Attention-Based Network for Weak Labels in Neonatal Seizure Detection

Dmitry Yu. Isaev DMITRY.ISAEV@DUKE.EDU
Department of Biomedical Engineering
Duke University
Durham, NC, USA

Dmitry Tchapyjnikov DMITRY.TCHAPYJNIKOV@DUKE.EDU
Department of Pediatrics, Department of Neurology
Duke University
Durham, NC, USA

C. Michael Cotten MICHAEL.COTTEN@DUKE.EDU
Department of Pediatrics
Duke University
Durham, NC, USA

David Tanaka DAVID.TANAKA@DUKE.EDU
Department of Pediatrics
Duke University
Durham, NC, USA

Natalia Martinez NATALIA.MARTINEZ@DUKE.EDU
Department of Electrical and Computer Engineering
Duke University
Durham, NC, USA

Martin Bertran MARTIN.BERTRAN@DUKE.EDU
Department of Electrical and Computer Engineering
Duke University
Durham, NC, USA

Guillermo Sapiro GUILLERMO.SAPIRO@DUKE.EDU
Department of Electrical and Computer Engineering
Departments of Biomedical Engineering, Computer Science, and Department of Mathematics
Duke University
Durham, NC, USA

David Carlson DAVID.CARLSON@DUKE.EDU
Department of Civil and Environmental Engineering
Department of Biostatistics and Bioinformatics
Duke University
Durham, NC, USA

Abstract

Seizures are a common emergency in the neonatal intesive care unit (NICU) among newborns receiving therapeutic hypothermia for hypoxic ischemic encephalopathy. The high

incidence of seizures in this patient population necessitates continuous electroencephalographic (EEG) monitoring to detect and treat them. Due to EEG recordings being reviewed intermittently throughout the day, inevitable delays to seizure identification and treatment arise. In recent years, work on neonatal seizure detection using deep learning algorithms has started gaining momentum. These algorithms face numerous challenges: first, the training data for such algorithms comes from individual patients, each with varying levels of label imbalance since the seizure burden in NICU patients differs by several orders of magnitude. Second, seizures in neonates are usually localized in a subset of EEG channels, and performing annotations per channel is very time-consuming. Hence models which make use of labels only per time periods, and not per channels, are preferable. In this work we assess how different deep learning models and data balancing methods influence learning in neonatal seizure detection in EEGs. We propose a model which provides a level of importance to each of the EEG channels - a proxy to whether a channel exhibits seizure activity or not, and we provide a quantitative assessment of how well this mechanism works. The model is portable to EEG devices with differing layouts without retraining, facilitating its potential deployment across different medical centers. We also provide a first assessment of how a deep learning model for neonatal seizure detection agrees with human rater decisions - an important milestone for deployment to clinical practice. We show that high AUC values in a deep learning model do not necessarily correspond to agreement with a human expert, and there is still a need to further refine such algorithms for optimal seizure discrimination.

1. Introduction

Seizures during the neonatal period are a common emergency in Neonatal Intensive Care Units (NICU). After a perinatal hypoxic-ischemic event, 30-60% of infants develop seizures (Kharoshankaya et al., 2016; Nash et al., 2011). Fifteen percent of infants with seizures die and an additional 50% experience significant disability, including cerebral palsy, intellectual disability, and future epilepsy (Ronen et al., 2007; Lai et al., 2013). In newborns, clinical seizures symptoms can be extremely subtle or not exist at all, thus requiring electroencephalographic (EEG) monitoring for seizure identification (Wietstock et al., 2016). At leading and major medical centers, seizure detection currently relies on a clinical neurophysiologist reviewing continuous EEG recordings at standard intervals (at our center currently every 4-6 hours) to identify seizures in the preceding time period. Because seizure screening occurs once every several hours, treatment delays are inevitable. This issue motivates the development of a continuous monitoring solution to decrease time to seizure identification and treatment as timely intervention is critical for positive outcomes. Given recent advances in automated seizure detection (Temko et al., 2011a; Ansari et al., 2019; O'Shea et al., 2020), the goal of creating machine learning software tools to automatically detect seizures and help clinicians to make decisions now seems more achievable than ever (Mathieson et al., 2016a,b; Temko et al., 2015).

To study our proposed learning framework, we focus on two sources of data. Recently, the Helsinki University Hospital has released a NICU dataset of neonatal seizures with three distinct raters ("Helsinki dataset" from now on) (Stevenson et al., 2019). Additionally, we have built a dataset from our own historical cache of patients, yielding 31 additional individuals, to build and evaluate the proposed methods ("Duke dataset"). Having these two datasets allows us to evaluate methods in the context of two centers' data and addi-

tionally evaluate how well the learned algorithms generalize to a new center, an important consideration in deployment.

This application comes with several important considerations. A first issue is that the data suffers from severe label imbalance, i.e., low proportion of seizure events, a noted issue in training machine learning models (Johnson and Khoshgoftaar, 2019). Additionally, the training data comes from several patients, each with highly varying levels of seizure rates (in the Duke dataset it varies from 0.09% to 24%). We propose to address this challenge as a group-label imbalance problem (controlling for class imbalance individually per patient, referred in our case as 'Patient-Class imbalance'), and explore best data subsampling practices for training in this scenario.

