

---

This is an electronic reprint of the original article.  
This reprint may differ from the original in pagination and typographic detail.

Tapani, Karoliina T.; Nevalainen, P.; Vanhatalo, Sampsa; Stevenson, Nathan J.

## Validating an SVM-based neonatal seizure detection algorithm for generalizability, non-inferiority and clinical efficacy

*Published in:*  
Computers in Biology and Medicine

*DOI:*  
[10.1016/j.combiomed.2022.105399](https://doi.org/10.1016/j.combiomed.2022.105399)

Published: 01/06/2022

*Document Version*  
Publisher's PDF, also known as Version of record

*Published under the following license:*  
CC BY

*Please cite the original version:*  
Tapani, K. T., Nevalainen, P., Vanhatalo, S., & Stevenson, N. J. (2022). Validating an SVM-based neonatal seizure detection algorithm for generalizability, non-inferiority and clinical efficacy. *Computers in Biology and Medicine*, 145, 1-10. Article 105399. <https://doi.org/10.1016/j.combiomed.2022.105399>

---

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.



# Validating an SVM-based neonatal seizure detection algorithm for generalizability, non-inferiority and clinical efficacy

Karoliina T. Tapani<sup>a,b,c,d,\*</sup>, Päivi Nevalainen<sup>a,e</sup>, Sampsa Vanhatalo<sup>a</sup>, Nathan J. Stevenson<sup>c</sup>

<sup>a</sup> BABA Center, Children's Hospital and Pediatric Research Center, Helsinki University Hospital and University of Helsinki, Helsinki, Finland

<sup>b</sup> Department of Neuroscience and Biomedical Engineering, Aalto University, Espoo, Finland

<sup>c</sup> Brain Modelling Group, QIMR Berghofer Medical Research Institute, Australia

<sup>d</sup> Department of Medical Physics, Kymenlaakso Central Hospital, Kotka, Finland

<sup>e</sup> Epilepsia Helsinki and Department of Clinical Neurophysiology, HUS Diagnostic Center, Helsinki University Hospital and University of Helsinki, Finland

## ARTICLE INFO

### Keywords:

neonatal EEG  
EEG monitoring  
Neonatal intensive care unit  
Seizure  
Support vector machine

## ABSTRACT

Neonatal seizure detection algorithms (SDA) are approaching the benchmark of human expert annotation. Measures of algorithm generalizability and non-inferiority as well as measures of clinical efficacy are needed to assess the full scope of neonatal SDA performance. We validated our neonatal SDA on an independent data set of 28 neonates. Generalizability was tested by comparing the performance of the original training set (cross-validation) to its performance on the validation set. Non-inferiority was tested by assessing inter-observer agreement between combinations of SDA and two human expert annotations. Clinical efficacy was tested by comparing how the SDA and human experts quantified seizure burden and identified clinically significant periods of seizure activity in the EEG. Algorithm performance was consistent between training and validation sets with no significant worsening in AUC ( $p > 0.05$ ,  $n = 28$ ). SDA output was inferior to the annotation of the human expert, however, re-training with an increased diversity of data resulted in non-inferior performance ( $\Delta\kappa = 0.077$ , 95% CI:  $-0.002$ - $0.232$ ,  $n = 18$ ). The SDA assessment of seizure burden had an accuracy ranging from 89 to 93%, and 87% for identifying periods of clinical interest. The proposed SDA is approaching human equivalence and provides a clinically relevant interpretation of the EEG.

## 1. Introduction

Clinical suspicion of neonatal seizures or an identified risk of seizures due to e.g. hypoxic-ischaemic encephalopathy (HIE), meningitis or stroke are key indications for electroencephalography (EEG) monitoring in neonatal intensive care units (NICU) [1]. To guide interventions and inform prognosis, reliable and accurate detection of seizures is essential [2,3]. Recent guidelines insist that neonatal seizures must have an EEG correlate as clinical recognition of seizures is confounded by a range of factors [4]. Neonatal seizure burden (the total accumulated duration of seizures) is associated with poor neurodevelopmental outcome [5]. As such, clinical practice aims to treat neonatal seizures with an array of anti-seizure medications to reduce the seizure burden over the first few days of life. As clinical experts are often not available on demand in the NICU, automated decision support systems to assist rapid clinical decision making are needed to reduce seizure burden as much as possible. Neonatal EEG monitoring generates a considerable amount of data and

its interpretation is a challenge that has led to the development of several automated seizure detection algorithms (SDA) [6–19].

While the performance of many of these algorithms has thoroughly been assessed within the training data, few of them have been validated on an independent data set [9,18,20–23]. This step is essential as it ensures that SDAs built on relatively small data sets of neonatal EEG generalize to the larger population of neonatal EEG recordings. While generalizability is important, it alone does not suffice to define SDA performance. An SDA must also be accurate and clinically useful. The complete assessment of SDAs, therefore, must address four important questions: 1) does the SDA performance generalize? 2) does the SDA generate an annotation that is non-inferior to the human expert, taking into account ambiguity in the annotations of human experts? 3) does the annotation of the SDA accurately reflect the seizure burden experienced by a neonate over a clinically relevant period of time (e.g. birth to 72–96 h post-natal age)? and 4) will the implementation of the SDA improve clinical practice?

\* Corresponding author. BABA Center and Department of Clinical Neurophysiology, Children's Hospital and Pediatric Research Center, Helsinki University Hospital and University of Helsinki, Helsinki, Finland.

E-mail address: [karoliina.tapani@aalto.fi](mailto:karoliina.tapani@aalto.fi) (K.T. Tapani).

<https://doi.org/10.1016/j.combiomed.2022.105399>

Received 16 December 2021; Received in revised form 21 February 2022; Accepted 9 March 2022

Available online 23 March 2022

0010-4825/© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

The SDA developed by Temko et al. (2011) has the most comprehensive assessment to date [24]. Promising early performance was validated on an independent data set and the annotation of the SDA was shown to provide an accurate reflection of seizure burden [23]. Non-inferiority of the SDA annotation to the annotation of the human expert was, however, not comprehensively answered and the performance of the SDA when clinically trialled was underwhelming [25]. Recent advances in SDAs and subsequent improvements in accuracy, hold out hope that modern SDAs will soon achieve a level of performance that will, ultimately, improve clinical practice.

In this paper we developed a novel process for validating a neonatal SDA by 1) assessing the generalizability of our previously developed SDA on an unseen validation data set; 2) determining if the SDA generates an annotation that is non-inferior to the human expert; and 3) determining if the annotation of the SDA accurately estimates the seizure burden on a neonate. In addition, we proposed several modifications to the SDA to improve its performance on unseen data. We also compared the performance of our SDA to a state-of-the-art SDA by Temko et al. (2011). While we do not have access to the implementation of the SDA of Temko et al. (2011), we will use our implementation of their feature set, trained on our data, as a benchmark.

## 2. Materials and methods

This study utilizes two cohorts of EEG recordings. The first data set (training data set) is used to develop and train the neonatal SDA. The second data set (validation data set) is used to validate the SDA. To test the ability of the SDA to generalize to unseen data, several performance measures evaluating differences between the SDA and human annotation of the EEG, were compared between the training and validation data sets. The smaller the difference in these performance measures the more generalized the SDA. The non-inferiority of the SDA annotation compared to the human annotation of the validation set was examined using measures of inter-observer agreement (IOA). Clinically relevant measures of seizure burden were compared between the SDA and human expert annotations to assess the clinical efficacy of the algorithm. Finally, the validation data set was split in two, with the first third ( $n = 10$ ) added to the training set and the latter two thirds ( $n = 18$ ) saved as a final test set. This process helps determining if more diverse training data allows improving SDA generalizability or non-inferiority.

