IEEE Journal of Translational Engineering in Health and Medicine

Received 7 April 2022; revised 6 July 2022; accepted 18 August 2022.
Date of publication 23 August 2022; date of current version 30 August 2022.

Digital Object Identifier 10.1109/JTEHM.2022.3201167

Ensemble Learning Using Individual Neonatal Data for Seizure Detection

ANA BOROVAC[1,2], STEINN GUDMUNDSSON[1], GARDAR THORVARDSSON[2], SAEED M. MOGHADAM [3], (Graduate Student Member, IEEE), PÄIVI NEVALAINEN[3,4], NATHAN STEVENSON[5], (Member, IEEE), SAMPSA VANHATALO[3], AND THOMAS P. RUNARSSON[1]

[1]Faculty of Industrial Engineering, Mechanical Engineering and Computer Science, University of Iceland, 107 Reykjavik, Iceland
[2]Kvikna Medical ehf., 110 Reykjavik, Iceland
[3]BABA Center, Pediatric Research Center, Department of Physiology, University of Helsinki, 00014 Helsinki, Finland
[4]HUS Diagnostic Center, Epilepsia Helsinki and Department of Clinical Neurophysiology, New Children's Hospital, Helsinki University Hospital, 00029 Helsinki, Finland
[5]Brain Modelling Group, QIMR Berghofer Medical Research Institute, Herston, QLD 4006, Australia

CORRESPONDING AUTHOR: A. BOROVAC (anb48@hi.is)

This work was supported in part by the Sigrid Juselius Foundation, and in part by the European Union's Horizon 2020 Research and Innovation Programme under Agreement 813483.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Research Review Board of the HUS Diagnostic Center.

ABSTRACT Objective: Sharing medical data between institutions is difficult in practice due to data protection laws and official procedures within institutions. Therefore, most existing algorithms are trained on relatively small electroencephalogram (EEG) data sets which is likely to be detrimental to prediction accuracy. In this work, we simulate a case when the data can not be shared by splitting the publicly available data set into disjoint sets representing data in individual institutions. Methods and procedures: We propose to train a (local) detector in each institution and aggregate their individual predictions into one final prediction. Four aggregation schemes are compared, namely, the majority vote, the mean, the weighted mean and the Dawid-Skene method. The method was validated on an independent data set using only a subset of EEG channels. Results: The ensemble reaches accuracy comparable to a single detector trained on all the data when sufficient amount of data is available in each institution. Conclusion: The weighted mean aggregation scheme showed best performance, it was only marginally outperformed by the Dawid–Skene method when local detectors approach performance of a single detector trained on all available data. Clinical impact: Ensemble learning allows training of reliable algorithms for neonatal EEG analysis without a need to share the potentially sensitive EEG data between institutions.

INDEX TERMS Convolutional neural network, distributed learning, ensemble learning, neonatal EEG, seizure detection algorithm.

I. INTRODUCTION

Seizures are common during perinatal period [1], and management of neonatal seizures requires timely detection and treatment to reduce ensuing brain damage [2]. The current gold standard for neonatal seizure detection is visual analysis by a human expert using a full-montage video electroencephalogram (EEG) [3]. Since such service is rarely available in neonatal intensive care units (NICUs), there is an urgent clinical need for automated neonatal seizure detection algorithm (NSDA) with human expert level accuracy.

Early automated NSDAs were based on *features*, quantitative descriptors of short, e.g. $10-16$ sec long, EEG segments and expert-defined threshold decision rules [4], [5], [6]. Hard-coded thresholds were later replaced by statistical techniques, such as linear discriminant analysis [7], support vector machines (SVMs) [8], [9], [10] and neural networks [11]. Recently, promising results have been obtained using convolutional neural networks (CNNs) [12], [13], [14].

Deep neural networks (DNNs) generally require a large amount of training data [15]. However, building a large and diverse enough neonatal EEG data set with high quality seizure annotations is time consuming, ambiguous [16], [17] and often limited due to strict regulations (e.g. the Privacy Rule of the U.S. Health Insurance Portability and Accountability Act (HIPAA), or the European General Data Protection Regulation (GDPR)) making data sharing between

institutions difficult, if not impossible [18], [19]. Challenges in sharing data have triggered growing interest in distributed approaches to statistical learning [20].

One approach that requires minimal sharing of information is model ensembling, i.e. models are trained locally at each institution and predictions on new data are aggregated (ensembled) from predictions made by the local models. This requires sharing only the models across the network of institutions rather than sharing the potentially sensitive, original biosignals. However, the procedures in model sharing need to be planned so that they mitigate the impact of possible inadvertent leaks of training data through a model [21], [22]. One solution to this problem is to have a *trusted agent* in charge of the models and an aggregation procedure. Compared to the federated learning [23], ensembling does not require communication between the institutions during the training phase (which may be difficult to set up) and it does not require the institutions to use the same model architecture. One institution could e.g. use a DNN, another an SVM and a third a decision tree classifier.

Once predictions on new data have been made there are a number of techniques by which they can be ensembled. If predictions are accompanied by probabilities they can be averaged [24], [25], if not, a commonly used method for label aggregation is to simply select the most frequent label, referred to as *majority vote* in the following. One could also put more weight on some predictions if they are a priori more trustworthy, otherwise, an estimate of each annotator performance can be used [26], [27], [28]. Dawid and Skene [29] used an expected maximization (EM) algorithm [30] to estimate annotator performance and provide consensus labels.

Ensemble learning has previously been used in neonatal seizure detection. In [31] stacking is used where different model types trained on the same data are combined. In [32] three identical NSDAs are trained on the same EEG data but using labels from different experts. In this work we use ensemble learning on disjoint data sets, to simulate the situation where institutions train NSDAs on locally available data. Depending on the training data available at each institution and its similarity to new data to be labelled, the local NSDAs are expected to vary in performance. The main contribution and novelty of this work is in the discovery of how such locally trained models can be aggregated with the aim of achieving performance comparable to a single state-of-the-art NSDA trained on the union of all local training data sets. For aggregation we compared the majority vote, the mean, the weighted mean (via stacking) and the Dawid–Skene expected maximization algorithm. We show that the weighted mean outperforms the other methods if the NSDAs in the ensemble are trained on very few patients and Dawid-Skene marginally outperforms the other methods when the local NSDAs are not much worse than the state-of-the-art NSDA. The NSDAs and ensembles are further validated on an independent data set consisting of more than 2100 hours of EEG recorded from a small subset of the channels used to train the classifiers.

II. METHODS AND PROCEDURES

Multiple local models, referred to as *local NSDAs* in the following, are trained on disjoint subsets of multi-channel EEG recordings, simulating a scenario where several hospitals train NSDAs individually, without sharing patient data. The trained detectors are then shared with a trusted agent. To classify a short EEG segment from a new patient as seizure/non-seizure, the trusted agent sends the segment through all the local NSDAs and the predictions are aggregated using one of the following schemes: majority vote, mean, weighted mean or the Dawid–Skene method. The methodology is summarized in figure 1.