Second, training datasets typically only provide "weak labels," meaning that only periods of time containing seizures are labeled without specifying the EEG channel exhibiting the seizure. However, as mentioned above, seizures in neonates are not typically whole-brain events and are often localized to individual brain regions, meaning that the seizure only appears in some of the measured EEG channels. Therefore, we would like a method to help localize a seizure to specific channels. Recent work has begun to utilize weak labels in CNNs (O'Shea et al., 2020; Ansari et al., 2019), but has yet to focus on effective localization, applicable to the task at hand. This goal is twofold: (*i*) we would expect that building this information into the method would improve performance; and (*ii*) downstream implementations would almost certainly require manual verification, and highlighting EEG channels which presumably exhibit seizures could accelerate this process. We address this challenge by building an attention-based Multi-Instance Learning (MIL) framework (Ilse et al., 2018). The MIL framework (Kraus et al., 2016; Wang et al., 2018b) is used to handle weak labeling, whereas the attention mechanism is used to highlight channels of interest for classification. We go further, evaluating the highlighting done by attention mechanism through comparing it with human per-channel seizure annotation to find out whether network "sees" the same thing as human does.

A third critical consideration is that previous studies have shown good performance metrics on in-house datasets (Temko and Lightbody, 2016; Tapani et al., 2019; Ansari et al., 2019), and only one study so far evaluated the results on an external dataset (O'Shea et al., 2020). However, in neonatal seizures, the inter-rater agreement is often relatively low (Stevenson et al., 2015, 2019). We explore both the pure AUC from our predictive metrics, but also evaluate how well the chosen algorithm would replicate a doctors' analysis using a variety of approaches and thresholds.

In the rest of this manuscript, we evaluate how well our proposed methods address these challenges in the context of the two mentioned real-world datasets. Overall, our performance when trained on Duke dataset is excellent (AUC $\simeq$ .970), and maintains relatively high performance when evaluated on untouched data from a different center (AUC $\simeq$ .925), despite a change in electrode layout and device between the two centers. These results show that there are still challenges to tackle on a universal solution, but point towards a potential continuous monitoring framework. In addition, we evaluated our methods versus multiple doctors, yielding algorithm-doctor agreement scores (Cohen's κ (Cohen, 1960)) only slightly lower than physician-physician inter-rater agreement.

Technical Significance. The proposed attention-MIL framework can help localize which channels are likely to indicate seizures, which we validate empirically. While previous studies had considered the weak labeling problem (O'Shea et al., 2020; Ansari et al., 2019), recovering seizure channels from weak labels can help accelerate downstream deployment and human validation. Additionally, we explore how the group-class imbalance affects the proposed algorithms during training. We provide additional metrics to explore how the algorithm matches human decision making at different parameter settings; while there are reports on matching algorithm and human performance (Temko et al., 2011b), prior studies with neural networks have focused primarily on AUC (O'Shea et al., 2017, 2020; Ansari et al., 2019). Finally, it is rare in the deep learning literature in this field to have a true second dataset collected in a different context; we posit that these results help reveal the true deployment utility of the learned algorithms.

Clinical Relevance. Intermittent review of continuous EEG recordings by a neurophysiologist inevitably leads to delays in seizure identification and treatment. A prior survey of neurophysiologists and neurointensivists showed that the frequency of reviewing EEGs varies widely: only 5% of surveyed physicians reviewed EEGs continuously, while 75% reviewed it two or more times per day (Gavvala et al., 2014). A similar survey demonstrated that 50% of responders reviewed EEGs two times a day or less (Abend et al., 2010). Higher seizure burden is independently associated with worse neurodevelopmental outcomes, both for hypoxic ischemic encephalopathy patients (Kharoshankaya et al., 2016; Glass et al., 2009), as well as in other pediatric critical care situations (Payne et al., 2014). In the NICU, the paucity of clinical signs suggestive of seizure in neonates results in most (if not all) seizures being identified on EEG after which the clinical team caring for the infant is informed. Decreasing time to seizure identification and treatment is therefore essential for reducing seizure burden and potentially improving clinical outcomes. The benefits of a fully continuous monitoring system are clear, as an automated detection system could flag potential seizures and lead to more timely seizure treatment. Such a system would need to capture most seizures and be highly specific since systems with high false positives are frequently ignored. A key component of this continuous monitoring system is the development of a reliable automatic detection procedure; in this manuscript, we present a machine learning approach to the automatic detection problem based upon datasets from two centers.