### 2.1. EEG acquisition and annotation

**Training data set:** This data set consists of EEG recordings from a cohort of 79 neonates with a variety of aetiologies. These neonates were recorded with a short duration paradigm and a whole scalp, 10–20 EEG system between 2010 and 2014. The EEG was recorded using a Nicolet One EEG monitor (Nicolet One, Natus Medical Incorporated, USA) with 19 scalp electrodes placed on positions Fp1, Fp2, F3, F4, F7, F8, T3, T3, C3, C4, T5, T6, P3, P4, O1, O2, Fz, Cz, Pz with sampling rate of 256Hz. The median duration of each recording was 74 min (IQR: 64–96 min). A bipolar derivation (double banana) was annotated by three clinical experts and used to train the SDA. The data set contains 460 seizures from 39 neonates (in consensus).

When training the SDA, it is noted that there is a preponderance of data in the non-seizure class compared to seizure class (458796 non-seizure epochs vs 20373 seizure epochs [16]). Class imbalance in the databases has been shown to be a problem when constructing adult SDAs [26]. However, this is not as significant a problem in neonatal SDAs, particularly SVM based methods, where seizures are more prevalent and the process of annotation is more concerned with false positives (seizures) than false negatives [19,27]. More details on the data set can be found in Ref. [28].

**Validation data set:** This data set consists of EEG recordings from a cohort of 28 neonates with stroke or HIE (Table 1). This validation set is a subset of the cohort first studied in Ref. [29].

**Table 1**

Table of patient characteristics for the validation data set ( $n = 28$ ). Data shown as  $n$  (%), or mean (standard deviation).

Patient characteristics ( $n = 28$ )	
Gestation age (weeks)	39.2( $\pm$ 2)
Birth weight (g)	3300( $\pm$ 550)
pH	7.2( $\pm$ 0.11)
Etiologies	
Stroke	25 (89%)
Hypoxic-ischemic encephalopathy	3 (11%)

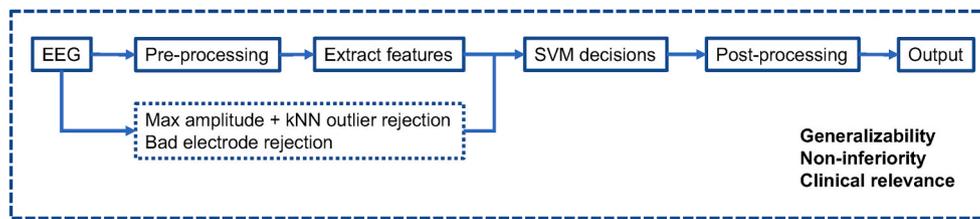
These neonates were recorded with long duration brain monitoring between 2011 and 2016. The signal was acquired at 250Hz using a Nicolet One EEG monitor (Nicolet One, Natus Medical Incorporated, USA) and four needle electrodes placed on positions F3, F4, P3, P4. The total length of all recordings was 86 days 16 h (median 77 h 50min, IQR: 48 h 20min - 94 h 35min). A three channel, bipolar derivation (F3–P3, F4–P4, P3–P4) was annotated by two clinical experts (PN, SV: both employed at Helsinki University Hospital) and processed by the SDA. The start and stop times of each seizure were noted. For annotations, the internationally approved neonatal seizure definition by Clancy et al. was used [30]. Clinical experts were blinded to each others annotations.

The training and validation sets were both collected in the neonatal intensive care unit (NICU) of the Children's Hospital, Helsinki University Hospital, Finland. The EEG was recorded as standard of care in our NICU. The use of the training EEG recordings were approved by the Ethics Committee of the Helsinki University Children's Hospital, Finland. The Institutional Research Review Board at Helsinki Children's Hospital approved the extraction and collation of the validation data set from hospital records and waived consent due to the study's retrospective and observational nature.

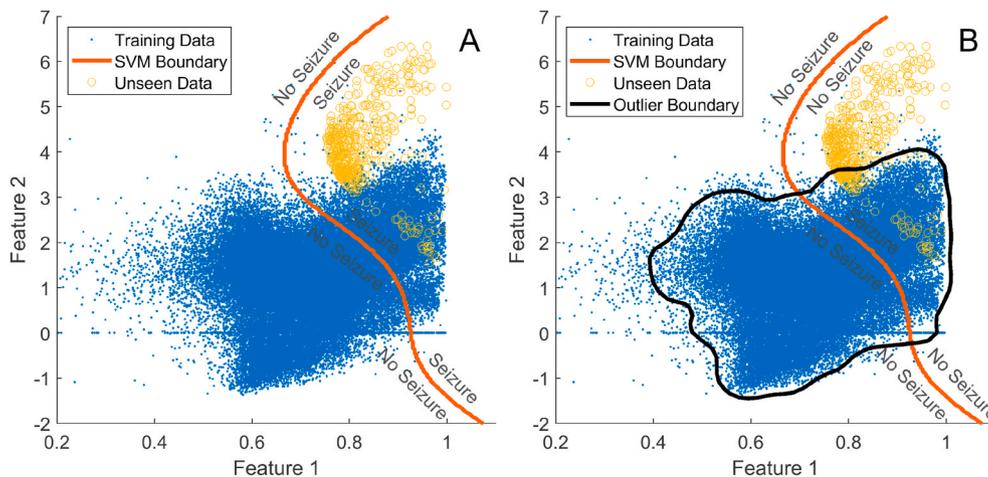
### 2.2. SDA

This SDA is based on a 'bag of features' calculated on 16s epochs of EEG. The combination of features to form a decision statistic was performed with a support vector machine (SVM) [16]. We used 22 summary EEG measures/features to represent a 16 s epoch of EEG. One additional feature to the previously published algorithm was maximum amplitude, which we used to discard epochs with high amplitude. To improve the robustness of the SDA, we incorporated an outlier detection system into our SDA. This was due to the relatively small size of our database and the expected difference in EEG quality between high density, short duration training data and low density, long duration validation data. Therefore we used a modified version of our previous system that was first proposed in Refs. [16,31]. Post-processing of the initial SDA output included eliminating outliers (we also eliminated EEG epochs detected as 'bad electrode' by the EEG monitor), applying a temporal moving average to each channel, taking the maximum value across channels and thresholding to form a binary (seizure/no seizure) decision. The binary decision was then extended in time using a collar. The parameters for outlier determination and seizure detection were optimized on training data by maximizing the agreement between the SDA and human annotation.

In summary, the algorithm works as follows. Neonatal EEG is first pre-processed before features are extracted from 16 s epochs. The features are combined by the SVM into a decision statistic which is then post-processed (median filter in time, maximum across channels and collaring) to form the final output of seizure and non-seizure. The post-processing steps, feature calculation, selection, and post-processing steps are explained in detail in the original SDA publication [16]. The flowchart of the SDA is presented in Fig. 1.



**Fig. 1.** Flowchart of the algorithm with outlier rejection. The first row presents the structure of the original SDA. In this validation study, the masking of outliers and bad electrodes (dotted box on the second row) was included for robustness on new data. The final evaluation of the SDA provides measures for generalizability, non-inferiority to human annotation and assessment of clinical relevance.



**Fig. 2.** The effect of outlier detection on SDA annotation. The blue dots denote most of the training samples in the database. A) The original SVM based SDA. The thick red line is the SVM decision seizure boundary for determining 'seizure' or 'no seizure' and the yellow circles represent incoming, unseen data (potential false detections). B) SVM based SDA with outlier detection. The black boundary is based on kNN distances of the training data; the updated decision boundary reduces the size of the feature space assigned to 'seizure'.

### 2.2.1. Outlier detection

Outlier detection identifies periods of the EEG recording that fall outside its experience; in our case the span of the feature set calculated on the training data. These methods are important to incorporate into clinical decision algorithms as the full variety of EEG patterns may never be captured in training data sets and when the detection of the rarer, clinically important class (seizure) is critical. The implementation of outlier detection, therefore, ensures that unusual EEG patterns (typically associated with artefact or rare neurological conditions) which fall outside the span of training data are dealt with differently. For our SDA, we adhere to the principle of *primum non nocere* and annotate 'no seizure' when an outlier is detected.