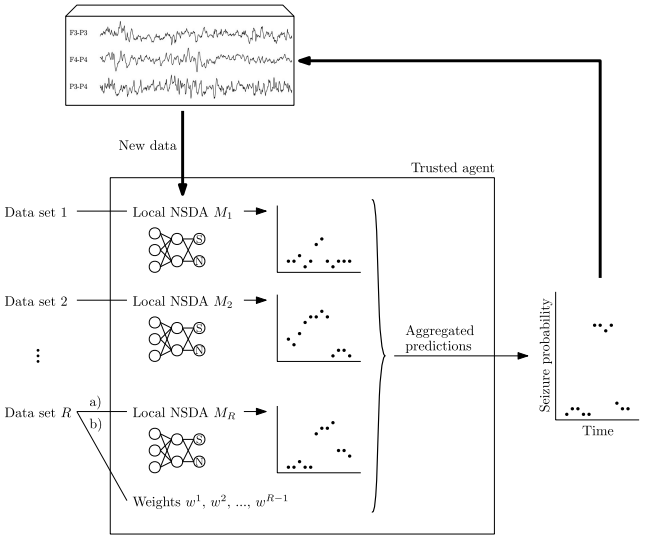

FIGURE 1. A schematic diagram of the proposed method. Each data set is used to train a local NSDAs or weights that are shared with a trusted agent. The trusted agent makes predictions on new data. Seizure predictions for new data are obtained a) by aggregating predictions made by R NSDAs using the majority vote, the mean or the Dawid-Skene method, or, b) by aggregating predictions made by $R-1$ local NSDAs using the weighted mean (weights are learned on the R^{th} data set).

For local NSDAs, we used DNNs which take EEG segments as input. The networks share the same architecture but have different network weights since they were trained on disjoint training sets.

A. AGGREGATION SCHEMES

In the following we consider a binary classification problem where the classes are labeled 0 and 1. Let D be a set of N predictions from R independent models

$$D = \left\{ \left(p_1^1, p_1^2, \ldots, p_1^R\right), \ldots, \left(p_N^1, p_N^2, \ldots, p_N^R\right) \right\},$$

where p_i^j is the estimated probability of model j of instance i belonging to class 1. By setting a threshold between the classes to 0.5, the predicted label of model j of instance i is given by

$$y_i^j = \begin{cases} 1; & \text{if } p_i^j \geq 0.5, \\ 0; & \text{otherwise.} \end{cases}$$

IEEE Journal of Translational Engineering in Health and Medicine

A simple way to aggregate multiple predictions for instance i, when models do not output their confidence (e.g. class probabilities), is to use majority vote, i.e. select the most frequent label. Here we use the mean of predicted labels,

$$\mu_i^{MV} = \frac{1}{R}\sum_{j=1}^{R} y_i^j; \quad i \in \{1, 2, \ldots, N\}. \tag{1}$$

When the models output class probabilities, which is e.g. the case when the models correspond to the neural networks, the predictions can be aggregated by taking the mean probability,

$$\mu_i^{M} = \frac{1}{R}\sum_{j=1}^{R} p_i^j; \quad i \in \{1, 2, \ldots, N\}. \tag{2}$$

As some of the models might perform better than others, a weighted mean can be used to emphasize the more accurate models. To get the final prediction in a range between 0 and 1, we used logistic regression,

$$\mu_i^{WM} = \sigma\left(\sum_{j=1}^{R} w^j p_i^j\right); \quad i \in \{1, 2, \ldots, N\}, \tag{3}$$

where $\sigma(x) = 1/(1+e^{-x})$. The weights for w^j are learned on a held out data set (see section II-D).

The fourth aggregation method evaluated here is the Dawid–Skene method. The method estimates the sensitivity and specificity of each model, together with consensus predictions μ^{DS}. For details of the method see appendix A. To predict the absence/presence of seizures from the above aggregation schemes, a threshold of 0.5 is used.

B. DATA

The EEG data used to train the NSDAs is a publicly available data set containing 79 approximately one hour long neonatal EEG recordings, measured with 19 Ag/AgCl electrodes positioned according to the 10-20 system [33]. An 18 channel montage is used, i.e. we derive channels Fp2-F4, F4-C4, C4-P4, P4-O2, Fp1-F3, F3-C3, C3-P3, P3-O1, Fp2-F8, F8-T4, T4-T6, T6-O2, Fp1-F7, F7-T3, T3-T5, T5-O1, Fz-Cz and Cz-Pz. The recordings are annotated by three EEG experts where each second in a recording is annotated as a seizure or non-seizure. We refer to this data set as 18-channel DS below.

The second, proprietary, data set (the 3-channel DS) consisting of EEG recordings of 28 neonates, is used as a held out test set to evaluate the aggregation schemes in a real world setting, i.e. detectors are trained on the 18-channel DS and tested on this data set. The data set is also used in [34] and is a subset of the data set used in [35]. Institutional Research Review Board of the HUS diagnostic center approved the use of this data, including a waiver of consent due to the study's retrospective and observational nature. Each recording spans from 19 hours to 7 days. The recordings were obtained using 4 needle electrodes (F3, F4, P3 and P4) with a common reference, instead of the full set of 19 electrodes used in the training data set. Neonatal recordings are typically performed with this reduced electrode set to allow easier maintenance in a long duration brain monitoring [36]. The three bipolar derivations (F3-P3, F4-P4 and P3-P4) are used for both two human expert annotators and as the detectors input.

Additional attributes of the data sets are given in table 2 in appendix B.

Each EEG recording is cut into 16 sec long segments with 12 sec overlap. Out of the 79 (28) recordings in 18-channel DS (the 3-channel DS), 38 (24) contain at least one seizure longer than 16 sec identified by three (two) human experts, meaning each of these recordings contain at least one consensus seizure segment. Segments containing more than 1 sec of zero voltage interval in at least one channel (disconnected electrode or pause in the recording) are left-out from the training and test sets. The signals are filtered with a 6th order Chebyshev Type 2 band-pass filter with cut-off frequencies of 0.5 Hz and 16 Hz, down-sampled to 32 Hz and rescaled to 16-bit integers. This is similar to the preprocessing in [10] and [13].

C. NEONATAL SEIZURE DETECTION ALGORITHM

Each NSDA is a neural network consisting of three components; a feature extractor, an attention layer and an output layer. The feature extractor is a CNN from [37]. The features are extracted from each EEG channel separately and are combined into a single feature channel by the attention layer [13]. The attention layer is used since expert labels are not specific to individual channels and neonatal seizures tend to be partial [3], i.e. localized in a small area of the brain and therefore only present in a subset of the recorded channels. The attention layer is also independent of the number of input feature channels making the detector independent of the number of recorded EEG channels. The output layer is a fully connected layer with two output nodes representing the two classes. A detailed description of the network architecture is given in appendix C.

To compare the aggregation schemes to current state-of-the-art NSDAs, we trained a neural network using all the recordings in the 18-channel DS containing at least one consensus seizure longer than 16 sec (P). This NSDA is referred to as the *baseline NSDA* in the following.

The local NSDAs use the same neural network architecture as the baseline NSDA but differ in the data used for training. The patients in P (patients containing a consensus seizure) are partitioned into $k = 3, 4, \ldots, 10$ subsets representing data sets in individual institutions. Partitioning is random such that each patient is in exactly one subset and there are at least three patients in every subset. The union of the k subsets is then P, the data set used as a training set for the baseline NSDA. By excluding patients without consensus seizures we ensure each subset has patients with seizures and eliminate the varying number of EEGs with normal brain activity in individual subsets, making the analysis more straightforward. As there can be a big difference between

the training set sizes, we obtain local NSDAs with different generalisation strengths and consequently with different performance strengths on unseen data. This is expected in practice. Even though the acquisition equipment is subject to international standards and the electrodes are positioned according to the 10-20 system, the EEG signals may vary considerably depending on the patient cohorts as the signals differ between neonates of different ages and conditions [38], [39]. Therefore, the detectors are expected to perform differently on unseen data.