Generalizable Insights about Machine Learning in the Context of Healthcare

Transferring models that apply to EEG data is difficult due to differences in equipment and clinical protocols used to perform data collection. While electrodes are usually placed according to international standards (e.g., the 10-20 placement system (American Encephalographic Society, 1994)), deployed systems differ between centers (e.g., different numbers of electrodes), yielding different dimensionalities of data. This is a challenge when transferring models between centers, hindering real-world applicability. Therefore, we focus on learning machine learning models that are robust to such differences, and we evaluate its multi-center capabilities by evaluating on data from multiple centers and electrode layouts. By comparing the automatic evaluation with doctor evaluations, we revealed the necessity to tune thresholds to specific data sources rather than solely considering AUC. Further-

more, for a high-stakes decision, we assert that it is critical to underpin the decision in an interpretable manner to facilitate human review. To address this challenge, our proposed model is agnostic to the amount of channels and highlights those channels that are likely to exhibit seizure activity for a given timeframe. We then evaluate how our highlighting system matches with human interpretation.

2. Cohort

2.1. Data Collection and Annotation

2.1.1. Duke dataset

Patients aged <30 days who received continuous EEG (cEEG) monitoring between 2012 and 2019 were first identified through the EEG database system utilized by Duke University Medical Center (Natus NeuroWorks®). Medical records were then manually reviewed and infants who were concurrently undergoing therapeutic hypothermia while being monitored on EEG were selected. A total of 154 patients were identified, 45 of whom developed seizures during cEEG monitoring, as assessed by an experienced epileptologist. After exclusion of corrupt files, cEEG data of 31 infants with seizures were available. This study of human subjects was approved by the Duke Health Institutional Review Board (Pro00100420).

Among 31 infants retained in the dataset, 42% (n=13) were female with a median gestational age of 39 weeks (Inter-Quartile Range (IQR) 38-40) at time of birth. Median time from birth to EEG placement was 9 hours (IQR 5-11). EEG recordings started at onset or soon after initiation of therapeutic hypothermia and recordings continued until 24 hours after rewarming. There were more seizures typically in the beginning and they decreased in frequency later in the recordings, yet entire recordings, regardless of therapeutic hypothermia phase, were used for algorithm development and training.

An experienced epileptologist from Duke University Medical Center annotated the dataset. Annotation for each seizure was provided in a separate table marking the beginning and end time of the seizures with 1 second resolution. In total, the dataset contained 2320 hours of recording with 50.81 hours of annotated seizures.

Summary of the Duke dataset is provided in Table 1 and Figure 1. Details can be found in Appendix A.1.

Since the system is intended to monitor all at-risk individuals, it is critical to maintain low false alarm rate. It is especially important on patients without seizures so that the system would not get ignored by practitioners. For an additional evaluation of our algorithm on such patients, we utilized 10 out of 154 newborns who underwent therapeutic hypothermia but did not develop seizures. This subset of patients had a median gestational age of 39 weeks (IQR 37-40) at time of birth and their median time from birth to EEG placement was 9 hours (IQR 5-12), and duration of each recording was 24 hours.

2.1.2. Subsample of the Helsinki Dataset for Cross-Dataset Validation

To get a better understanding of the generalizability of an algorithm, it is important to evaluate it in a variety of environments. For that purpose, we used data and annotations from the Helsinki dataset (Stevenson et al., 2019). We selected patients that had seizures by consensus of 3 raters (total of 39 patients, 53% (n=21) female, 41% (n=16) male, gender

Table 1: Summary of seizure amount, duration, and total recording duration in the Duke dataset. The dataset consists of 31 patients with continuous EEG recordings (minimum duration of 24 hours), which is typical of the multiple-day seizure monitoring protocols utilized in many NICU settings.

	Amount of seizures	Total hours	Total seizure hours	Seizure rate
Total	1778	2320.00	50.81	-
Mean	57.35	74.84	1.64	2.81%
Std	71.95	35.31	2.41	4.91%

not provided for 2 patients). Median gestational age for this subsample was 39 weeks (IQR 38-40). Summary for the subsample is provided in Table 2 and Figure 1.

Table 2: Summary of seizure amount, duration, and total recording duration in a subset of 39 patients from the Helsinki dataset who had seizures by consensus.

	Amount of seizures	Total hours	Total seizure hours	Seizure rate
Total	343	60.12	10.91	-
Mean	8.80	1.54	0.28	18.60%
Std	11.2	0.71	0.38	21.09%