The effect of outlier detection on seizure annotation is shown in Fig. 2. The example is based on a 2 feature SVM. In 2A, the original SVM shows the separation between the two classes, 'seizure' and 'no seizure'. In 2B the black outlier boundary is included and is based on distances of the 3 nearest neighbours within the training data. Any data falling outside the decision boundary is labelled as 'no seizure' resulting in a smaller decision space for 'seizure'.

As not all features in our original feature set are best suited for outlier detection, we use both general feature space distances,  $k$ -nearest neighbours (kNN) search, and feature specific (EEG amplitude) distances. If an incoming data point is either greater than a threshold Euclidean distance from the  $k$ -nearest neighbours in the training data set or greater than a threshold EEG amplitude, then an outlier is detected and no seizure annotated.

We used the `knnsearch` function in Matlab [32], which outputs ascending Euclidean distances of an incoming data point to an array of training data points. In the case of our SVM-based detector, this data point is represented by a vector of features computed on a single epoch of EEG. The number of distances measured, that is the number of neighbours, is defined by parameter  $k$ .

We complemented our thresholding of feature space distances with a

stricter, separate threshold on EEG amplitude. The number of neighbours, distance threshold for the kNN search, EEG amplitude threshold, decision collar and decision threshold were, therefore, optimized on the training data set.

### 2.3. Analysis of algorithm performance

The most commonly used measure in SDA performance assessment is the area under the curve of the receiver operator characteristics (AUC), a second by second comparison of the algorithm annotation to the human annotation. It combines two important measures, true positive rate (or sensitivity) and false detection rate (or specificity) on multiple detection thresholds. These measures can be defined with the numbers of true positives/seizure (TP), true negatives/non-seizure (TN), false positives (FP) and false negatives (FN). Sensitivity is defined by  $TP/(TP + FN)$  and specificity by  $TN/(TN + FP)$ . In addition to these temporal measures we assessed seizure detection rate (SDR), that is the number of correctly detected seizure events, and false detections per hour (FD/h), which is the number of falsely detected seizure events divided by the duration of the recording in hours. These measures can be supplemented with measures such as the Kappa family of statistics; we use Cohen's kappa as we only compare between two annotations at a time [33].

We calculated performance measures on individual neonates and also on a single, concatenated, consensus of human expert annotation for comparison. As each data set contained neonates with no seizure, a concatenated annotation avoids complications of per neonate assessments (sensitivity/AUC and SDR cannot be calculated on neonates who do not have seizure). It also provides a more useful assessment of event based measures as several recordings in the training data set had limited events on relatively short recordings resulting in poor quantization per neonate.

### 2.3.1. Implementation of outlier detection

The cross-validation results of the original SDA and SDA with outlier detection were compared using statistical testing. Differences between AUC, sensitivity and specificity calculated on a per neonate basis were tested using Wilcoxon signed rank test (for paired data) due to non-normal distribution of these values across training and validation cohorts. AUC and sensitivity can only be calculated on neonates with seizure, while specificity was calculated on all neonates in the training and validation cohorts. All tests were two-sided and the level of significance was 0.05.

Analysis of concatenated recordings was also performed using bootstrapped statistics. In this case, a concatenated annotation from each data set was generated using a random sampling of neonate annotations (sampling with replacement). The difference between performance measures were calculated on each re-sampled annotation. The difference in performance measures was calculated for 1000 resamplings and used to estimate a 95% confidence interval of differences. If the difference spanned zero, there was no significant difference in performance between data sets. This statistical method was used to compare performance where applicable (we cannot compare bootstrapped distributions between different data sets).

As outlier parameters were optimized using IOA between the algorithm and human experts, we used Cohen's kappa as the primary measure of performance differences between the SDA, with and without, outlier detection.

### 2.3.2. Generalizability of the SDA

We measured generalization by comparing performance measures between cross-validation assessments of the training dataset and the validation dataset. If there was no statistically significant difference in these measures between the training and validation sets then we claim that the SDA *generalizes* to a larger population of neonates. The primary measure of generalizability was statistical testing of AUC.

Differences between AUC, sensitivity and specificity calculated on a per neonate basis were tested similarly to the previous section, with the exception that a Mann-Whitney U tests (for unpaired data) was used. If the difference in AUCs is not significant, the SDA generalizes on unseen data.

### 2.3.3. Non-inferiority to human annotation

The limitation of AUC as a measure of performance is that it does not inherently determine if the SDA annotation of the EEG is 'good enough'. This is particularly valid in instances where there is ambiguity in the 'gold standard' used to calculate the AUC such as the manual annotation of the EEG by human experts. We, therefore, determined the non-inferiority of the SDA to the annotations of the human expert using bootstrapped measures of IOA. Each expert and the SDA were compared to each other using Cohen's kappa statistic. The difference between the kappa of SDA/human expert pairs and human/human pairs was estimated using a bootstrap (1000 iterations). If the difference spanned zero then the SDA/human kappa was assumed to be equivalent/non-inferior to the human/human kappa. This analysis uses the complete annotation of each human expert and avoids the requirement of consensus annotations.

We also investigated SDA errors at the annotation and feature level. At the annotation level, we compared general summary measures of the annotation such as the total seizure burden, seizure number and seizure duration. At the feature level, we determined if some features were systematically different between the training and validation data sets.

### 2.3.4. Clinical efficacy

While non-inferiority provides an effective means to compare the annotation of an SDA to the annotation of human experts, it does not evaluate the practical clinical impact of any discrepancies. To this end,

we determined the clinical efficacy of SDA by comparing clinically relevant interpretations of SDA and human annotations of the EEG. The first set of measures compares a continuous assessment of seizure burden. This measure is considered clinically important as it provides a clear quantitative target for treatment.

While some clinicians are aggressively treating any amount of seizures with a variety of anti-epileptic medications [1,34,35], there is also substantial variance among clinicians with respect to the level of seizure burden that is considered to prompt treatment. We compared hourly assessments of seizure burden using the correlation coefficient and a diagnostic binary decision of periods of clinical targets and periods of no clinical target in an EEG recording. We define a clinical target as a 2-h period with two 30s seizures or 3 min of accumulated seizure [4]. We then compare the number of clinical targets defined by the human annotation to the number defined by SDA annotation using sensitivity and specificity.

We also evaluate the prognostic utility of the SDA. Recent work has shown that high total seizure burden (> 45 min) and high maximum hourly seizure burden (> 13 min per hour) in a long duration recording of EEG are associated with poor neurodevelopmental outcome in neonates with HIE [5]. Note, that a high hourly seizure burden is typically used to define status epilepticus (>30 min per hour). We use 13 min per hour as an alternate definition of high seizure burden due to its association with poor developmental outcome [5]. We compare the SDA annotation to the human consensus annotation for the prediction of high or low seizure burden in each neonate using measures of sensitivity and specificity. We also show how differences in human interpretations affect these measures by averaging sensitivity and specificity between experts, using expert 1 as the 'gold standard' and using expert 2 as the 'gold standard'.

### 2.4. Re-training

Finally, we investigated the incorporation of additional data from the validation set into the training set on SDA performance. We extracted seizure and non-seizure epochs from the first 10 neonates in the validation data set. The epochs for each class were randomly sampled from consensus segments, so that the size of this new training set was comparable to the size of the original training set. The non-seizure epochs were randomly sampled from the second half of the recordings only, as the majority of false detections occurred in this time period (see Results). To increase diversity in the training data set, not the total number of samples, every second sample from the new and from the original training data was used to re-train the SVM. Generalizability and non-inferiority measures were calculated on the re-trained SVM to determine the effectiveness of added diversity in the training data.

## 3. Results

### 3.1. Implementation of outlier detection and generalization of the SDA

The incorporation of an outlier detection significantly improved the agreement between the SDA output and the consensus annotation of the human expert (Table 2). Slight, but significant reductions in per infant AUC were offset by significant improvements in specificity and false detections per hour.

