the EEG is transformed to the δ (0.5-4 Hz) and θ (4-8 Hz) frequency bands using a discrete wavelet transform. Then, the potential activities are defined where the energy of the signal increases significantly compared to the background energy. Finally, if autocorrelation analysis shows a periodic activity in the detected potential activities, that part of the EEG is marked as oscillatory type seizure [21].

II) Improved heuristic method: This is an improved version of the aforementioned heuristic method which was described in [5] and in Appendix A of [6]. In this version, instead of splitting the EEG in small epochs for detecting the spikes, an adaptive thresholding method is proposed. Furthermore, the autocorrelation analysis of the oscillatory detector has been improved by extracting more distinctive features.

III) Multi-stage classifier: The main goal of this method described in [12], is using a machine learning technique to reduce the number of false alarms of the mentioned heuristic method. This algorithm is composed of two stages. In the first one, all potential seizures are detected by the heuristic method described in [21]. In the second one, each detection (i.e. the so-called potential seizure), was split into 8 sec epochs with 50% overlap and 50 features are extracted for each epoch. Then, a pre-trained support vector machine with probabilistic output is applied. The probabilities of different epochs of each potential detection are compared with a threshold and aggregated with a majority voting operator. As a result, some detections of the heuristic method are removed by the second stage as false detections. It has been shown that this post-processing stage can significantly improve the performance of the heuristic algorithm.

Note that all parameters or variables of these three used algorithms (such as trained-classifiers, the hyper-parameters, and final thresholds) were exactly defined based on their original papers and were not retrained or retuned for this paper. Furthermore, they were not subjected to any changes

during the performance measurement for different raters or different metrics.

IV. Results

A. Accuracy

As it is mentioned in section III-C, the 'ideal reference' of each metric is defined by averaging that metric over the raters, which shows the averaged raters' opinions. The accuracy of a substitute metric is corresponding to proximity of such metric to its ideal reference. Therefore, in order to compare the accuracy of the classical metrics using majority voting with the proposed metrics, first, all epoch/event-based metrics were measured based on each rater's labels individually. In this case, since only one rater is used for each metric, the classical and proposed metrics are equal. Then, the average of each metric was measured and used as an 'ideal reference'.

Second, the classical metrics using majority voting and proposed metrics are computed based on all raters' labels and compared with the ideal reference. The results are listed in Table I where the boldface items of the last two columns represent the values which are closer to the ideal reference (are more accurate).

Note that for measuring the epoch-based metrics, the epochs were selected as 1sec windows with no overlap. However, this length does not play a major role in the performance measurement.

The results shown in Table I indicate that the proposed metrics produce results which are, in general, more similar to the ideal reference (12 out of 18). Because of the fact that the ideal reference is not always practically accessible, as it was previously mentioned, the proposed metrics can be a more accurate substitute for the average performance assessment than classical metrics using majority voting.

B. Robustness

In order to compare the robustness of the two approaches for the assessment of the performance of the classifiers, majority voting and the proposed one, a realistic experiment and a bootstrapping test were designed. For the former, the secondary raters were asked to rescore 65% of the events for the second time. The interval between the scoring times was 1 year on average. Given that there are many dubious seizures in the used database, some events were labeled differently in the second trial (17% of the scorings, 847 out of 4980). For instance, the disagreed event presented in Fig. 4 was labeled as seizure by the 3rd secondary rater in the second trial, while it had been labeled as non-seizure primarily by this rater in the first trial. It is expected that a robust metric tolerates this change. To reveal the robustness of the metrics, the performance was computed using the new labels and it was compared with the results obtained with the initial labels. Table I lists the differences between the two trials for each metric, separately. The boldface values indicate the items with smaller change. As it is shown, in 16 cases out of 18, the proposed metrics have smaller difference, which indicates that the proposed metrics are more robust to small change in the labels.