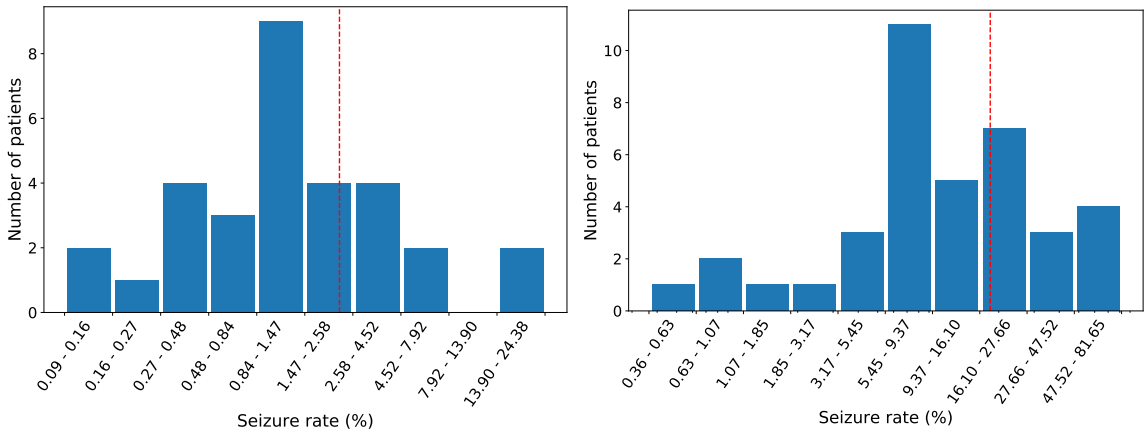

Figure 1: Histogram of seizure rate per patient in the Duke dataset (left) and the Helsinki dataset (right) on a log-scale. Red dotted line is prevalence of seizures over entire dataset (2.2% for the Duke dataset, 18.1% for the Helsinki dataset)

2.2. Data Extraction

To make results comparable with existing literature, all the data in this paper was extracted and preprocessed with the routine outlined in (Temko et al., 2011a) using the publicly available code from (Tapani et al., 2019).

EEG electrode setup for Duke dataset was based on the international 10-20 placement system modified for neonates as recommended by the American Clinical Neurophysiology Society Guidelines (Shellhaas et al., 2011; Kuratani et al., 2016). EEG recordings were initially collected with a sampling frequency of 256Hz, using 9 electrodes. As is standard practice, bipolar derivations (differences between time-series from neighboring electrodes) were computed, resulting in the following 12 data channels (a.k.a. the 'double banana' montage): Fp1-C3, C3-O1, Fp2-C4, C4-O2, Fp1-T3, T3-O1, Fp2-T4, T4-O2, T3-C3, C3-Cz, Cz-C4, C4-T4 . Notch filtering (at 60Hz for the Duke dataset, and at 50Hz for the Helsinki dataset), high-pass filtering at 0.5 Hz, low-pass filtering at 16 Hz and down-sampling to 32Hz was performed. Then data for each patient was split into subsequent 8-second chunks with 4 seconds overlap (referred from here on as epochs). Any period with data losses in the recording (a small minority of data) was removed.

2.3. Feature Choices

For the i-th patient after preprocessing we had $N_{ep,i}$ epochs with dimension $(N_c, 256)$ where N_c is number of bipolar channels (12 for the Duke dataset, 18 for the Helsinki dataset), and 256 is the amount of timepoints per 8 seconds on 32Hz downsampled data. For the deep learning approaches, here developed and investingated, this data format was directly used. The Support Vector Machine approach here tested relies on human-engineered features, so each epoch of data was converted to 55 features per channel. These features follow (Temko et al., 2011a), and are representative of frequency domain, time domain, and information theory based characteristics of the signals.

3. Methods

In our work we compare two novel deep learning approaches and one classical approach (SVM), and investigate how the choice of data balancing techniques influences overall performance over the algorithm. We use AUC on leave one patient out cross-validation (LOO CV) as the main performance metric to evaluate how our performance generalizes to new individuals. Furthermore, we explore how well best performing algorithms generalize using cross-center validation. Specifically, we use the publicly available Helsinki dataset (Stevenson et al., 2019) and take the best performing model on full Duke dataset and evaluate its performance on the Helsinki dataset. We additionally assessed the performance of a publicly available SVM model pre-trained on the Helsinki dataset (Temko et al., 2011a; Tapani et al., 2019) on Duke dataset. Finally, we analyze how well one of our models can identify seizure activity per channel using only per-epoch labels for training, and how performance is associated with inter-rater agreement.

3.1. Machine Learning Models

3.1.1. Deep Learning Models

Our methods are based on the Convolutional Neural Network due to its widespread success in signal processing task. We primarily focused on two architectures. These architectures use a per-electrode (or per-channel) feature extractor with weights shared across all electrodes. Our feature extractor is based on Inception blocks (Szegedy et al., 2015) for their multi-scale filtering. We hypothesize that this structure might help in classification due to the evolution of seizures in frequency. In preliminary experiments we saw an improvement in performance of this architecture over the standard CNN filter approach. After adapting the Inception block to one-dimensional data, our feature extractor had 8,514 trainable parameters. Additional details on the feature extractor can be found in Appendix A.2. Below, we discuss the structure of our two proposed networks, and their visualization can be seen in Figure 2.

In our first deep learning network (DL1), the outputs of the per-channel feature extractor are concatenated and then passed through a dense layer. Since the number and order of channels is fixed, using a dense layer overall helps the classification since the channels are not independent (at least because channels are bipolar derivations of raw electrode signals); however, this should be carefully addressed since seizure activity appears in different channels for different patients.