The SDA had consistent performance between training and validation sets with our primary measure for generalizability, the AUC, suggesting that the algorithm generalizes to unseen data (Table 2). That is, there was no significant reduction in performance between the cross-validated training data set and the validation data. The decrease in sensitivity was significant, while there was no statistical difference for specificity. The proposed SDA outperformed an SDA using the Temko feature set (Table 2).

**Table 2**

The performance of a neonatal SDA with outlier detection. The neonatal SDA with outlier detection was applied to the original dataset (10-fold cross-validation) and compared to the original SDA without outlier detection. The neonatal SDA was also applied to an independent validation set and compared to an SDA based on the Temko feature set (Temko FS). Median (IQR) are presented for AUC, sensitivity (Sens) and specificity (Spec). The value with 95% CIs in brackets are presented for cAUC, cSDR, cFD/h and cKappa (estimated on concatenated recordings). <sup>a</sup> and <sup>b</sup> denote a significant increase and decrease, respectively, between the original SDA and the SDA with outlier detection. <sup>c</sup> denotes a significant increase in SDA performance between the training and validation data sets. <sup>d</sup> denotes significant decrease in performance between proposed SDA and SDA based on Temko features. Primary measures of performance are shown in bold.  $n_t$  is the total number of neonates in each dataset and  $n_s$  is the number of neonates with consensus seizure in each dataset.

	Original Training (10-fold CV)	Original Training (10-fold; outlier detection)	Validation (outlier detection)	Validation (Temko FS)
	$n_t = 79, n_s = 39$	$n_t = 79, n_s = 39$	$n_t = 28, n_s = 24$	$n_t = 28, n_s = 24$
AUC	0.992 (0.938–0.998)	0.986 (0.908–0.995) <sup>b</sup>	<b>0.960 (0.896–0.980)</b>	0.918 (0.823–0.961) <sup>d</sup>
Sens	0.823 (0.341–0.975)	0.784 (0.363–0.958)	0.605 (0.218–0.753) <sup>c</sup>	0.385 (0.061–0.674) <sup>d</sup>
Spec	0.993 (0.979–1)	0.996 (0.982–1) <sup>a</sup>	0.998 (0.997–0.999)	0.998 (0.994–0.999) <sup>d</sup>
cAUC	0.956 (0.931–0.979)	0.955 (0.928–0.977)	0.967 (0.953–0.974)	0.949 (0.931–0.961) <sup>d</sup>
cSDR	0.750 (0.582–0.851)	0.685 (0.531–0.791)	0.761 (0.631–0.841)	0.657 (0.522–0.750) <sup>d</sup>
cFD/h	2.290 (1.312–3.897)	1.967 (0.964–3.478) <sup>b</sup>	0.357 (0.167–0.690)	0.302 (0.135–0.513)
cKappa	0.668 (0.464–0.749)	<b>0.672 (0.474–0.764)<sup>a</sup></b>	0.729 (0.656–0.789)	0.623 (0.533–0.694) <sup>d</sup>

**3.2. Non-inferiority to human annotation**

The SDA provides an annotation of the EEG that is slightly inferior to human experts (Table 3;  $\Delta\kappa$  fails to span 0). Example EEG segments that

**Table 3**

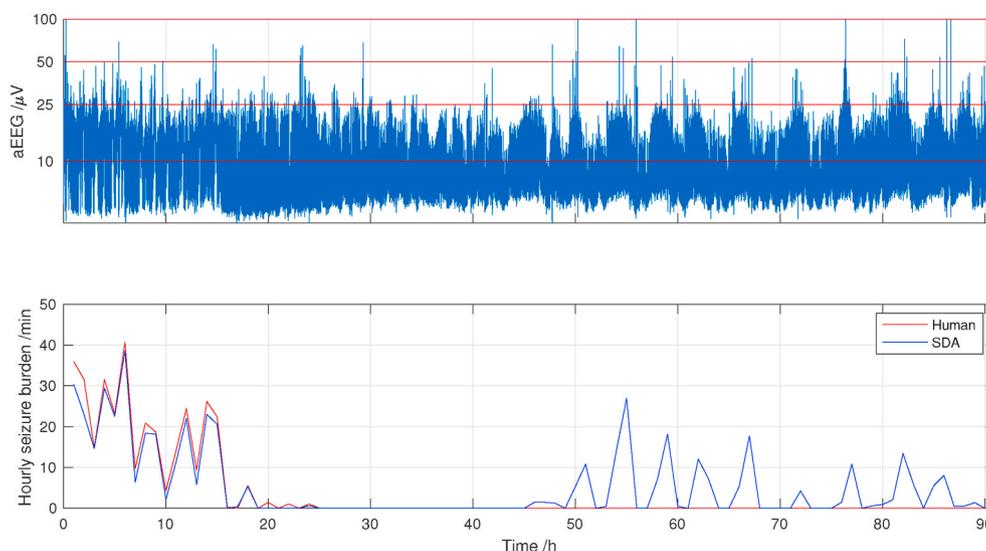
Inter-observer agreement between human and algorithm based annotations of seizure/non-seizure in the validation set of EEG. E1 is human expert 1, E2 is human expert 2, Consensus is the consensus annotation of human experts,  $\Delta\kappa$  is difference in IOA between a single human expert and the SDA and the IOA between human experts (results are presented as median, 95% confidence interval), FS is feature set.

Inter-observer Agreement	E1	E2	Consensus
SDA (proposed FS)	0.591	0.703	0.729
SDA (Temko FS)	0.511	0.600	0.623
E1	–	0.792	0.805
E2	–	–	0.982
SDA $\Delta\kappa$ (proposed FS)	<b>0.197</b> (0.149–0.269)	<b>0.085</b> (0.011–0.176)	–
SDA $\Delta\kappa$ (Temko FS)	<b>0.277</b> (0.220–0.364)	<b>0.187</b> (0.113–0.289)	–

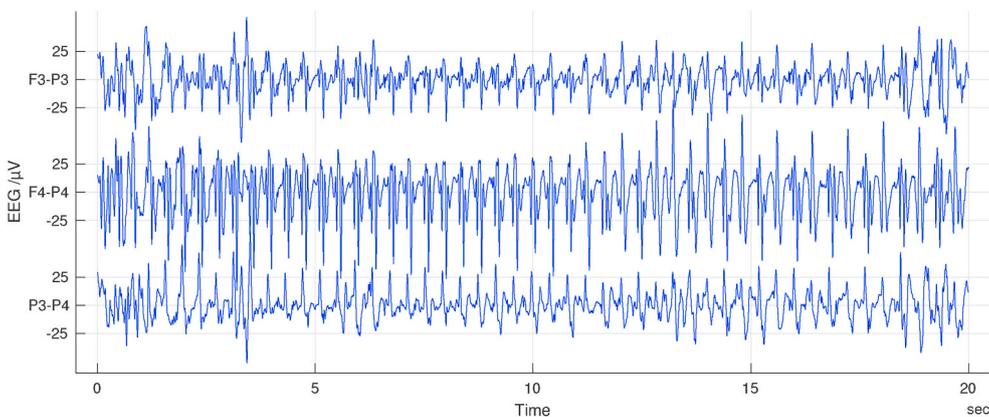
were correctly detected as seizure, missed or falsely labelled as seizure are shown in Fig. 4.

At the feature level, the major differences between training and validation data sets were between the smoothed nonlinear energy operator feature within EEG epochs without seizure (training data: median [IQR] 39.2 [15.2–74.8], validation data: median [IQR] 69.9 [43.5–111.4]; values are not normalized). Increases in the nonlinear energy imply increases in high frequency energy in seizure signals.