D. TRAINING

After partitioning the training set, each NSDA (baseline NSDA and local NSDAs) is trained on 16 sec long EEG segments corresponding to the consensus seizures and non-seizure segments. To avoid complications due to class imbalance [13], [40], the training sets are balanced prior to training by sub-sampling the non-seizure segments. Segments with disagreements between the human experts and partly seizure/non-seizure segments are not included in the training sets. Cross entropy is used as the loss function. The Adam optimizer is used to optimize the network weights using an initial learning rate of 0.001 which is then halved every 10 epochs. The NSDAs are trained for 30 epochs with a mini-batch size of 32. Hyper-parameters, learning rate and number of epochs, are tuned empirically, from observing the behavior of the loss function during the training of the baseline NSDA. A small mini-batch size is chosen due to a small amount of data used in some local NSDAs. For the weighed mean aggregation scheme, the weights $w^j, j \in \{1, 2, \ldots, R\}$, are learned using a stacking classifier [28]. A logistic regression classifier is trained using the data from one randomly selected local NSDA in each experiment. This local NSDA is not used in an ensemble for making predictions on a test patient. Therefore, non-overlapping data sets are used for training the local NSDAs and the logistic regression classifier. Also, the training data of the local NSDAs would not need to be shared in practice as the input of the logistic regression classifier is just a set of seizure probabilities estimated by the local NSDAs and these can be provided by the trusted agent.

All the deep learning code used in the experiments is implemented using PyTorch 1.7.1 [41] and run on an NVIDIA GTX 1080 Ti GPU. For logistic regression, we use the scikit-learn [42] implementation with default hyper-parameters. The code is available at github.com/anaborovac/Distributed-NSDA.

E. PERFORMANCE

To avoid overlap between training and test data when evaluating classifier performance on the 18-channel DS, leave-one-subject-out cross-validation is used. This entailed training 38 baseline NSDAs, 38 sets of local NSDAs and 38 sets of logistic regression classifiers, leaving out data from one subject (patient) at a time. The experiment is repeated 10 times, resulting in $10 \cdot 38 \cdot (3+4+\cdots+10) = 19760$ local NSDAs and $10 \cdot 38 \cdot (1+1+\cdots+1) = 10 \cdot 38 \cdot 8 = 3040$ logistic regression classifiers.

Data from each left-out patient is sent through the corresponding baseline NSDA and local NSDAs. Predictions from the baseline NSDAs are compared to human expert labels to obtain performance metrics. Predictions from the local NSDAs are first aggregated using one of the aforementioned aggregation schemes: majority vote (1), mean (2), weighted mean (3) and the Dawid–Skene method (appendix A) to obtain the final predictions and these are then compared to human expert labels.

Two sets of performance metrics are calculated, metrics based on the success/failure in classifying individual 16 sec long segments, and event-based metrics which indicate whether a seizure is detected at all, or whether a seizure is falsely reported. The segment-based metrics are sensitivity (SE), specificity (SP) and the area under the receiver operating characteristic curve (AUC). These metrics are calculated from segments without disagreements between human experts and segments with either seizure either non-seizure activity for the whole segment duration. The event-based metrics are seizure detection rate (SDR), false detections per hour (FD/h) and the mean false detection duration (MFDD) [43]. A consensus seizure is considered to be detected if it is detected at any point in time and a seizure is considered as a false detection if it did not overlap with any (consensus or not) seizure labelled by the human experts. Definitions of the metrics are provided in appendix D. Metrics calculated on each patient separately are summarized by their means and medians.

Before the event-based metrics are calculated a post-processing step is in order since segments overlap. Besides a few segments at the beginning and end of each recording, for each 4 sec long segment there are 4 overlapping 16 sec long segments. Prediction for a 4 sec segment is obtained by averaging predictions from overlapping 16 sec long segments [44], [45]. Seizures with duration less than 10 sec are excluded and considered normal brain activity as by definition seizures are longer than 10 sec [46].

For studying the segment-based level of agreement between the local NSDAs we use Gwet's first-order agreement coefficient (AC1) [47]. Compared to the often used Cohen's (Fleiss') κ [13], [48], [49], Gwet's AC1 is less prone to the paradoxes associated with highly imbalanced data [50], [51].

Performance on the 3-channel DS is evaluated in the same manner as for the 18-channel DS, i.e. the metrics are calculated for each patient separately and then summarized with the mean and the median. The baseline NSDA is trained using all 38 patients in P (no patients are left-out), and the union of the training sets for the local NSDAs also contain all 38 patients in P. This results in additional $1 + 10 \cdot (3 + 4 + \cdots + 10) = 521$ NSDAs and $10 \cdot (1 + 1 + \cdots + 1) = 10 \cdot 8 = 80$ logistic regression classifiers.

IEEE Journal of Translational Engineering in Health and Medicine

III. RESULTS

To assess the clinical usefulness of the aggregation schemes they are compared to a baseline NSDA which is trained on data from all 38 patients in P (in a leave-one-subject-out setting for evaluation on the 18-channel DS). The baseline NSDA thus corresponds to the situation where a single agent has access to all the training data (P), a situation which is expected to be favorable compared to aggregating predictions from multiple models trained on disjoint subsets of the same data.

A. BASELINE NSDA

Table 1 compares the performance of the baseline detector to other NSDAs found in the literature. All detectors are neural networks and were trained or tested using the 18-channel DS. The difference between the mean (0.92) and median (0.98) AUC values for the baseline NSDA calculated on the 18-channel DS is mainly due to the presence of respiratory and heart rate artefacts and low seizure burden in some of the recordings.

TABLE 1. Comparison of the area under the curve (AUC) values found in the literature. Each reference uses a different proprietary data set. All NSDAs, except [13], were trained using the 18-channel DS. Superscript L denotes leave-one-subject-out testing and superscript C denotes AUC value on concatenated recordings from the data set.

		AUC	
		18-channel DS	Proprietary DS
Isaev et al. [13]	mean	0.92	0.97^{L}
O'Shea et al. [14]	mean	0.96^{C}	0.99^{L}
Stevenson et al. [49]	median	0.99^{L}	
Baseline NSDA	median	0.98^{L}	0.93
	mean	0.92^{L}	0.92

The performance of an NSDA on an independent test set is usually worse than performance estimates obtained from a held out training data. Such a decrease can be attributed to several factors, including differences in patient cohorts, seizure prevalence, the number of available EEG channels, the human experts that annotated the EEG [48], and training data not representing the general population. For example, the mean AUC decreased from 0.97 to 0.92 in [13] and from 0.99 to 0.96 in [14]. We observe a similar drop in performance when the baseline detector was tested on a proprietary the 3-channel DS. Detailed validation of the NSDA performance is available in table 3 in appendix E.

In summary, the baseline NSDA gives comparable results to the state-of-the-art NSDAs and performs well on recordings which include only a small subset of the channels used in training.

B. AGGREGATION SCHEMES

Here we evaluate the different aggregation schemes and compare them to the baseline NSDA and to the average performance of the local NSDAs. If the baseline performance can be reached with an aggregation scheme, it would indicate that the data does not need to be shared during the training of an NSDA to obtain a detector with state-of-the-art performance. The four aggregation schemes, majority vote, mean, weighted mean and the Dawid–Skene method were evaluated on the 18-channel DS and the 3-channel DS for $k = 3, 4, \ldots, 10$ local NSDAs. Results for the majority vote are not shown since in all cases majority vote was slightly outperformed by the mean aggregation scheme (see figure 7 in appendix E).

With an increasing number of local NSDAs the average performance of an individual detector gradually gets worse (figure 2). This is explained by the fact that the number of patients behind each local NSDA is becoming smaller since the total number of patients in the combined training sets is constant (37 for the 18-channel DS and 38 for the 3-channel DS). Consequently there is an increased risk of overfitting in individual detectors. The size of the local training sets is quantified with the mean median number of patients in the training set. E.g., if four local NSDAs are used and the mean median is 8.1, then on average there are at least nine patients in the training of two of the local NSDAs.

Figure 2 shows that the AUC, seizure detection rate and false detection rate behave similarly across both data sets for all the aggregation schemes, but there is considerably more variability for the 3-channel DS. All the aggregation schemes give AUC values that are similar to the baseline value. However, the aggregation schemes differ in terms of seizure detection rate and false detections per hour.