In contrast, in our second deep learning network (DL2), the output of the per-channel feature extractor is passed through an attention-MIL layer, as outlined in Ilse et al. (2018). We built upon this framework with the intention that channels exhibiting seizures should be given more weight, which could both improve modeling and facilitate communication of the results. After the attention layer, a weighted average of the features is passed through a dense layer. Thus, this model is agnostic to channel interaction, facilitating portability to any channel layout. This is a desired feature for a generalizable seizure detection algorithms, since EEG setup can vary in different NICUs (Ansari et al., 2019), and such configuration allows the model to be used in different NICUs without retraining, and also to jointly learn from multi-center weakly labeled datasets. This is also critical in our cross-center validation, because the two centers use different electrode layouts.

To be more specific, the attention-MIL layer in DL2 takes as an input a bag of $\{h_k\}$ features ($h_k \in \mathbb{R}^{1\times 48}$ in our case), $k = 1...N_c$ (with N_c the number of channels), and outputs

$$\mathbf{z} = \sum_{k=1}^{N_c} a_k \mathbf{h_k},$$

where

$$a_k = \frac{\exp\left(\mathbf{w}^\top \tanh \mathbf{V}\mathbf{h_k}^\top\right)}{\sum_{j=1}^{N_c} \exp\left(\mathbf{w}^\top \tanh \mathbf{V}\mathbf{h_j}^\top\right)},$$

and $\mathbf{w} \in \mathbb{R}^{L\times 1}$ and $\mathbf{V} \in \mathbb{R}^{L\times 48}$. $\mathbf{w}$ and $\mathbf{V}$ are the learned weights, and L is the inner dimension of the attention-MIL layer. For this work we selected $L = 32$.

Critically, attention-MIL weights can be used as a proxy for whether channels exhibit seizure activity, and help clinicians understand "where to look at," i.e., which channels contributed most to the detection of seizure.

In total, DL1 and DL2 had 27,011 and 11,683 trainable parameters respectively. In each experiment we trained a network for 25,000 steps with a batch size of 256.

We implemented our DL models in the Keras framework (Chollet and others, 2015) with TensorFlow GPU backend, and run them on a desktop with 6-core i7 Processor with 64Gb of RAM and GeForce 1080 Ti GPU. Both code and pre-trained DL2 model on Duke dataset will be made available at https://github.com/dyisaev/seizure-detection-neonates.

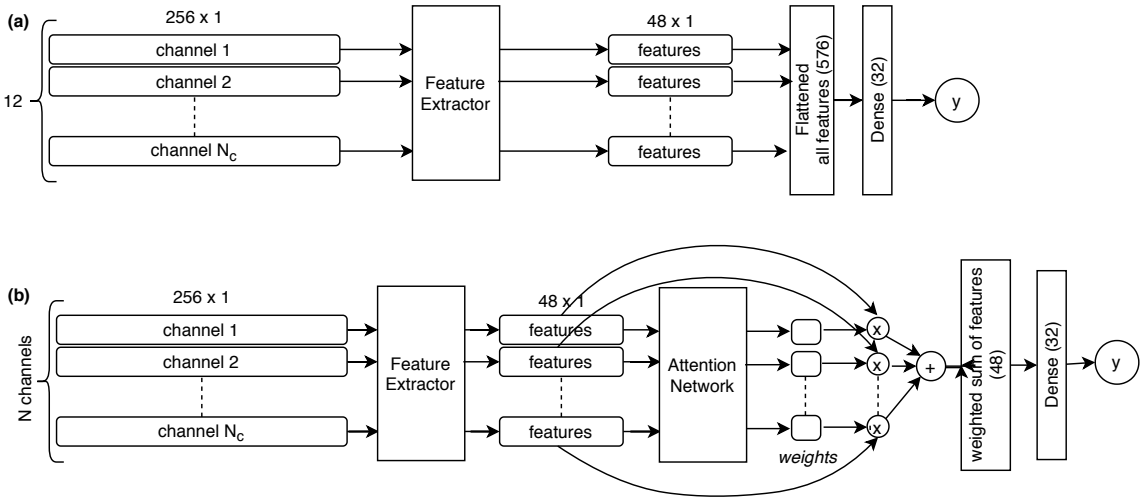

Figure 2: Graphical schema of the two Deep Learning architectures studied in this paper. DL1 model (a) , DL2 model (b). Both models share same per-channel feature extractor module, described in Appendix A.2. Feature extractor weights are the same for all channels.

3.1.2. Classical ML Models - Support Vector Machines

To compare the proposed and studied deep learning approaches with classical ML approaches, we selected a model which has shown good results in previous publications (Temko et al., 2011a; Tapani et al., 2019; O'Shea et al., 2020). We replicated the exact procedure of feature extraction using publicly available code (Tapani et al., 2019), training the model and predicting seizure. The model used radial basis function SVM based on 55 features (Temko et al., 2011a). The model trains on 55x1 features, representing 8-second recording segment per channel. It takes advantage of strong labels, combining only data from channels marked as 'seizure' in seizure samples and data from random channels from non-seizure segments during training. The model predicts seizure per-epoch if at least one channel exhibits seizure; predictions are done per channel, smoothed with moving average of 3 consecutive segments, and finally overall time segment prediction is done by max-pooling per-channel predictions.