In practice, it is not possible to repeat the rescoring procedure by the same expert EEG readers for many times, not only because it is time-consuming and expensive, but also because of the memorizing effect. Therefore, a bootstrapping test is designed, as described in the previous section, and used to repeat the rescoring for statistical comparison. As a result, Fig. 5 shows the PDFs of all epoch/event-based metrics for the three algorithms separately. In this figure, it is shown that the PDFs of the proposed weighted metrics (dashed line) are generally more similar to the reference PDFs (continuous line) than the majority voting technique (dotted line). Furthermore, for most PDFs, the standard deviations of the proposed metrics are smaller than those of the majority voting, which confirm their higher robustness.

V. Discussion

The development of a classifier for neonatal seizure detection with adequate performance is undeniably needed in NICUs. However, measuring the performance of such a system is problematic since the gold standard for the interpretation of the EEG signals is not yet clearly defined. In other words, several expert clinical neurophysiologists may score a database differently, which results in different performances for one algorithm. This problem is exacerbated when dubious seizures are present in a database containing recordings from neonates with severe neurological problems, such as severe HIE. The best and simplest solution for such ambiguity of ground truth is averaging the performance over all the expert raters. This averaged performance is the ideal reference since it expresses the averaged satisfaction of the raters if they use that classifier individually at their centers. However, in practice, the scores of all raters are not always available for all events, especially in large and multi-center datasets. For instance, the used database in this study was collected over 10 years in two centers and there is no guarantee that all present experts who scored it are available in the future. (Note that in this paper, all events were scored by the primary and secondary raters to be able to compare the performance of the proposed metrics with the ideal ones, although it is not required for the proposed metrics as mentioned.) Furthermore, for training some seizure detection algorithms, the label of each event should be explicitly defined. In that case, a common approach is using majority voting to aggregate different labels of each event. However, since majority voting does not use the minority votes,