Figure 3 shows the seizure probability estimates returned by local NSDAs for an hour-long recording, together with probability estimates obtained with the ensemble methods. All the aggregation schemes result in AUC close to one, although they detect only 3 out of 7 consensus seizures. The missed seizures are short in duration and they are clearly visible in the figure (as white bands) since the corresponding probabilities are higher than for the non-seizure segments.

The SDR in figure 2 behaves similarly for both data sets. For all values of k tested, the Dawid–Skene method is comparable to the baseline NSDA, while for the mean and the weighted mean aggregation schemes, fewer seizures were detected with an increased number of local NSDAs. Recall that when there are few NSDAs, each NSDA detects almost as many seizures as the baseline detector. The mean aggregation scheme performed slightly worse than the weighted mean and both performed notably worse than the Dawid–Skene method for more than four local detectors. Moreover, the average SDR of the local NSDAs is comparable to the values corresponding to the mean aggregation scheme. With the weighted mean a larger number of seizures are detected for $k \geq 8$ ($k = 10$) on the 18-channel DS (3-channel DS), for smaller k the mean and the weighted mean aggregation schemes return comparable seizure detection rates.

Moreover, in figure 2 we observe that all aggregation schemes result in a lower number of FD/h than the average local NSDA. The average FD/h of the local NSDAs are noticeably higher for the 3-channel DS than for the

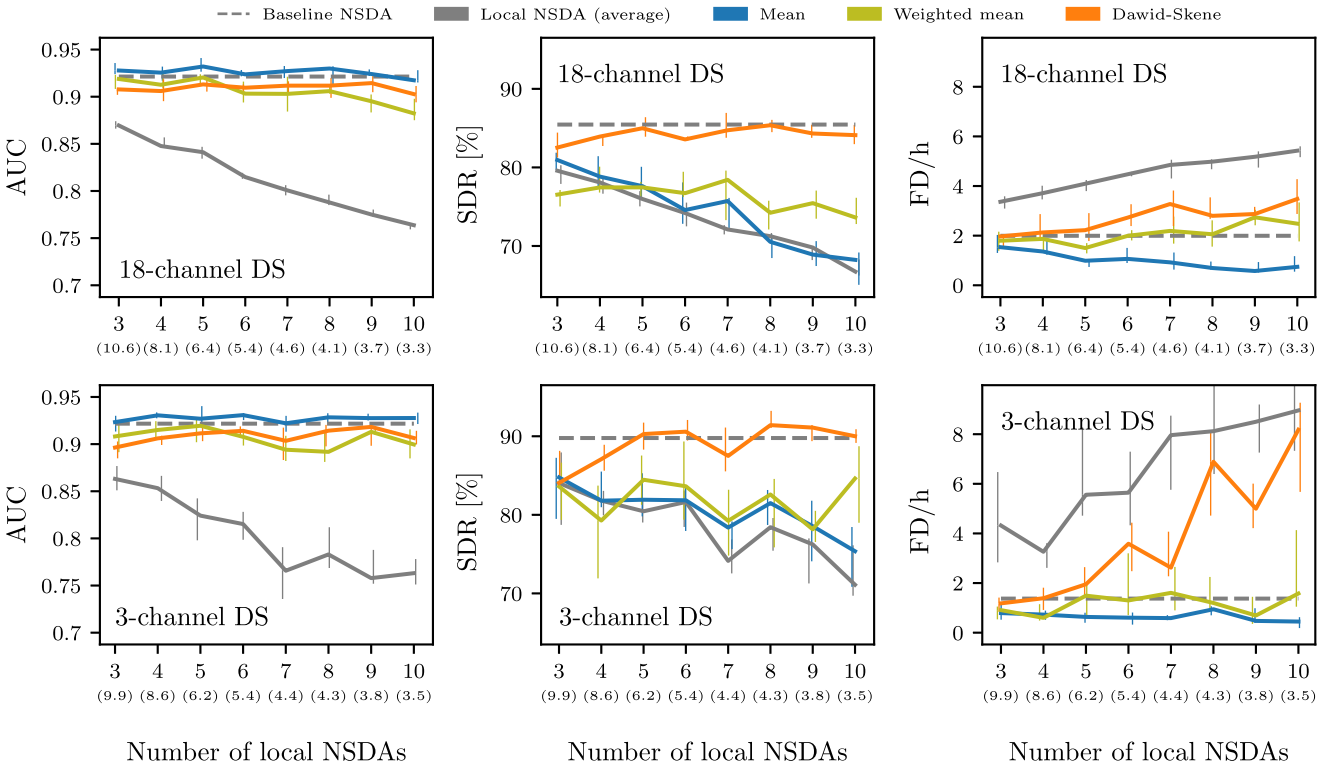

FIGURE 2. Average area under the curve (AUC), seizure detection rate (SDR) and false detections per hour (FD/h) as a function of the number of local NSDAs used in the aggregation schemes. The solid lines represent the medians of ten runs together with interquartile ranges denoted with vertical lines. The grey dashed line represents the average metric of the baseline NSDA. The average (across ten runs) mean median number of patients in each NSDA is shown in parentheses.

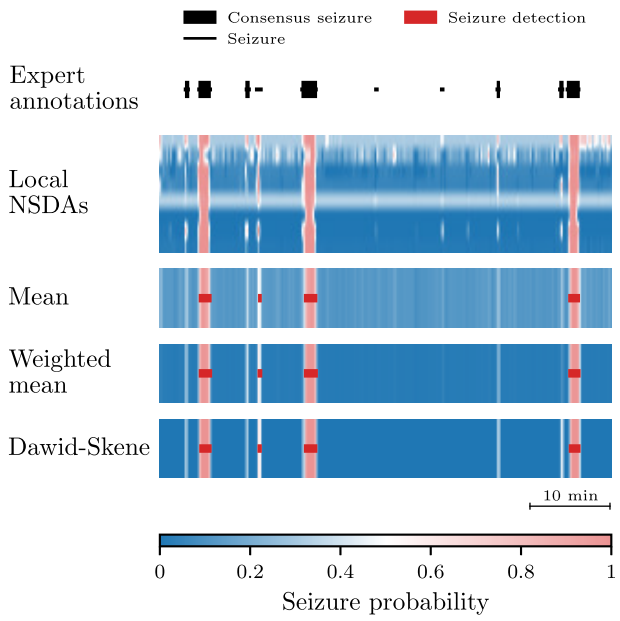

FIGURE 3. An example of aggregated predictions from eight local NSDAs. The area under the curve is 1.0 for the mean and the weighted mean and 0.99 for the Dawid–Skene method. All aggregation schemes detect 42.9 % of consensus seizures and they do not falsely detect any seizure.

18-channel DS. One possible explanation is that the recordings in the 3-channel DS are much longer and on average just 3.5 % of a recording corresponds to a seizure activity. The mean aggregation scheme has a lower false detection rate than the baseline NSDA and the FD/h decreases steadily with increasing number of local NSDAs. This may be a result of low level of agreement between the local NSDAs for the large k (figure 5 in appendix E). So, even though an individual local NSDA falsely detects a large number of seizures, the aggregated prediction filtered them out or was below the 0.5 threshold. This may on the other hand caused problems with the Dawid–Skene method, i.e. the FD/h increased slowly on the 18-channel DS and rapidly on the 3-channel DS with increasing number of local NSDAs. In contrast, the logistic regression classifier determining the weights for the weighted mean aggregation scheme successfully detected local NSDAs with high/low false detection rate for all k tested.

We observed low false detection rates for the mean and weighted mean aggregation schemes and therefore investigated whether the false detections are short or long in duration. We did not observe big differences between the aggregation schemes (10 - 30 sec) and different values of local NSDAs (figure 6 in appendix E).