3.2. Data Balancing Approaches

Previous literature suggests the detrimental effect of class imbalance on CNN performance (Buda et al., 2018), and no work so far has fully explored the influence of class balancing on the classification performance in neonatal seizures. Moreover, imbalance in seizure burden varies across patients (see Figure 1). Thus, we tested each method with 3 types of balancing approaches: No balancing (simply subsampling all available training epochs); Class balancing (keeping the proportion of classes (labels of seizure/non-seizure) in each minibatch equal); Patient-Class balancing (keeping the proportion of (Patients x Classes) partitions equal in each training minibatch). While Class balancing addresses the problem of algorithms seeing much more negative (non-seizure) than positive (seizure) examples, there is still a problem of algorithms seeing much more positive examples from patients with high seizure burden in this approach. Patient-Class balancing intends to address that, and is expected to provide better generalization.

3.3. Post-processing

We can use post-processing procedures to reduce short false positive periods and link together longer seizures, providing a slight boost to AUC. We explicitly specify in the reported results if post-processing was used, which includes probability reweighting (to adjust for true class prevalence in the dataset, see Appendix A.3) and transforming the outputs to improve robustness (see Appendix A.4).

4. Results

4.1. Evaluation Approach/Study Design

We selected area under the receiver operating curve (AUC) on leave-one-patient-out (LOO) cross-validation as a main measurement of model performance. We also explored the influence of post-processing, so we estimated performance in 2 ways. First, we assessed AUC when prediction is evaluated on each epoch of LOO patient's data. Second, we applied the post-processing procedures for the best performing model with different thresholds for computing AUC and evaluated AUC on each second of the LOO patient's data.

For cross-center validation, we selected best performing model on Duke dataset and the publicly available SVM model trained on the Helsinki dataset (referred as SVM_T in Tapani et al. (2019)). We computed AUC per patient on the Helsinki dataset for 39 patients that had seizures by consensus. However, the 3 raters of the Helsinki dataset disagreed on precise beginnings and ends of seizure periods regions, thus we used only the regions where all 3 raters agree for computing[1]. This is directly comparable with AUCs reported in previous work (O'Shea et al., 2020).

We also assessed Cohen's κ (Cohen, 1960) between the proposed algorithm output and a human rater for the Duke dataset (or a consensus of 3 raters for the Helsinki dataset), as well as the sensitivity and specificity dependance on the selected decision threshold given our data and algorithm.

1. We expect that this would increase the AUC over using a single rater's labels alone.

To evaluate how well our best algorithm (trained only on patients with seizures) performs on patients without seizures, we computed specificity and number of seizures detected on a previously unseen set of 10 patients' recordings from NICU of Duke University Medical Center deemed as non-seizure by the same epileptologist who annotated the Duke dataset. All recordings were 24 hours long. This requires selection of the decision threshold, which we set as the probability of positive class over the entire training dataset (see Appendix A.3 for derivations).

To assess how well the attention-MIL mechanism of DL2 model captures seizure channels, we performed AUC analysis of attention-MIL scores on Duke dataset for seizure epochs. We computed the AUC value between the scores and human annotation in two settings: (a) per channel and epoch ('Attention AUC') - each individual channel was assigned a positive or negative label based on the epileptologist per-electrode labels, and AUC was calculated using the channel-specific prediction (i.e., prediction if attention only used that channel); and (b) per epoch ('Attention AUC per epoch') - if at least one channel exceeding the decision threshold in an epoch is deemed a seizure by the human rater, then we consider the epoch as true positive, and we compute true and false positive rates. To the best of our knowledge, this is the first quantitative assessment of how well the deep learning algorithm trained using weak (per-epoch) labels is able to provide per-channel annotations.

4.2. Results on Machine Learning Approaches on Different Balancing Techniques

Results of different balancing techniques and their influence on the deep learning approaches are summarized in Table 3. To measure significance of difference between each pair of approaches we performed Wilcoxon paired signed-rank test (Wilcoxon, 1945), see Appendix A.5. Class Balancing approach on DL2 model outperformed all other approaches, resulting in AUC of 0.950.

Note that the SVM approach does not operate on weak labels, and so is limited by the availability of per-channel labels.

Table 3: Results of different balancing approaches and their influence on the performance on Duke dataset (average AUC on leave one patient out cross-validation). Standard deviation (SD) is shown in parentheses. Results do not include post-processing routine.

Model	No Balancing	Class Balancing	Patient-Class Balancing
DL1	**0.933 (0.055)**	0.923 (0.070)	0.911 (0.086)
DL2	0.923 (0.057)	**0.950 (0.041)**	**0.943 (0.051)**
SVM	0.822 (0.063)	0.772 (0.061)	0.765 (0.058)

To further explore the influence of post-processing on the results, we performed post-processing on the best performing model (Class-balanced DL2). With post-processing the model achieved the average AUC of 0.970 (SD: 0.033).