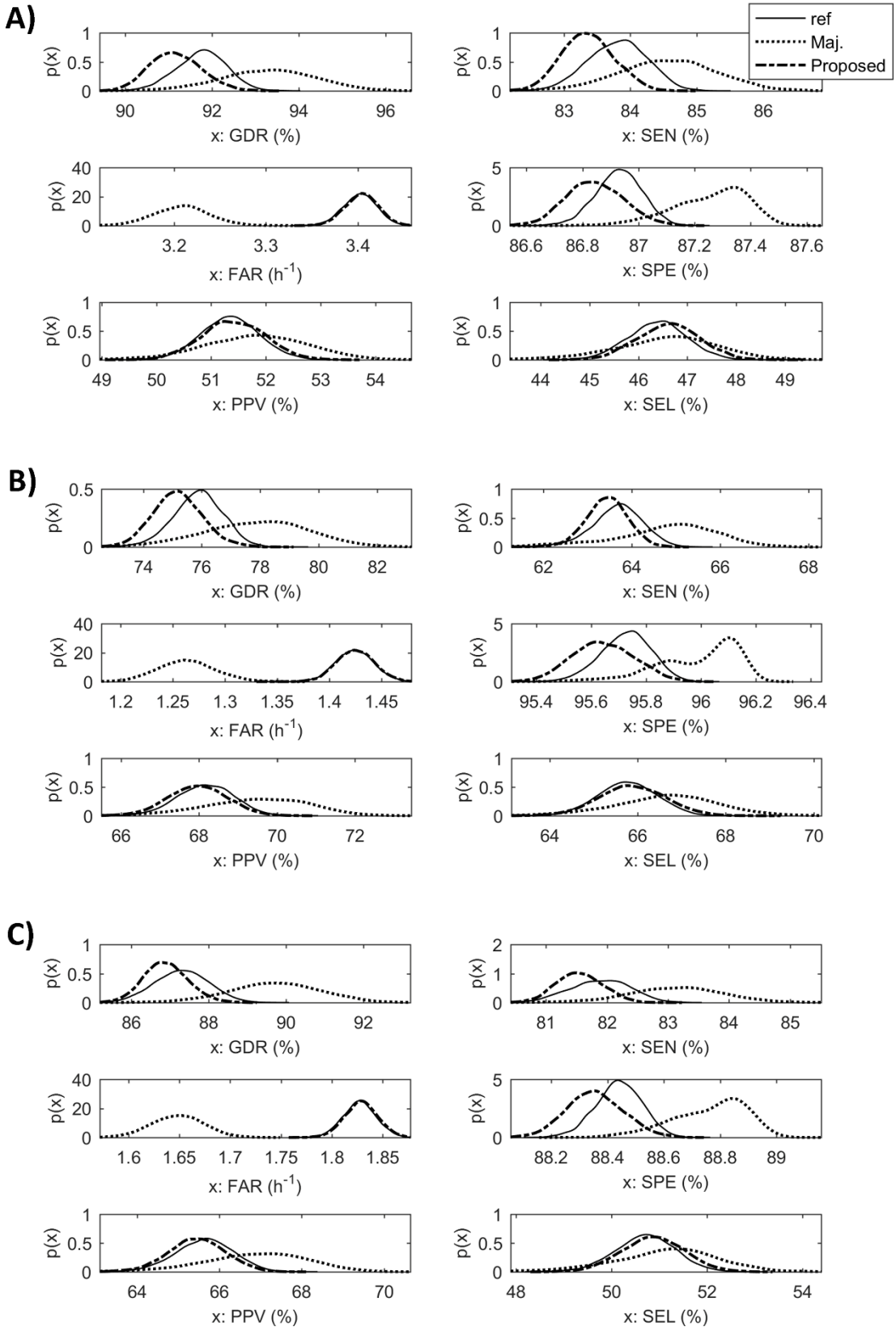

Fig. 5. Probability density function of GDR, FAR, PPV, SEN, SPE, and SEL when the detector is A) the heuristic algorithm proposed in [21], B) the improved heuristic algorithm proposed in [5], and C) the multi-stage classifier proposed in [12]. In all traces, continuous, dotted, and dashed lines are respectively corresponding to the PDFs of the reference, classical metrics using majority voting, and proposed weighted metrics

it does not measure the overall performance and seems to be different from the defined ideal reference. Therefore, a new framework taking all raters' opinions into account was needed for measuring the performance of neonatal seizure detectors. To this end, in this paper, a new set of performance measurement metrics was proposed and compared with the majority voting approach.

The main feature of the proposed weighted metrics is the fact that it takes minority votes into account besides the majority votes. The events with larger agreement in the votes have higher influence on the proposed metrics while in majority voting there is no difference between two events having 100% and (50%+1) of votes. Therefore, the overall performance measured by the newly proposed metrics is closer to the ideal reference than majority voting, as it has been practically confirmed by the results provided in this manuscript. In addition, the dubious seizures may be scored differently when a rater scores them several times. Moreover, these dubious seizures are usually controversial and have less agreement. Hence, it is probable that whenever a dubious seizure is rescored, the majority vote changes which results in a decrease of the robustness of the majority voting technique. Nevertheless, since minority votes also play a role in the proposed metrics, changing a label has a small effect on the metrics, thereby increasing the robustness besides the accuracy.

The other advantage of the proposed metrics, compared to the majority voting approach, is that when the raters have different levels of expertise, they can influence the metrics differently. To this end, each rater can have a specific weight which corresponds to his/her experience, knowledge, agreement with others, etc. This weight can be easily used in the calculation of the μ_x using weighted averaging. For instance, the membership value of being seizure (μ_S) of an event scored as (S_z, S_z, N_{Sz}) by 3 raters having (1, 0.75, 0.5) weights is 0.78 ($= \frac{1+0.75}{1+0.75+0.5}$). As a result, the labels of more experienced or preferred raters have more impact on all the modified performance measurement metrics than those of the other raters. In this paper, for the ideal reference, it has been assumed that the raters have the same experience, which rarely happens in real practice. In that case, the mentioned weighted averaging technique can also be used for the ideal metrics. Note that for this solution, the levels of expertise should be quantified which is not always straightforward.