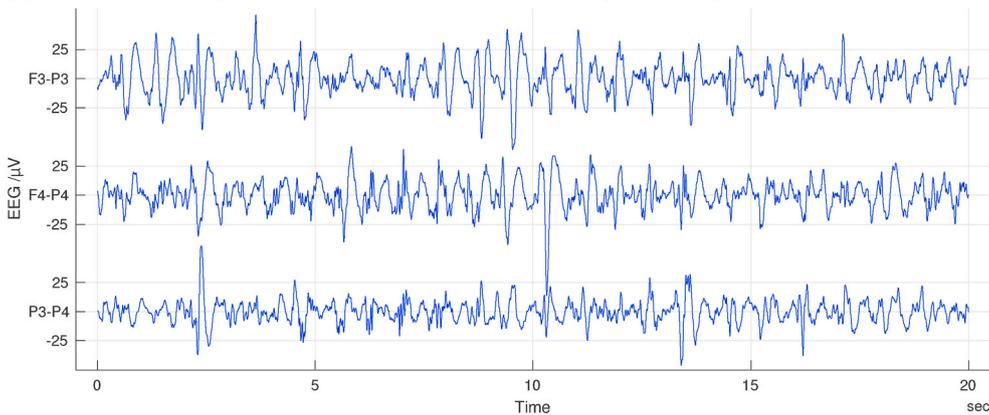
At the annotation level, the SDA has a noticeable reduction in seizure detection rate and, in general, it generates an excessive number of seizures compared to the consensus seizure annotation. This is in the context of an SDA annotation that is, nevertheless, more conservative and diffuse with its annotation; the SDA has lower estimates of seizure burden measures (maximum and total, Table 4). The SDA also tends to generate more false detections during the second half of recordings compared to the first. A representative example is shown in Fig. 3. In general, the median false detection rate was 0.13 per hour (IQR: 0.03–0.17) during the first half of the recording and 0.20 per hour (IQR: 0.06–0.42) during the second half of the recording. The second half of recordings is more likely to contain EEG patterns associated with recovery and, therefore, more normal EEG such as sleep states. These patterns may not be prevalent in the training data.



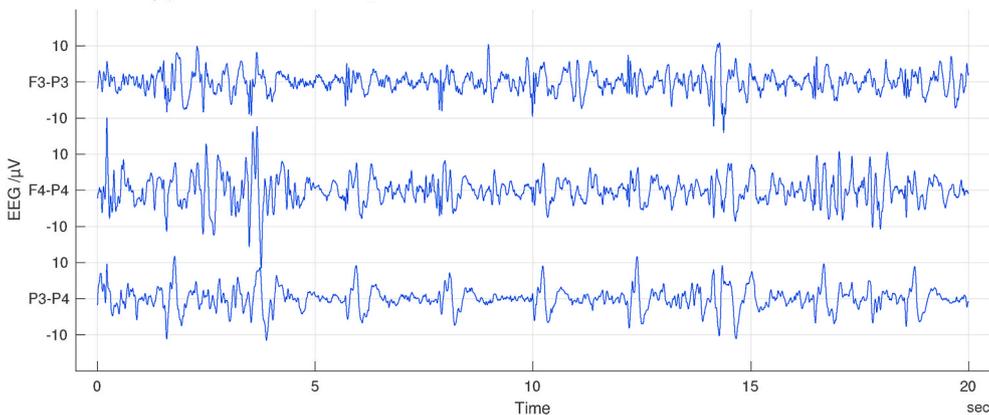
**Fig. 3.** An example aEEG and seizure burden time course. Note the appearance of sleep-wake-cycling after 50 h. This will introduce normal EEG patterns such as mixed frequency activity, high voltage synchronous activity and trace alternant, that will not be prevalent in the original training data as those recordings were typically taken earlier in the seizure burden time course.



(a) True detection on patient 2. The seizure is on all channels, but most prominent on channel F4-P4.



(b) False detection on patient 34. Channel F3-P3 raises the false alarm.



(c) Missed seizure on patient 1. Seizure is on all channels, but predominantly on channel P3-P4.

**Fig. 4.** Examples of a true detection, a false detection and a missed seizure.

**Table 4**

Seizure summary statistics of seizures annotated by SDAs and the consensus of human experts. Results are presented as median (IQR) summarized across neonates with seizures. Seizure duration is presented as median (IQR) over all annotated seizures.<sup>1</sup> 24 neonates had seizures annotated by both human consensus and the SDA (proposed FS) and <sup>2</sup> 23 by consensus and SDA (Temko FS).

Seizure statistics	Consensus	SDA (proposed FS)	SDA (Temko FS)
Seizure patients	24	28 <sup>1</sup>	27 <sup>2</sup>
Seizure number	33 (11–68)	29 (17–119)	14 (8–120)
<b>Seizures (total)</b>	<b>1362</b>	<b>2313</b>	<b>1749</b>
Seizure duration (s)	96 (55–167)	44 (32–84)	64 (52–104)
Burden (max; min/h)	13 (3–29)	8 (3–20)	9 (2–21)
Burden (total; min)	42 (11–163)	28 (10–138)	24 (8–201)

### 3.3. Clinical efficacy

To assess the clinical efficacy of the SDA, we used more clinically focused metrics (Table 5). The key component of this assessment is evaluation of the short-term seizure burden per hour (Fig. 5). The correlation between human consensus and SDA for short-term seizure burden per hour on concatenated data was 0.89 (95% CI: 0.80–0.96). The correlation between human experts assessment of short-term seizure burden was 0.95 (95% CI: 0.93–0.98).

The hourly seizure burden is commonly used as a target to initiate treatment in clinical studies. The SDA identifies epochs that exceed a target threshold with an accuracy of 87% (Table 5a). The sensitivity of the SDA is 83% and the specificity is 87% which are in a similar range to

**Table 5**

Agreement in clinically relevant measures between the SDA and human consensus annotation of the EEG on all neonates. POI - period of clinical interest (hourly epochs of EEG that exceed a treatment threshold for seizures), nPOI - not a POI, H - human annotation (consensus), S - algorithm annotation, HB - high seizure burden, LB - low seizure burden.

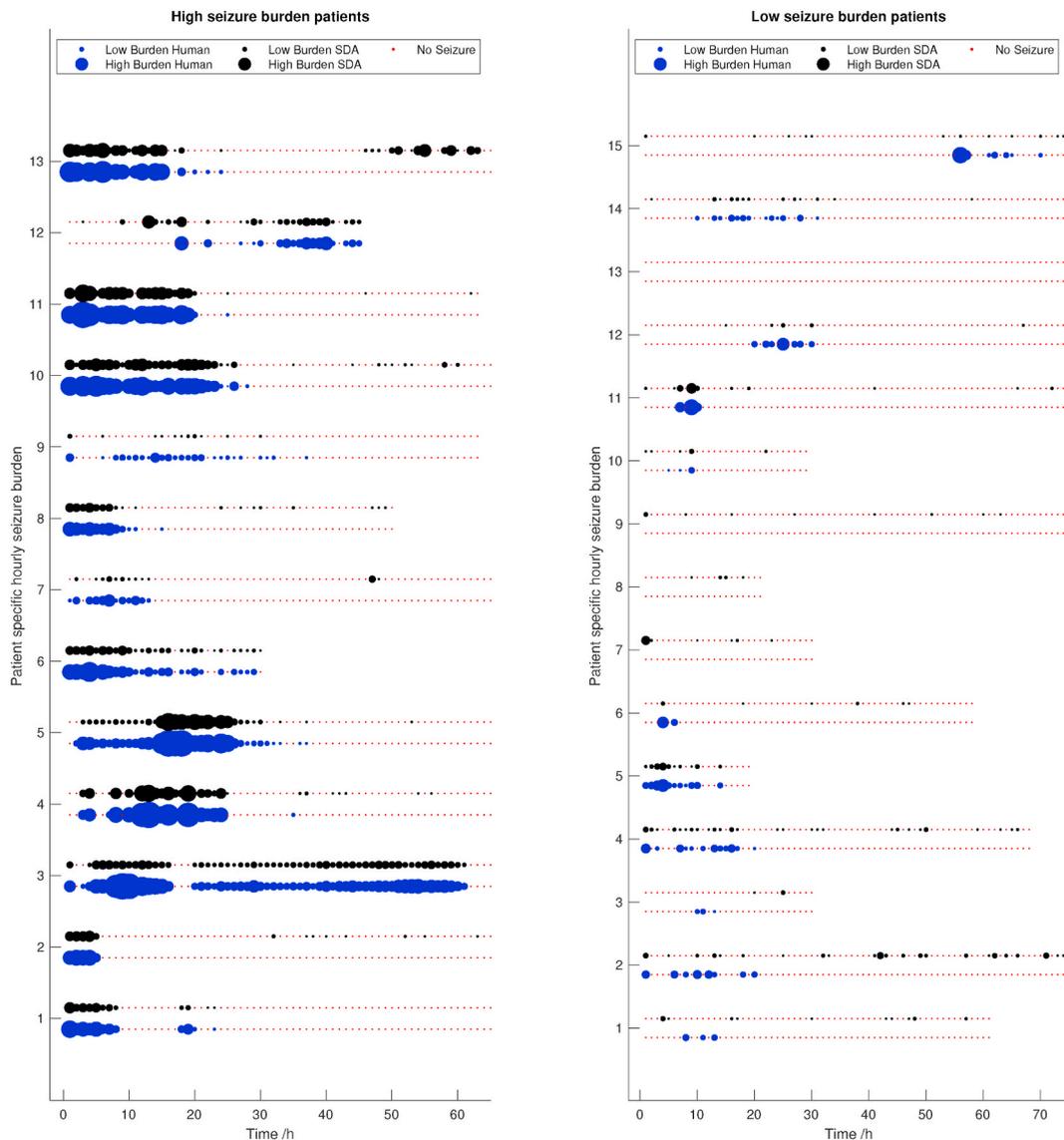
(a) Period of Clinical Interest (POI)		
	POI (S)	nPOI (S)
POI (H)	151	30
nPOI (H)	103	716
(b) Total Seizure Burden		
	HB (S)	LB (S)
HB (H)	11	2
LB (H)	0	15
(c) Maximum Hourly Seizure Burden		
	HB (S)	LB (S)
HB (H)	11	3
LB (H)	0	14