4.3. Results on Cross-Dataset validation

The results for cross-dataset AUC presented in Table 4 were achieved including the post-processing routine, which was the same for both datasets. The drop in performance between datasets was less for DL2 than for SVM_T, showing significant promise for DL2 to generalize to new centers.

We note that the SVM_T shows significantly higher performance than the SVM trained on our own data. Again, the SVM does not operate on weak labels, and so is limited by the availability of per-channel labels, which was higher in the Helsinki dataset. Additionally, the Helsinki dataset labeled positive and negative channels on the montage (bipolar derivations) whereas Duke dataset labeled individual electrodes that was expanded to the montage. This difference in labeling could explain this performance difference because the algorithms actually operate on the montage.

Table 4: Results of cross-dataset validation as measured by average AUC on per-patient evaluation of models. Evaluation on the same dataset is done via Leave One Patient Out cross-validation. Uncertainties given are the SD over patients.

Model	Trained On	Duke dataset	Helsinki dataset
DL2 (Class balance)	Duke dataset	0.970 (0.033)	0.925 (0.099)
Pre-trained SVM (SVM_T)	Helsinki dataset	0.826 (0.117)	0.923 (IQR 0.869–0.990) [2]

4.4. Association of per-patient AUC scores with inter-rater agreement

We wanted to evaluate how well our proposed method works relative to a typical rater. To do this, we calculated the average inter-rater agreement for each patient using Cohen's κ on the Helsinki dataset. We then compared this value to the AUC calculated on each patient, shown in Figure 3. It is clear from the picture that as Cohen's κ grows, both AUC grows and variability in AUC reduces. In other words, when human raters agree with each other, we largely agree with them. Spearman's ρ correlation coefficient between average κ and AUC is 0.56 ($p<0.001$), showing a strong statistical relationship.

4.5. Agreement between the algorithm and a human rater

Using a threshold of 0.022 (corresponding to a .5 threshold corrected for the prevalence in the Duke dataset), we calculated the agreement with a human rater on Duke dataset using Cohen's κ, and our algorithm gave a median value of 0.517 with an IQR of 0.313-0.671 on the per-patient agreement with human rater on Duke dataset. Median value was 0.59 with an IQR of 0.119-0.769 of the agreement with a consensus of 3 raters on the Helsinki dataset. Because these values are dependent on the chosen threshold, we wanted to evaluate how much the choice of threshold impacts the achieved performance. We visualize the median and IQR of Cohen's kappa, sensitivity and specificity compared to a varying decision threshold, in Figure 4. It is clear from the graph that optimal thresholds are

2. Reported in (Tapani et al., 2019), SD was not provided

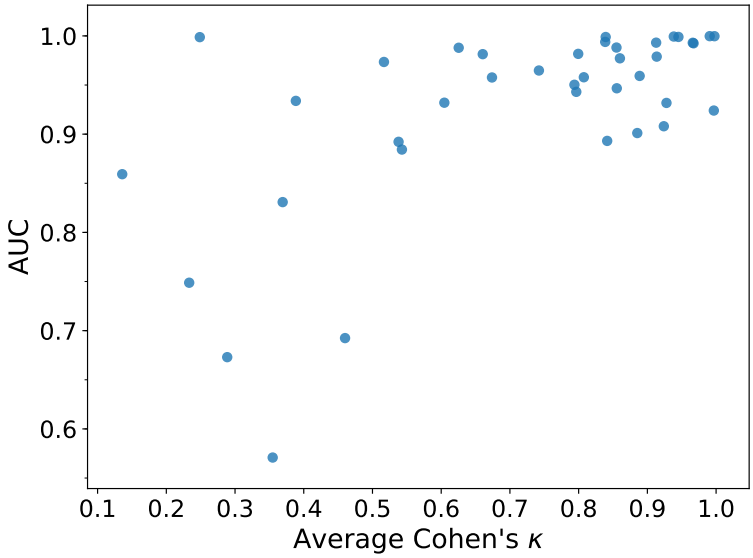

Figure 3: Scatterplot of average Cohen's κ of inter-rater agreement on the Helsinki dataset vs cross-dataset AUC on patients with consensus seizures/non-seizures on the Helsinki dataset

different for the two datasets. If we select an optimal threshold based on Cohen's κ on the Duke dataset (green dotted vertical line, Figure 4 (top row, left image)) we get a serious drop in sensitivity and specificity on the Helsinki dataset.

4.6. Validation on non-seizure patients

We next evaluated the false positive rate on patients where an epileptologist did not mark any seizures. Specifically, we used 10 patients, each of which had 24-hours of non-seizure recordings, from Duke University Medical Center. At the chosen threshold level, the algorithm flagged 589 seizures, with median of 42 seizures per recording (IQR 25.5-84.5). Median specificity per patient was 0.98 (IQR 0.94-0.99). Median duration of the detections was 30 seconds (IQR 28-32.5), which, given that 16 seconds is a collaring length in post-processing, provides a 'raw' detection length of 3-4 consecutive epochs. These results show the importance of decision threshold selection, post-processing, and adapting the algorithm to background noise (Temko et al., 2013). The level of false positives is, of course, related to the seizure threshold, as can be seen more broadly in Figure 4.