To summarise, all aggregation schemes tested here perform better than the average local NSDA and are comparable to the baseline NSDA for $k \in \{3, 4\}$. This shows that the overfitting by local models noted earlier is offset by aggregating their predictions. This is in line with published reports on ensemble methods such as Random Forests which aggregate predictions from multiple models individually overfitting the data. The decrease in performance for

IEEE Journal of Translational Engineering in Health and Medicine

larger values of k is mainly a result of training the local NSDAs on smaller training sets that do not capture the general population. The (weighted) mean aggregation scheme detects fewer seizures than the baseline detector, however the false detection rate is comparable, if not lower. The Dawid–Skene method successfully detects the same number of seizures as the baseline NSDA for any number of local NSDAs, but the false detection rate is compromised for $k \geq 6$. Predictions obtained with the Dawid–Skene are difficult to explain [52], [53], only a few local NSDAs with poor performance may have caused unexpected and undesired aggregated prediction [54].

IV. CONCLUSION

In this work we have shown that an NSDA based on a convolutional neural network together with an attention layer can accurately detect seizures, even if the data is obtained with different types of electrodes (scalp vs needle) and significantly lower number of channels than it was used for training. All the performance metrics of the NSDAs unsurprisingly dropped when training sets contained data from only a few patients. For aggregation of such NSDAs the weighted mean aggregation scheme performed best. Compared to the Dawid–Skene method, it successfully detected local NSDAs with high false detection rates and seizure detection rate was not as compromised as it was for the mean aggregation scheme. When a larger number of patients was included in the training of individual local NSDAs, i.e. when the number of local NSDAs was few, the Dawid–Skene method marginally outperformed the other aggregation schemes. It had a higher seizure detection rate and the false detections per hour was comparable to the (weighted) mean aggregation scheme. Independent of the number of local NSDAs, the majority vote was slightly outperformed by the mean aggregation scheme and all aggregation schemes performed better than the average individual (local) NSDA.

The experiments suggest that data does not need to be shared between institutions. It takes approx. 15 seconds to process one hour of 18-channel EEG with 10 local detectors, which is fast enough to be used in an online setting in the clinic. By utilizing GPU optimized code in the preprocessing steps and a fast version of the Dawid-Skene aggregation method [55], one hour of EEG could be processed in less than 2 seconds.

To confirm the findings reported here in a real-world setting, data from multiple institutions would be required. A large data set would also allow a detailed study on the number of local NSDAs needed to reach the desirable classification performance and whether a mixture of different types of NSDAs improves or degrades the overall performance.

APPENDIX A
DAWID-SKENE METHOD

The Dawid–Skene method was initially used to estimate the performance of human annotators [29]. Here the method is used to estimate the performance of models (local NSDAs) and obtain consensus judgement amongst them. The method is as follows. From a given set D of model predictions, the task is to estimate consensus labels $\{\mu_i\}_{i=1}^{N}$, the sensitivity α^j and specificity β^j of predictive model $j \in \{1, 2, \ldots, R\}$. Let Y_e denote the multivariate random variable

$$Y_e = (Y_1^1, Y_1^2, \ldots, Y_1^R, \ldots, Y_N^1, Y_N^2, \ldots, Y_N^R),$$

where random variable Y_i^j denotes the label given to instance i by model j. Furthermore, let T_i denote a random variable corresponding to the true label of instance i for which

$$P[T_i = 1] = t_i = t; \quad i \in \{1, 2, \ldots, N\}.$$

Assuming that model labels are independent and that conditional probability of Y_i^j on T_i follows Bernoulli distribution with parameters α^j and β^j, respectively:

$$\begin{aligned} a_i &= P_\alpha\left[Y_i^1, Y_i^2, \ldots, Y_i^R | T_i = 1\right] \\ &= \prod_{j=1}^{R} (\alpha^j)^{y_i^j} (1-\alpha^j)^{1-y_i^j}; \quad i \in \{1, 2, \ldots, N\}, \\ b_i &= P_\beta\left[Y_i^1, Y_i^2, \ldots, Y_i^R | T_i = 0\right] \\ &= \prod_{j=1}^{R} (\beta^j)^{1-y_i^j} (1-\beta^j)^{y_i^j}; \quad i \in \{1, 2, \ldots, N\}. \end{aligned}$$

To simplify the notation, let $\theta = (t, \alpha, \beta)$ denote the parameters to be estimated. Assuming that instances are sampled independently, the likelihood function for Y_e is [29], [56]:

$$\begin{aligned} P_\theta[Y_e] &= \prod_{i=1}^{N} P_\theta[Y_i^1, Y_i^2, \ldots, Y_i^R] \\ &= \prod_{i=1}^{N} \Bigg(\underbrace{P_\theta[Y_i^1, Y_i^2, \ldots, Y_i^R | T_i = 1]}_{a_i} \underbrace{P_\theta[T_i = 1]}_{t} \\ &\qquad + \underbrace{P_\theta[Y_i^1, Y_i^2, \ldots, Y_i^R | T_i = 0]}_{b_i} \underbrace{P_\theta[T_i = 0]}_{1-t} \Bigg) \\ &= \prod_{i=1}^{N} \left(a_i t + b_i (1-t)\right). \end{aligned} \tag{4}$$

Dawid and Skene used the EM algorithm to identify a local maximum of the likelihood function. The true labels are estimated by maximizing the likelihood function using estimated values for the sensitivity and specificity of each annotator, and the prior probability of class 1 (t), i.e. seizure. The algorithm has two main steps [29].

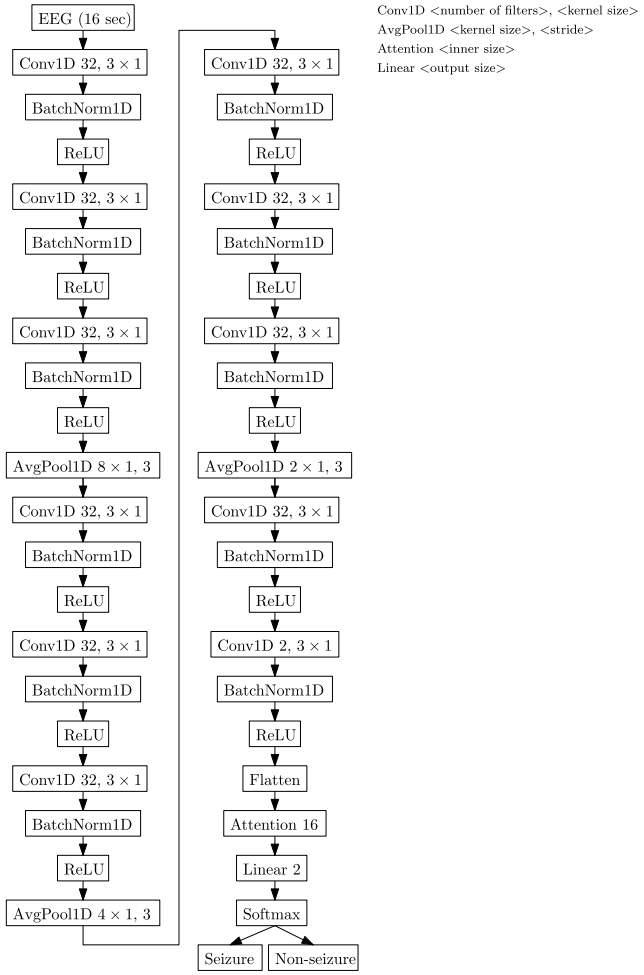

FIGURE 4. **Architecture of the NSDA with a total of 29352 learnable parameters. Other parameters were set to default PyTorch values.**