inter-observer agreement between human experts (sensitivity 89% and specificity 97%) for this measure.

The sensitivity and specificity of the SDA for detecting high total seizure burden (> 45 minutes; compared to the consensus annotation) are 85% and 100%, respectively (Table 5b). The sensitivity and specificity of human experts are 85% and 79%, respectively. Finally, for detecting high maximum hourly seizure burden (> 13 minutes per hour; compared to consensus), the sensitivity and specificity for the SDA are 79% and 100% (Table 5c), while the sensitivity and specificity of human experts are 94% and 92%, respectively. Importantly, the SDA does not detect high total or hourly seizure burden in a low seizure burden neonate.

### 3.4. Re-training

Incorporating seizure and non-seizure data from the first 10 recordings in the validation set into the training data set improved AUC and Kappa on the remaining 18 neonates in the validation set (Table 6). Improved sensitivity was offset by a slight reduction in specificity. These improvements were sufficient to generate an annotation that was non-inferior to the human annotation ( $\Delta\kappa = 0.077$ , 95% CI:  $-0.002-0.232$ ,



**Fig. 5.** Hourly (short-term) seizure burden evolution over time on human consensus and SDA annotation. Each marker represents an hour segment and the marker thickness the hourly seizure burden. Time series were cut after no more events occurred for visualization purposes. Red dots stand for no seizure.

**Table 6**

The performance of the re-trained neonatal SDA. The re-trained neonatal SDA was applied to the original dataset (10-fold cross-validation) and the truncated validation set ( $n = 18$ ). These values were compared to the original SDA applied to the truncated validation set and a re-trained SDA based on the Temko feature set (Temko FS). Median (IQR) results are presented for AUC, sensitivity (Sens) and specificity (Spec). The value with 95% CIs in brackets are presented for cAUC, cSDR, cFD/h and cKappa (estimated on concatenated recordings). <sup>a</sup> and <sup>b</sup> denote a significant increase and decrease, respectively, between Validation and Validation (retrained), <sup>c</sup> and <sup>d</sup> denote a significant increase and decrease, respectively, between Validation (retrained) and Validation (retrained Temko FS), and <sup>e</sup> and <sup>f</sup> denote a significant decrease between Training (10-fold CV) and Training (outlier detection; 10-fold CV) in Table 2  $n_t$  is the number of infants in the cohort and  $n_s$  is the number of infants with consensus seizure in the cohort.

	Original Training (10-fold CV) $n_t = 79, n_s = 39$	Validation $n_t = 18, n_s = 14$	Validation (retrained) $n_t = 18, n_s = 14$	Validation (retrained Temko FS) $n_t = 18, n_s = 14$
AUC	0.979 (0.930–0.996)	0.965 (0.880–0.980)	0.976 (0.932–0.983) <sup>d</sup>	0.953 (0.867–0.974) <sup>d</sup>
Sens	0.816 (0.433–0.947)	0.519 (0.156–0.759)	0.648 (0.311–0.837) <sup>d</sup>	0.395 (0.057–0.645) <sup>d</sup>
Spec	0.991 (0.973–1) <sup>f</sup>	0.998 (0.997–0.999)	0.997 (0.996–0.998) <sup>b</sup>	0.998 (0.997–0.999) <sup>c</sup>
cAUC	0.964 (0.944–0.980)	0.964 (0.932–0.978)	0.974 (0.953–0.984) <sup>d</sup>	0.956 (0.912–0.971) <sup>d</sup>
cSDR	0.667 (0.515–0.778)	0.679 (0.447–0.868)	0.729 (0.447–0.876)	0.600 (0.307–0.763)
cFD/h	1.943 (0.918–3.331)	0.414 (0.145–0.976)	0.332 (0.147–0.610)	0.191 (0.092–0.303)
cKappa	0.680 (0.507–0.769)	0.722 (0.571–0.795)	0.735 (0.560–0.808)	0.662 (0.484–0.726)

$n = 18$ ). In contrast, the original SDA, when applied to the truncated validation data set retained its inferior annotation of seizures ( $\Delta\kappa = 0.084$ , 95%CI: 0.003–0.226,  $n = 18$ ). Re-training did not significantly change primary performance measures when evaluated within a 10-fold cross-validation on the original cohort of 79 neonates (Table 6). Specificity was slightly, but significantly, reduced.

Re-training an SDA based on the Temko feature set showed slight, but not significant, improvements in performance compared to a version trained on the original cohort of 79 neonates. Compared to the proposed, retrained SDA, SDA retrained with Temko features had inferior performance with all measures except specificity and false detections per hour (Table 6). The annotation of the EEG by the retrained SDA with the Temko feature set remained inferior to the human expert annotation ( $\Delta\kappa = 0.145$ , 95%CI of 0.083–0.287,  $n = 18$ ).

#### 4. Discussion

In this study, we have evaluated the performance of our neonatal SDA on an independent validation data set consisting of limited channel, long duration EEG recordings. The performance of the SDA was not significantly different between cross-validation on the training data set and the validation data set, using the primary generalizability measure, the AUC. While the overall performance of the SDA remained consistent, it was, nevertheless, inferior to the human expert annotation of the validation set. The inferiority of the SDA annotation was mitigated by retraining on a larger data set containing data portions of the validation data set.

There is a limited number of validation studies performed on published neonatal SDAs [9,20–23]. These studies provide seizure detection rates of 53–66% and false detections of 0.04–2.3 per hour; values that are compatible with our SDA performance 76% at 0.35 false detections per hour. The limitations of these studies are that 1) they do not test the generalizability of their SDAs and 2) they do not evaluate the non-inferiority of the SDA annotation compared to that of the human expert taking into account the inherent ambiguity in the annotation of multiple human experts. The algorithm with the highest congruence between training and validation sets is that of Temko et al. with a seizure detection rate of 71% at 0.25 false detection per hour on the training set and a seizure detection rate of 75% at 0.36 false detection per hour on a validation set [23]. This algorithm was further shown to potentially aid clinical seizure recognition with earlier detection of seizure onset and improved temporal localization of clinical target periods [23]. It has recently been clinically trialled and shown to improve the recognition of hourly seizures, but not the identification on neonates with seizure, which may help optimize the timing of treatment [25]. We have shown that our feature set is capable of outperforming the feature set of Temko

et al. that underlies this algorithm, when trained on identical data.