Moreover, it seems logical that detecting the seizures which are agreed by different experts is more important than controversial events. Hence, if we can employ the scores of raters in developing or training the classifiers in order to prefer the highly agreed events to the poorly ones, it may increase the overall performance. Another important finding of the proposed technique is that each event has three membership values expressing their agreement. If the training method accepts weights for the training data, e.g. weighted least squares support vector machines [27], the μ_S which corresponds to the agreement of being seizure is a good option to be used as the weight. Furthermore, if the classifier supports rewarding and penalizing strategies, like reinforcement learning, μ_{Ns} can be used as a penalty to correct the model.

TABLE I

SIMILARITY OF REFERENCE WITH PROPOSED METRICS AND CLASSICAL METRICS USING MAJORITY VOTING

Detector		Metric	Ideal Reference [b]	Classical [c]	Proposed [d]
Heuristic Detector [21]	Event	[W]GDR (%)	92.1	95.0	**91.6**
		[W]FAR (h^{-1})	3.17	3.14	**3.17**
		[W]PPV (%)	53.7	59.2	**53.9**
	Epoch	[W]SEN (%)	77.5	**78.1**	74.3
		[W]SPE (%)	89.7	**90.5**	88.1
		[W]SEL (%)	50.4	58.7	**52.7**
Improved Heuristic Detector [5]	Event	[W]GDR (%)	73.0	**77.8**	67.1
		[W]FAR (h^{-1})	1.60	1.30	**1.60**
		[W]PPV (%)	62.2	70.5	**63.5**
	Epoch	[W]SEN (%)	55.2	**56.5**	48.6
		[W]SPE (%)	97.0	97.65	**96.9**
		[W]SEL (%)	63.5	74.8	**65.2**
Multi-stage Detector [a] [12]	Event	[W]GDR (%)	82.3	89.6	**82.2**
		[W]FAR (h^{-1})	1.47	1.31	**1.47**
		[W]PPV (%)	62.4	71.5	**64.8**
	Epoch	[W]SEN (%)	73.2	**76.0**	69.9
		[W]SPE (%)	91.1	**92.1**	89.6
		[W]SEL (%)	55.8	65.7	**58.2**

[a] The multistage classifier when its defined threshold (TH) equals 0.3
[b] Averaged metrics over all raters
[c] Classical metrics using Majority Voting technique
[d] Proposed weighted metrics

TABLE II
DIFFERENCE BETWEEN THE 1ST AND 2ND SCORING TRIALS

Detector		Metric	Δ Classical	Δ Proposed
Heuristic Detector [21]	Event	[W]GDR (%)	**0.47**	0.84
		[W]FAR (h^{-1})	**0.22**	0.25
		[W]PPV (%)	6.10	**1.79**
	Epoch	[W]SEN (%)	2.84	**0.30**
		[W]SPE (%)	2.41	**0.40**
		[W]SEL (%)	5.33	**1.88**
Improved Heuristic Detector [5]	Event	[W]GDR (%)	2.33	**0.11**
		[W]FAR (h^{-1})	0.65	**0.31**
		[W]PPV (%)	3.59	**0.95**
	Epoch	[W]SEN (%)	5.82	**0.58**
		[W]SPE (%)	1.49	**0.26**
		[W]SEL (%)	5.00	**0.78**
Multi-stage Detector [12]	Event	[W]GDR (%)	1.45	**0.20**
		[W]FAR (h^{-1})	0.19	**0.14**
		[W]PPV (%)	5.02	**3.04**
	Epoch	[W]SEN (%)	3.35	**0.04**
		[W]SPE (%)	2.39	**0.31**
		[W]SEL (%)	5.94	**2.81**