Expectation step: calculate the expected value of a true label knowing labels made by predictive models,

$$\begin{aligned}\mu_i &= \mathbb{E}[T_i|Y_i^1, Y_i^2, \ldots, Y_i^R] \\ &= P_\theta[T_i = 1|Y_i^1, Y_i^2, \ldots, Y_i^R] \\ &= \frac{P_\theta[Y_i^1, Y_i^2, \ldots, Y_i^R|T_i = 1]P_\theta[T_i = 1]}{P_\theta[Y_i^1, Y_i^2, \ldots, Y_i^R]} \\ &\qquad \text{(Bayes' theorem)} \\ &= \frac{a_i t}{a_i t + b_i(1-t)}; \quad i \in \{1, 2, \ldots, N\}. \end{aligned} \tag{5}$$

Maximization step: estimate t, α^j and β^j that maximize the likelihood function (4),

$$t = \frac{\sum_{i=1}^N \mu_i}{N}, \tag{6}$$

$$\alpha^j = \frac{\sum_{i=1}^N \mu_i y_i^j}{\sum_{i=1}^N \mu_i}; \quad j \in \{1, 2, \ldots, R\}, \tag{7}$$

$$\beta^j = \frac{\sum_{i=1}^N (1-\mu_i)(1-y_i^j)}{\sum_{i=1}^N (1-\mu_i)}; \quad j \in \{1, 2, \ldots, R\}. \tag{8}$$

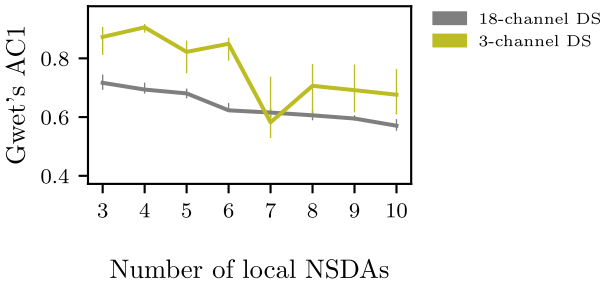

FIGURE 5. **Average Gwet's AC1 between local NSDAs for 18-channel DS and 3-channel DS. The solid lines represent the medians of ten runs together with interquartile ranges denoted with vertical lines.**

TABLE 2. **A summary of the data sets used in the study. Numbers inside parentheses represent standard deviation. Means for recordings are calculated across patients containing at least one consensus seizure longer than 16 sec (duration of one EEG segment).**

	18-channel DS	3-channel DS
Number of patients	79	28
Number of patients with consensus seizures $\geq$ 16 sec	38	24
Gestational age (weeks)	39.3 (2.1)	39.2 (2.0)
Number of derived EEG channels	18	3
Total recordings duration (hours)	111.9	2149.4
Mean recording duration (hours)	1.4 (0.6)	76.4 (35.8)
Total number of consensus seizures	344	1387
Total duration of consensus seizures (hours)	11.0	65.3
Mean duration of consensus seizures (minutes)	1.9 (2.7)	2.8 (6.0)
Mean fraction of recording containing seizures (%)	31.8 (26.4)	5.3 (5.6)
Mean fraction of recording containing consensus seizures (%)	19.1 (20.9)	3.5 (4.0)

In the special case when all the μ_i's are either 0 or 1, then t is the estimated ratio of positive instances and α^j (β^j) is an estimated ratio of correctly predicted positive (negative) examples by expert j, i.e. the estimated sensitivity (specificity) of expert j.

Input: $D, \epsilon = 10^{-5}, k_{max} = 5000$
Output: μ^{DS}
initialize $\mu^{DS} = \mu^M$
compute $\theta^{(0)}$ using equations (6), (7) and (8)
$k = 0$
repeat
k = k + 1
compute μ^{DS} using equation (5)
compute $\theta^{(k)}$ using equations (6), (7) and (8)
until $|\log P_{\theta^{(k-1)}}[Y_e] - \log P_{\theta^{(k)}}[Y_e]| < \epsilon \quad or \quad k \geq k_{max}$

APPENDIX B
DATA INFORMATION

See table 2.

IEEE Journal of Translational Engineering in Health and Medicine

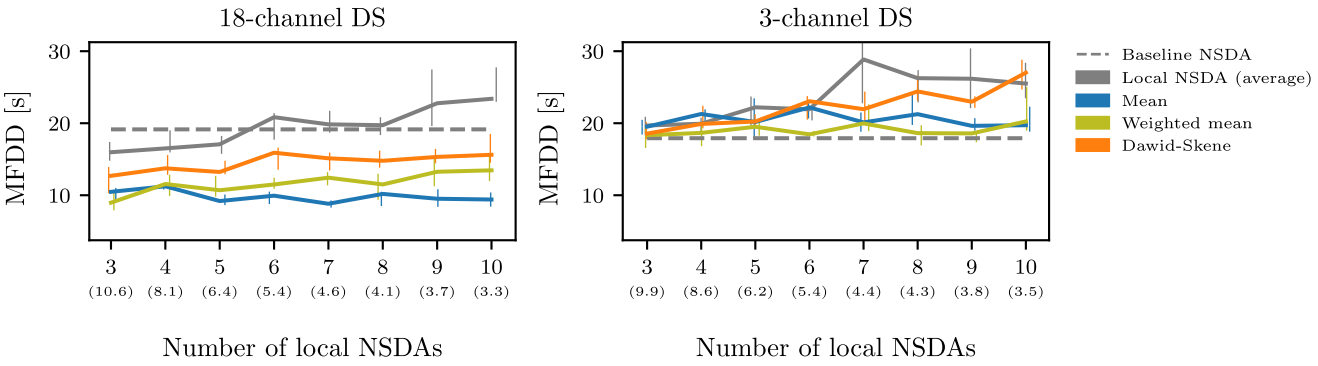

FIGURE 6. Average mean false detection duration (MFDD) as a function of the number of local NSDAs used in the aggregation schemes. The solid lines represent the medians of ten runs together with interquartile ranges denoted with vertical lines. The grey dashed line represents the average MFDD of the baseline NSDA.

APPENDIX C
ARCHITECTURE OF THE NSDA

In this work the NSDAs are deep neural networks consisted of three components, a feature extractor [37], an attention layer [13] and an output layer (figure 4). We used PyTorch implementation of layers for the feature extractor and for the output layer. Using PyTorch notation, the attention layer was implemented as follows. If an input to the attention layer is of size (N, C_{in}, L) then the output is of size (N, L) and can be described as

$$\text{out}(N_i) = \sum_{k=0}^{C_{in}-1} a_k \text{input}(N_i, k);$$

$$a_k = \frac{\exp\left(w^T \tanh\left(V\text{input}(N_i, k)^T\right)\right)}{\sum_{j=0}^{C_{in}-1} \exp\left(w^T \tanh\left(V\text{input}(N_i, j)^T\right)\right)},$$

where $V \in \mathbb{R}^{L\times<\text{inner size}>}$ and $w \in \mathbb{R}^{L\times 1}$ are learnable parameters.

APPENDIX D
PERFORMANCE METRICS

A. SEGMENT-BASED METRICS

Segment-based metrics were calculated based on 16 sec long EEG segments. A true positive (TP) is a correctly predicted seizure segment, a true negative (TN) is a correctly predicted non-seizure segment, a false positive (FP) is an incorrectly predicted non-seizure segment and a false negative (FN) is an incorrectly predicted seizure segment.

- Sensitivity (ratio of correctly predicted seizure intervals):

$$\text{SE} = \frac{\text{TP}}{\text{TP} + \text{FN}} \cdot 100.$$

- Specificity (ratio of correctly predicted non-seizure intervals):

$$\text{SP} = \frac{\text{TN}}{\text{TN} + \text{FP}} \cdot 100.$$

- Area under the receiver operating characteristics curve (AUC). The receiver operating characteristics curve describes SE depending on 1-SP.

TABLE 3. Accuracy of the baseline model. Area under the curve (AUC), sensitivity (SE), specificity (SP), seizure detection rate (SDR), false detections per hour (FD/h) and mean false detection duration (MFDD) are computed as the mean and median over all the patients with seizures.