Initially, the SDA provided an inferior annotation of seizures compared to human experts on the validation data set. The most obvious failing of the SDA was an excessive annotation of seizures. The SDA annotated twice as many seizures as the human experts although the majority of these seizures were multiple detections within a single human annotated seizure. Nevertheless, a considerable proportion were periods annotated by human experts as 'no seizure'. Reassuringly, visual inspection of false detections showed EEG with repetitive characteristics, activity that is commonly associated with seizure (see Fig. 4).

There are two key differences between the training and validation data sets that may explain increased errors in the validation set 1) reduced electrode density and increased recording duration of EEG in the validation set, and 2) different brain injury etiologies in the two data sets. Firstly, EEG monitoring with a reduced electrode set is commonly used in the NICU as it offers less disruption for neonates, is compatible with visualizations such as the aEEG and reduces the demand of cot-side resources. While these recordings capture the majority of seizures and have electrodes located on regions that are more robust to artefact, we have shown that low density recordings result in a different annotation to standard EEG recordings and have a reduced prevalence of seizure [28,36–38]. This means that the manifestation of neonatal seizure may be more diffuse as the distance between seizure loci and recording electrodes will be higher. Long duration EEG also contains a wider variety of normal EEG patterns that may not be seen in short duration, higher density recordings performed early to confirm the presence of seizures. This explains why the 'no seizure' feature values within the validation data set have a larger spread compared to the training data. Secondly, the validation set contains seizures mostly from a single aetiology (stroke) whereas the training set contained seizures from several aetiologies (HIE, stroke, meningitis) [28]. Seizures in neonates with stroke may have different morphology to seizures from other aetiologies and stroke was poorly represented in the training data set. This also means that recordings in the validation set will not contain prolonged periods of isoelectric EEG and burst-suppression that are apparent in the EEG recordings of neonates with conditions such as HIE. Despite these potential limitations, the SDA had consistent, general performance between the data sets and non-inferiority to human annotation could be achieved by incorporating a subset of the validation data set into the training set.

Nevertheless, the SDA annotation provides an accurate ( $\sim 90\%$ ) representation of key interpretations of the annotations such as identifying periods of EEG that meet treatment thresholds and definitions of high seizure burden that are associated with poor prognosis. These values may exceed current clinical practice when on-demand EEG annotation is not available, however, the risk of false identifications may

outweigh increased sensitivity. A risk that is minimized for a non-inferior annotation.

Bag of features approaches to neonatal seizure detection are being superseded by deep neural networks [18]. We and others have shown that deep convolutional neural networks provide similar performance to the proposed 'bag of features' approach when applied to our training set [18,31] with deep convolutional networks providing more consistent performance over several data sets [18]. These methods will, nevertheless, face similar challenges such as lack of diversity within training data sets, subjectivity in the gold standard of human expert annotation (the training target) and performance assessment.

A potentially controversial component of our study is the use of 10s as a minimum duration of seizure. The original choice of a 10s minimum seizure duration was always arbitrary and acknowledges that as evolution is an important component of determining the presence of seizure on an EEG recording, there must be some sort of minimum duration to judge such an evolution [30]. This minimum seizure duration has been abolished in the latest recommendations on the Task Force for Neonatal Seizure [4] but its abolition is disputed [39]. With the lack of biophysical evidence for a minimum seizure duration, it is best to turn to other forms of evidence such as studies of inter-observer agreement. These works provide the only data-driven evidence base for a minimum seizure duration and suggest that 30s may be a feasible option [38]. This, more lenient, definition of minimum seizure duration has the potential to improve the accuracy of SDAs. Furthermore, it is a simple way to achieve SDA annotations that generalize to the larger population of neonates and are non-inferior to the visual interpretation of the human expert.

## 5. Conclusion

We validated a neonatal SDA on an independent data set of neonatal EEG. We show consistent performance between training and validation data sets, but an SDA annotation that is inferior to the human expert. This annotation is, nevertheless, sufficient to provide highly accurate estimation of seizure burden and identification of clinical targets. Increasing the diversity of the training set resulted in a non-inferior SDA annotation for the left-out part of the validation data set. This raises key open questions for neonatal SDAs: 1) how much data is required and from what population of neonates should it be sampled to train neonatal SDAs that are non-inferior to the human expert, 2) is non-inferior annotation compared to human experts the ultimate target and 3) if and when validated non-inferiority is achieved, will it improve the clinical recognition of seizures and improve health outcomes for critically ill neonates. Once the algorithm has achieved a clinically acceptable level of performance, it will be trialed and finally, implemented. This process ensures that the algorithm positively affects clinical practice.

## Declaration of competing interest and conflict of interest

The authors have no competing interests to declare and no conflicts of interest.

## Author contribution statement

KTT contributed to study design, developed analysis tools, performed data processing and scripting for results. PN contributed to data collection, collation and annotation of the EEG recordings. SV contributed to study design, data collection, collation and annotation of the EEG recordings. NJS contributed to study design and algorithm development. KTT, NJS, PN, and SV contributed to the writing and editing of the paper manuscript.

## Acknowledgements

Funding for this research was provided by the Finnish Cultural

Foundation (00181077), Finnish Pediatric Foundation, the Finnish Academy (313242, 288220, 321235), Aivosäätiö, Neuroscience Center at University of Helsinki, Helsinki University Hospital, Helsinki University Hospital Research Funds, HUS and University of Helsinki Researcher position, Sigrid Juselius Foundation and European Union's Horizon 2020 Research and Innovation Programme (H2020-MCSA-ITN-813483 and H2020-MCSA-IF-656131). The study sponsors had no involvement in the study design, nor in the collection, analysis or interpretation of data. Additionally we acknowledge the computational resources provided by the Aalto Science-IT project.