In addition, since the most alarming systems in hospitals only supported on/off inputs, it was previously expected that the output of seizure detectors should be either seizure or non-seizure, S_z/N_{Sz}. On the contrary, new generations of alarming systems, such as colorful lighting systems, can be used in NICUs and according to the probability of detected seizures, change the color/luminance of the light. Although some published methods produced probability as output, like [7], [11], [12], they still applied thresholding combined with classical performance metrics. Hence, they do not measure the actual performance of their probability-based systems. Nevertheless, in the proposed metrics, the defined $O^x(y)$ functions showing the overlap between the events and detections can easily be replaced by the probability. If this is the case, the metrics take the probability of being a seizure into account and measure the overall performance of probability-based outputs.

However, a disadvantage of the proposed metrics compared to majority voting is that whenever a rater is not experienced enough, his/her opinion always influences the

metrics as well as the experienced ones since in the proposed metrics the minority votes are always important. However, using majority voting can remove an outlier rater if present. Although using the aforementioned weights for the raters can partially solve this problem, finding the correct weights might be another challenge.

VI. Conclusion

The purpose of the current study was to determine a new framework for measuring the overall performance of automated neonatal seizure detectors in the presence of raters' disagreement. To this end, the typically used metrics were extended in order to take the agreement of raters on each scored/detected event into account and were tested on cEEG data from 81 neonates, recorded in 2 medical centers. A bootstrapping test was also applied on the data and supports the empirical results. The significant finding to emerge from these results is that the proposed metrics are more accurate and robust than majority voting because the minority votes also play a role in the metrics. The proposed metrics are generic and applicable for every pattern recognition 2-class problem in which the ground truth is differently defined by several experts. Furthermore, the new metrics automatically remove 'dubious'-scored events and allow having different raters for different parts of the database and for missing scores.

References

[1] J. J. Volpe, *Neurology of the Newborn*, 5th ed. Philadelphia: Saunder WB, 2008.

[2] A. M. E. Bye and D. Flanagan, "Spatial and Temporal Characteristics of Neonatal Seizures," *Epilepsia*, vol. 36, no. 10, pp. 1009–1016, Oct. 1995.

[3] D. M. Murray, G. B. Boylan, I. Ali, C. A. Ryan, B. P. Murphy, and S. Connolly, "Defining the gap between electrographic seizure burden, clinical expression and staff recognition of neonatal seizures," *Arch. Dis. Child. - Fetal Neonatal Ed.*, vol. 93, no. 3, pp. F187–F191, May 2008.

[4] J. M. Rennie, G. Chorley, G. B. Boylan, R. Pressler, Y. Nguyen, and R. Hooper, "Non-expert use of the cerebral function monitor for neonatal seizure detection," *Arch. Dis. Child.-Fetal Neonatal Ed.*, vol. 89, no. 1, pp. F37–F40, 2004.

[5] W. Deburchgraeve, "Development of an automated neonatal EEG seizure monitor," PhD Thesis, Katholieke Universiteit of Leuven, Leuven, 2010.

[6] M. De Vos *et al.*, "Automated artifact removal as preprocessing refines neonatal seizure detection," *Clin. Neurophysiol.*, vol. 122, no. 12, pp. 2345–2354, Dec. 2011.

[7] A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. Boylan, "EEG-based neonatal seizure detection with Support Vector Machines," *Clin. Neurophysiol.*, vol. 122, no. 3, pp. 464–473, Mar. 2011.

[8] N. J. Stevenson, J. M. O'Toole, L. J. Rankine, G. B. Boylan, and B. Boashash, "A nonparametric feature for neonatal EEG seizure detection based on a representation of pseudo-periodicity," *Med. Eng. Phys.*, vol. 34, no. 4, pp. 437–446, May 2012.