		Segment-based metrics		
		AUC	SE [%]	SP [%]
18-channel DS	median	0.98	90.46	97.21
	mean	0.92	79.52	93.69
3-channel DS	median	0.93	78.00	98.23
	mean	0.92	70.54	97.40
		Event-based metrics		
		SDR [%]	FD/h	MFDD [s]
18-channel DS	median	100.0	0.91	12.00
	mean	85.45	1.99	19.15
3-channel DS	median	95.55	0.97	15.82
	mean	89.77	1.37	17.92

B. EVENT-BASED METRICS

Event-based metrics are in comparison with the segment-based metrics focused on each predicted seizure and not just 16 sec long segments. Three event-based metrics were used [43]:

- Seizure detection rate (SDR):

$$\text{SDR} = \frac{\text{DS}}{\text{CS}} \cdot 100,$$

where DS is a number of detected consensus seizures and CS is a number of consensus seizures. A seizure was considered to be detected if it was detected at any time of its duration.

- False detections per hour (FD/h):

$$\text{FD/h} = \frac{\text{IDS}}{\text{D}},$$

where IDS is a number of incorrectly detected seizures and D is duration of data in hours. A seizure was considered to be incorrectly detected if it was not overlapping with any seizure annotated by the experts.

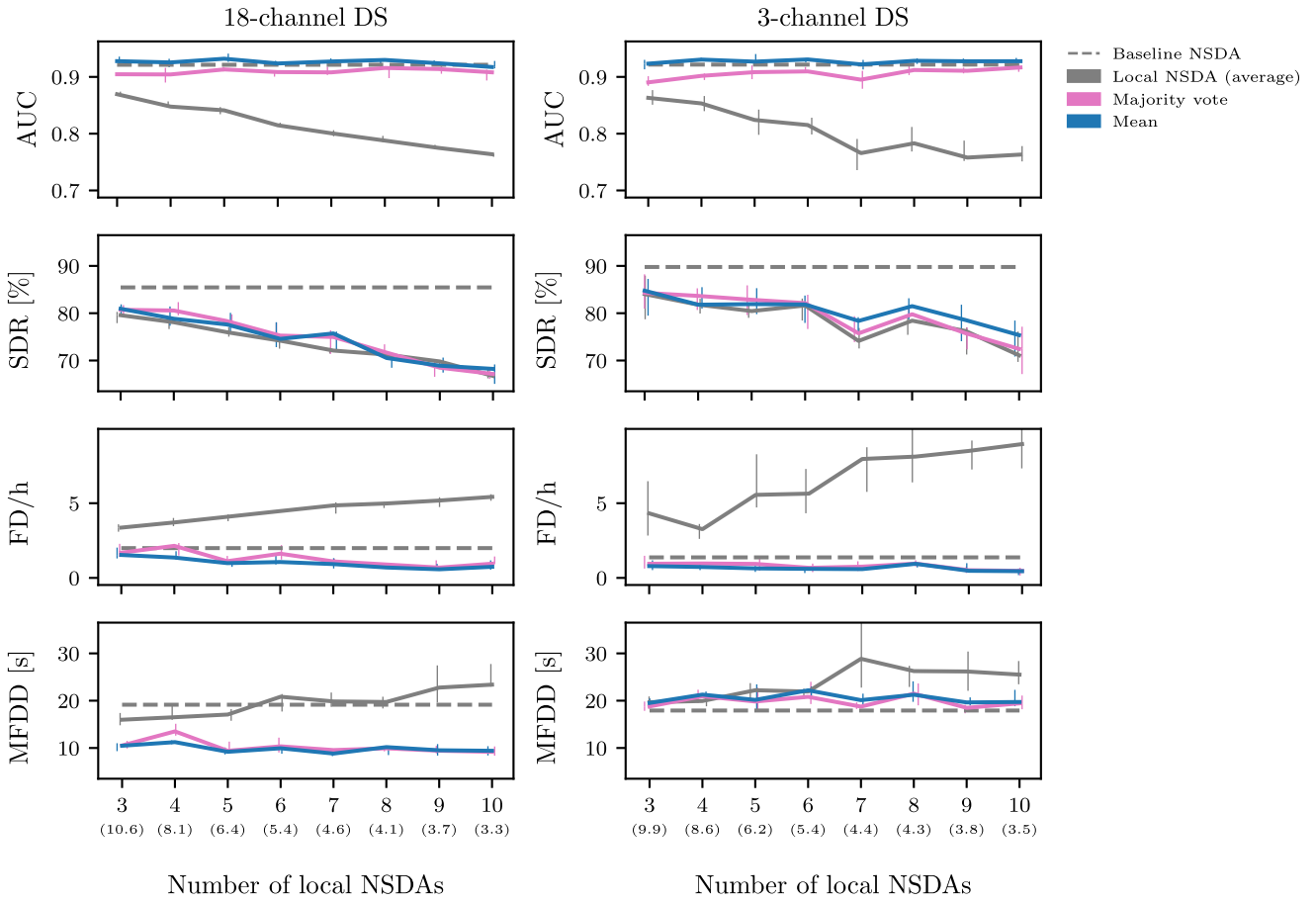

FIGURE 7. Average area under the curve (AUC), seizure detection rate (SDR), false detections per hour (FD/h) and false detection duration (MFDD) as a function of the number of local NSDAs used in the aggregation schemes. The solid lines represent the medians of ten runs together with interquartile ranges denoted with vertical lines. The grey dashed line represents the average metric of the baseline NSDA. The average (across ten runs) mean median number of patients in each NSDA is shown in parentheses.

- Mean false detection duration (MFDD):

$$\text{MFDD} = \begin{cases} 0; & \text{if IDS} = 0 \\ \frac{\text{DIDS}}{\text{IDS}}; & \text{otherwise} \end{cases},$$

where DIDS is a sum of durations of incorrectly detected seizures in seconds and IDS is a number of incorrectly detected seizures.

APPENDIX E
ADDITIONAL RESULTS

See table 3 and figures. 6 and 7.

REFERENCES

[1] H. C. Glass *et al.*, "Risk factors for EEG seizures in neonates treated with hypothermia: A multicenter cohort study," *Neurology*, vol. 82, no. 14, pp. 1239–1244, Apr. 2014.

[2] S. T. Björkman, S. M. Miller, S. E. Rose, C. Burke, and P. B. Colditz, "Seizures are associated with brain injury severity in a neonatal model of hypoxia–ischemia," *Neuroscience*, vol. 166, no. 1, pp. 157–167, Mar. 2010.

[3] R. M. Pressler *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*, vol. 62, no. 3, pp. 615–628, Mar. 2021.

[4] 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, 1992.

[5] S. Nagasubramanian, B. Onaral, and R. Clancy, "On-line neonatal seizure detection based on multi-scale analysis of EEG using wavelets as a tool," in *Proc. 19th Annu. Int. Conf. IEEE Eng. Med. Biol. Society. Magnificent Milestones Emerg. Opportunities Med. Eng.*, vol. 3, Jun. 1997, pp. 1289–1292.

[6] 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.

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

[8] R. Ahmed, A. Temko, W. P. Marnane, and G. Lightbody, "Exploring temporal information in neonatal seizures using a dynamic time warping based SVM kernel," *Comput. Biol. Med.*, vol. 82, pp. 100–110, Mar. 2017.

[9] 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.

[10] 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.

[11] H. Hassanpour, M. Mesbah, and B. Boashash, "Time–frequency based newborn EEG seizure detection using low and high frequency signatures," *Physiol. Meas.*, vol. 25, no. 4, p. 935, 2004.

[12] A. H. Ansari, P. J. Cherian, A. Caicedo, G. Naulaers, M. De Vos, and S. Van Huffel, "Neonatal seizure detection using deep convolutional neural networks," *Int. J. Neural Syst.*, vol. 29, no. 4, May 2019, Art. no. 1850011.