## References

- [1] J. Rennie, G. Boylan, Treatment of neonatal seizures, *Arch. Dis. Child. Fetal Neonatal Ed.* 92 (2) (2007) F148–F150.
- [2] S. Björkman, S. Miller, S. Rose, C. Burke, P. Colditz, Seizures are associated with brain injury severity in a neonatal model of hypoxia-ischemia, *Neuroscience* 166 (1) (2010) 157–167.
- [3] S. Miller, J. Weiss, A. Barnwell, D. Ferriero, B. Latal-Hajnal, A. Ferrer-Rogers, N. Newton, J. Partridge, D. Glidden, D. Vigneron, et al., Seizure-associated brain injury in term newborns with perinatal asphyxia, *Neurology* 58 (4) (2002) 542–548.
- [4] R.M. Pressler, M.R. Cilio, E.M. Mizrahi, S.L. Moshé, M.L. Nunes, P. Plouin, S. Vanhatalo, E. Yozawitz, L.S. de Vries, K. Puthenveetil Vinayan, et al., The ILAE classification of seizures and the epilepsies: modification for seizures in the neonate. position paper by the ILAE task force on neonatal seizures, *Epilepsia* 62 (3) (2021) 615–628.
- [5] L. Kharoshankaya, N.J. Stevenson, V. Livingstone, D.M. Murray, B.P. Murphy, C. E. Ahearne, G.B. Boylan, Seizure burden and neurodevelopmental outcome in neonates with hypoxic-ischemic encephalopathy, *Dev. Med. Child Neurol.* 58 (12) (2016) 1242–1248.
- [6] A. Liu, J. Hahn, G. Heldt, R. Coen, Detection of neonatal seizures through computerized EEG analysis, *Electroencephalogr. Clin. Neurophysiol.* 82 (1) (1992) 30–37.
- [7] J. Gotman, D. Flanagan, J. Zhang, B. Rosenblatt, Automatic seizure detection in the newborn: methods and initial evaluation, *Electroencephalogr. Clin. Neurophysiol.* 103 (3) (1997) 356–362.
- [8] P. Celka, P. Colditz, A computer-aided detection of EEG seizures in infants: a singular-spectrum approach and performance comparison, *IEEE T Bio-Med. Eng.* 49 (5) (2002) 455–462.
- [9] L. Smit, R. Vermeulen, W. Fetter, R. Strijers, C. Stam, Neonatal seizure monitoring using non-linear EEG analysis, *Neuropediatrics* 35 (2004) 329–335, 06.
- [10] M.A. Navakatikyan, P.B. Colditz, C.J. Burke, T.E. Inder, J. Richmond, C. E. Williams, Seizure detection algorithm for neonates based on wave-sequence analysis, *Clin. Neurophysiol.* 117 (6) (2006) 1190–1203.
- [11] W. Deburchgraeve, P. Cherian, M. De Vos, R. Swarte, J. Blok, G. Visser, P. Govaert, S. Van Huffel, Automated neonatal seizure detection mimicking a human observer reading EEG, *Clin. Neurophysiol.* 119 (11) (2008) 2447–2454.
- [12] J. Mitra, J.R. Glover, P.Y. Ktonas, A.T. Kumar, A. Mukherjee, N.B. Karayannis, J. D. Frost Jr., R.A. Hrachovy, E.M. Mizrahi, A multi-stage system for the automated detection of epileptic seizures in neonatal EEG, *J. Clin. Neurophysiol.* 26 (4) (2009) 218–226.
- [13] A. Temko, E. Thomas, W. Marnane, G. Lightbody, G. Boylan, EEG-based neonatal seizure detection with support vector machines, *Clin. Neurophysiol.* 122 (3) (2011) 464–473.
- [14] A. Ansari, P. Cherian, A. Dereymaeker, V. Matic, K. Jansen, L. De Wispelaere, C. Dielman, J. Vervisch, R. Swarte, P. Govaert, et al., Improved multi-stage neonatal seizure detection using a heuristic classifier and a data-driven post-processor, *Clin. Neurophysiol.* 127 (9) (2016) 3014–3024.
- [15] A. O'Shea, G. Lightbody, G. Boylan, A. Temko, Neonatal seizure detection using convolutional neural networks, in: 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP), IEEE, 2017, pp. 1–6.
- [16] K.T. Tapani, S. Vanhatalo, N.J. Stevenson, Time-varying EEG correlations improve automated neonatal seizure detection, *Int. J. Neural Syst.* 29 (2019), 1850030, 04.
- [17] A.H. Ansari, P.J. Cherian, A. Caicedo, G. Naulaers, M. De Vos, S. Van Huffel, Neonatal seizure detection using deep convolutional neural networks, *Int. J. Neural Syst.* 29 (2019), 1850011, 04.
- [18] A. O'Shea, G. Lightbody, G. Boylan, A. Temko, Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture, *Neural Network.* 123 (2020) 12–25.
- [19] D.Y. Isaev, D. Tchapyjnikov, C.M. Cotten, D. Tanaka, N. Martinez, M. Bertran, G. Sapiro, D. Carlson, Attention-based network for weak labels in neonatal seizure detection, *Proc. Mach. Learn. Res.* 126 (2020) 479.
- [20] J. Gotman, D. Flanagan, B. Rosenblatt, A. Bye, E. Mizrahi, Evaluation of an automatic seizure detection method for the newborn EEG, *Electroencephalogr. Clin. Neurophysiol.* 103 (3) (1997) 363–369.
- [21] R. Lawrence, A. Mathur, S.N.T. Tich, J. Zempel, T. Inder, A pilot study of continuous limited-channel aEEG in term infants with encephalopathy, *J. Pediatr.* 154 (6) (2009) 835–841.
- [22] P. Cherian, W. Deburchgraeve, R. Swarte, M. De Vos, P. Govaert, S. Van Huffel, G. H. Visser, Validation of a new automated neonatal seizure detection system: a clinician's perspective, *Clin. Neurophysiol.* 122 (8) (2011) 1490–1499.

- [23] S.R. Mathieson, N.J. Stevenson, E. Low, W.P. Marnane, J.M. Rennie, A. Temko, G. Lightbody, G.B. Boylan, Validation of an automated seizure detection algorithm for term neonates, *Clin. Neurophysiol.* 127 (1) (2016) 156–168.
- [24] A. Temko, E. Thomas, W. Marnane, G. Lightbody, G. Boylan, Performance assessment for EEG-based neonatal seizure detectors, *Clin. Neurophysiol.* 122 (3) (2011) 474–482.
- [25] A.M. Pavel, J.M. Rennie, L.S. de Vries, M. Blennow, A. Foran, D.K. Shah, R. M. Pressler, O. Kapellou, E.M. Dempsey, S.R. Mathieson, et al., A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial, *Lancet Child Adolesc. Health* 4 (10) (2020) 740–749.
- [26] C. Sun, H. Cui, W. Zhou, W. Nie, X. Wang, Q. Yuan, Epileptic seizure detection with EEG textural features and imbalanced classification based on EasyEnsemble learning, *Int. J. Neural Syst.* 29 (10) (2019) 1950021.
- [27] K. Tapani, S. Vanhatalo, N. Stevenson, Incorporating spike correlations into an SVM-based neonatal seizure detector, *EMBC* 65 (2017) 322–325.
- [28] N. Stevenson, K. Tapani, L. Lauronen, S. Vanhatalo, A dataset of neonatal EEG recordings with seizure annotations, *Sci. Data* 6 (1) (2019) 1–8.
- [29] P. Nevalainen, M. Metsäranta, S. Toiviainen-Salo, T. Lönnqvist, S. Vanhatalo, L. Lauronen, Bedside neurophysiological tests can identify neonates with stroke leading to cerebral palsy, *Clin. Neurophysiol.* 130 (5) (2019) 759–766.
- [30] R.R. Clancy, A. Legido, The exact ictal and interictal duration of electroencephalographic neonatal seizures, *Epilepsia* 28 (5) (1987) 537–541.
- [31] N. Stevenson, K. Tapani, S. Vanhatalo, Hybrid neonatal EEG seizure detection algorithms achieve the benchmark of visual interpretation of the human expert, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 5991–5994.
- [32] J.H. Friedman, J.L. Bentley, R.A. Finkel, An algorithm for finding best matches in logarithmic expected time, *ACM Trans. Math Software* 3 (3) (1977) 209–226.
- [33] J. Cohen, A coefficient of agreement for nominal scales, *Educ. Psychol. Meas.* 20 (1) (1960) 37–46.
- [34] L.G. van Rooij, L. Hellström-Westas, L.S. de Vries, Treatment of neonatal seizures, *Semin. Fetal Neonatal Med.* 18 (2013) 209–215. Elsevier.
- [35] R. Wickström, B. Hallberg, M. Bartocci, Differing attitudes toward phenobarbital use in the neonatal period among neonatologists and child neurologists in Sweden, *Eur. J. Paediatr. Neurol.* 17 (1) (2013) 55–63.
- [36] N.J. Stevenson, L. Lauronen, S. Vanhatalo, The effect of reducing EEG electrode number on the visual interpretation of the human expert for neonatal seizure detection, *Clin. Neurophysiol.* 129 (1) (2018) 265–270.
- [37] L. Webb, M. Kauppila, J.A. Roberts, S. Vanhatalo, N. Stevenson, Automated detection of artefacts in neonatal EEG with residual neural networks, *Comput. Methods Progr. Biomed.* (2021), 106194.
- [38] N.J. Stevenson, R.R. Clancy, S. Vanhatalo, I. Rosén, J.M. Rennie, G.B. Boylan, Interobserver agreement for neonatal seizure detection using multichannel EEG, *Ann. Clin. Transl. Neurol.* 2 (11) (2015) 1002–1011.
- [39] R.A. Shellhaas, Neonatal seizures reach the mainstream: the ILAE classification of seizures in the neonate, *Epilepsia* 62 (3) (2021) 629–631.