[9] E. M. Thomas, A. Temko, W. P. Marnane, G. B. Boylan, and G. Lightbody, "Discriminative and Generative Classification Techniques Applied to Automated Neonatal Seizure Detection," *IEEE J. Biomed. Health Inform.*, vol. 17, no. 2, pp. 297–304, Mar. 2013.

[10] J. G. Bogaarts, E. D. Gommer, D. M. W. Hilkman, V. H. J. M. van Kranen-Mastenbroek, and J. P. H. Reulen, "EEG Feature Pre-processing for Neonatal Epileptic Seizure Detection," *Ann. Biomed. Eng.*, vol. 42, no. 11, pp. 2360–2368, Aug. 2014.

[11] S. B. Nagaraj, N. J. Stevenson, W. P. Marnane, G. B. Boylan, and G. Lightbody, "Neonatal Seizure Detection Using Atomic Decomposition With a Novel Dictionary," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 11, pp. 2724–2732, Nov. 2014.

[12] A. H. Ansari *et al.*, "Improved multi-stage neonatal seizure detection using a heuristic classifier and a data-driven post-processor," *Clin. Neurophysiol.*, vol. 127, no. 9, pp. 3014–3024, Sep. 2016.

[13] A. Liu, J. S. Hahn, G. P. Heldt, and R. W. Coen, "Detection of neonatal seizures through computerized EEG analysis," *Electroencephalogr. Clin. Neurophysiol.*, vol. 82, no. 1, pp. 30–37, Jan. 1992.

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

[15] M. A. Navakatikyan, P. B. Colditz, C. J. Burke, T. E. Inder, J. Richmond, and C. E. Williams, "Seizure detection algorithm for neonates based on wave-sequence analysis," *Clin. Neurophysiol.*, vol. 117, no. 6, pp. 1190–1203, Jun. 2006.

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

[17] S. R. Mathieson *et al.*, "Validation of an automated seizure detection algorithm for term neonates," *Clin. Neurophysiol.*, vol. 127, no. 1, pp. 156–168, Jan. 2016.

[18] L. S. Smit, R. J. Vermeulen, W. P. F. Fetter, R. L. M. Strijers, and C. J. Stam, "Neonatal Seizure Monitoring Using Non-Linear EEG Analysis," *Neuropediatrics*, vol. 35, no. 6, pp. 329–335, Nov. 2004.

[19] P. J. Cherian *et al.*, "Validation of a new automated neonatal seizure detection system: A clinician's perspective," *Clin. Neurophysiol.*, vol. 122, no. 8, pp. 1490–1499, Aug. 2011.

[20] S. Vanhatalo, "Development of neonatal seizure detectors: An elusive target and stretching measuring tapes," *Clin. Neurophysiol.*, vol. 122, no. 3, pp. 435–437, Mar. 2011.

[21] W. Deburchgraeve *et al.*, "Automated neonatal seizure detection mimicking a human observer reading EEG," *Clin. Neurophysiol.*, vol. 119, no. 11, pp. 2447–2454, Nov. 2008.

[22] A. Dereymaeker *et al.*, "Interrater agreement in visual scoring of neonatal seizures based on majority voting on a web-based system: The Neoguard EEG database," *Clin. Neurophysiol.*, vol. 128, no. 9, pp. 1737–1745, Sep. 2017.

[23] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*, Boca Raton, FL: Chapman & Hall. CRC press, 1993.

[24] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Application*. Cambridge University Press, 1997.

[25] B. Efron, "Second Thoughts on the Bootstrap," *Stat. Sci.*, vol. 18, no. 2, pp. 135–140, May 2003.

[26] A. W. Bowman and A. Azzalini, *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. OUP Oxford, 1997.

[27] J. A. K. Suykens, J. De Brabanter, L. Lukas, and J. Vandewalle, "Weighted least squares support vector machines: robustness and sparse approximation," *Neurocomputing*, vol. 48, no. 1–4, pp. 85–105, Oct. 2002.