[13] D. Y. Isaev *et al.*, "Attention-based network for weak labels in neonatal seizure detection," *Proc. Mach. Learn. Res.*, vol. 126, p. 479, Aug. 2020.

[14] A. O'Shea, G. Lightbody, G. Boylan, and A. Temko, "Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture," *Neural Netw.*, vol. 123, pp. 12–25, Mar. 2020.

[15] Y. Bengio, I. Goodfellow, and A. Courville, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2016.

[16] A. Malone, C. A. Ryan, A. Fitzgerald, L. Burgoyne, S. Connolly, and G. B. Boylan, "Interobserver agreement in neonatal seizure identification," *Epilepsia*, vol. 50, no. 9, pp. 2097–2101, Sep. 2009.

[17] N. J. Stevenson, R. R. Clancy, S. Vanhatalo, I. Rosén, J. M. Rennie, and G. B. Boylan, "Interobserver agreement for neonatal seizure detection using multichannel EEG," *Ann. Clin. Translational Neurol.*, vol. 2, no. 11, pp. 1002–1011, Nov. 2015.

[18] J. Eicher, R. Bild, H. Spengler, K. A. Kuhn, and F. Prasser, "A comprehensive tool for creating and evaluating privacy-preserving biomedical prediction models," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, pp. 1–14, Dec. 2020.

[19] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, "Federated learning," *Synth. Lectures Artif. Intell. Mach. Learn.*, vol. 13, no. 3, pp. 1–207, 2019.

[20] M. Kirienko *et al.*, "Distributed learning: A reliable privacy-preserving strategy to change multicenter collaborations using AI," *Eur. J. Nucl. Med. Mol. Imag.*, vol. 48, no. 12, pp. 1–14, 2021.

[21] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 1322–1333.

[22] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The secret revealer: Generative model-inversion attacks against deep neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 253–261.

[23] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[24] K. Chang *et al.*, "Distributed deep learning networks among institutions for medical imaging," *J. Amer. Med. Inform. Assoc.*, vol. 25, pp. 945–954, Aug. 2018.

[25] A. Tuladhar, S. Gill, Z. Ismail, and N. D. Forkert, "Building machine learning models without sharing patient data: A simulation-based analysis of distributed learning by ensembling," *J. Biomed. Informat.*, vol. 106, Jun. 2020, Art. no. 103424.

[26] D. Tao, J. Cheng, Z. Yu, K. Yue, and L. Wang, "Domain-weighted majority voting for crowdsourcing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 163–174, Jan. 2018.

[27] T. Tian and J. Zhu, "Max-margin majority voting for learning from crowds," in *Proc. Nips*, 2015, pp. 1621–1629.

[28] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.

[29] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *J. Roy. Statist. Soc. C, Appl. Statist.*, vol. 28, no. 1, pp. 20–28, 1979.

[30] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B, Methodol.*, vol. 39, no. 1, pp. 1–22, 1977.

[31] Y. Pan, H. Li, L. Liu, Q. Li, X. Hou, and B. Dong, "AEEG signal analysis with ensemble learning for newborn seizure detection," in *Proc. Int. Workshop Multiscale Multimodal Med. Imag.* Cham, Switzerland: Springer, 2019, pp. 76–84.

[32] M. A. Tanveer, M. J. Khan, H. Sajid, and N. Naseer, "Convolutional neural networks ensemble model for neonatal seizure detection," *J. Neurosci. Methods*, vol. 358, Jul. 2021, Art. no. 109197.

[33] N. J. Stevenson, K. Tapani, L. Lauronen, and S. Vanhatalo, "A dataset of neonatal EEG recordings with seizure annotations," *Sci. Data*, vol. 6, no. 1, Mar. 2019, Art. no. 190039.

[34] K. T. Tapani, P. Nevalainen, S. Vanhatalo, and N. J. Stevenson, "Validating an SVM-based neonatal seizure detection algorithm for generalizability, non-inferiority and clinical efficacy," 2022, *arXiv:2202.12023*.

[35] P. Nevalainen, M. Metsäranta, S. Toiviainen-Salo, T. Lönnqvist, S. Vanhatalo, and L. Lauronen, "Bedside neurophysiological tests can identify neonates with stroke leading to cerebral palsy," *Clin. Neurophysiol.*, vol. 130, no. 5, pp. 759–766, May 2019.

[36] G. B. Boylan, N. J. Stevenson, and S. Vanhatalo, "Monitoring neonatal seizures," *Seminars Fetal Neonatal Med.*, vol. 18, no. 4, pp. 202–208, Aug. 2013.

[37] A. OrShea, G. Lightbody, G. Boylan, and A. Temko, "Investigating the impact of CNN depth on neonatal seizure detection performance," in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 5862–5865.

[38] R. A. Hrachovy and E. M. Mizrahi, *Atlas of Neonatal Electroencephalography*. Cham, Switzerland: Springer, 2015.

[39] A. M. Husain, "Review of neonatal EEG," *Amer. J. Electroneurodiagnostic Technol.*, vol. 45, no. 1, pp. 12–35, Mar. 2005.

[40] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[41] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, H. Wallach, H. Larochelle, A. Beygelzimer, F. D. Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035.

[42] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Jan. 2011.

[43] 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.

[44] T. M. Ingolfsson *et al.*, "Towards long-term non-invasive monitoring for epilepsy via wearable EEG devices," 2021, *arXiv:2106.08008*.

[45] U. Pale, T. Teijeiro, and D. Atienza, "Systematic assessment of hyperdimensional computing for epileptic seizure detection," 2021, *arXiv:2105.00934*.

[46] T. N. Tsuchida *et al.*, "American clinical neurophysiology society standardized EEG terminology and categorization for the description of continuous EEG monitoring in neonates: Report of the American clinical neurophysiology society critical care monitoring committee," *J. Clin. Neurophysiol.*, vol. 30, no. 2, pp. 161–173, 2013.

[47] K. L. Gwet, *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Oxford, U.K.: Advanced Analytics, 2014.

[48] A. Borovac, S. Gudmundsson, G. Thorvardsson, and T. P. Runarsson, "Influence of human-expert labels on a neonatal seizure detector based on a convolutional neural network," in *Proc. NeurIPS*, Dec. 2021, pp. 1–5.

[49] N. Stevenson, K. Tapani, and S. Vanhatalo, "Hybrid neonatal EEG seizure detection algorithms achieve the benchmark of visual interpretation of the human expert," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 5991–5994.

[50] A. R. Feinstein and D. V. Cicchetti, "High agreement but low Kappa: I. The problems of two paradoxes," *J. Clin. Epidemiol.*, vol. 43, no. 6, pp. 543–549, Jan. 1990.

[51] N. Wongpakaran, T. Wongpakaran, D. Wedding, and K. L. Gwet, "A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples," *BMC Med. Res. Methodol.*, vol. 13, no. 1, pp. 1–7, Dec. 2013.

[52] S. Ibrahim, X. Fu, N. Kargas, and K. Huang, "Crowdsourcing via pairwise co-occurrences: Identifiability and algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[53] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan, "Spectral methods meet EM: A provably optimal algorithm for crowdsourcing," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 3537–3580, 2016.

[54] C. Miao, Q. Li, L. Su, M. Huai, W. Jiang, and J. Gao, "Attack under disguise: An intelligent data poisoning attack mechanism in crowdsourcing," in *Proc. World Wide Web Conf.*, 2018, pp. 13–22.

[55] V. B. Sinha, S. Rao, and V. N. Balasubramanian, "Fast dawid-skene: A fast vote aggregation scheme for sentiment classification," 2018, *arXiv:1803.02781*.

[56] V. C. Raykar *et al.*, "Learning from crowrds," *J. Mach. Learn. Res.*, vol. 11, no. 4, pp. 1–26, 2010.