4.7. Attention Network Visualization

Finally, we evaluated the performance of the attention mechanism of the DL2 network, which is summarized in Table 5. It is worth noting that for the Duke dataset per-channel annotations were provided per electrode, while for the Helsinki dataset per-channel annotations were provided per bipolar derivations. Thus, if an electrode was marked as 'seizure' in an epoch, then we considered all bipolar derivations including that electrode as seizures. As

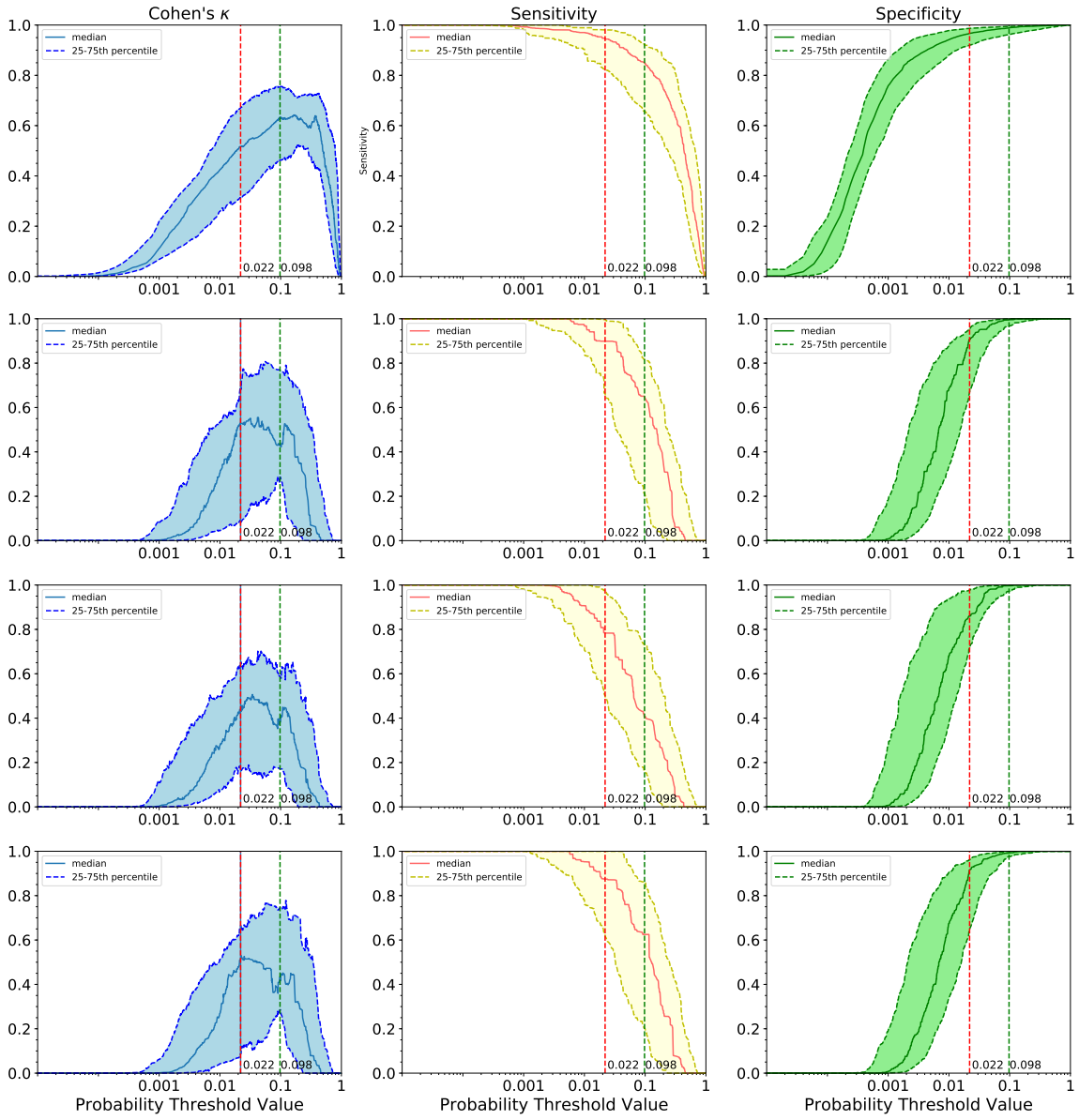

Figure 4: Ranges of Cohen's κ, sensitivity, and specificity as decision threshold changes for DL2 Class balanced model. Top row: Results on the Duke dataset LOO; Three bottom rows: Results on the Helsinki dataset, for rater 1, rater 2, and rater 3 respectively. Red dotted vertical line - threshold corresponding to a .5 threshold corrected for the true prevalence (Duke dataset). Green dotted vertical line - empirical optimal threshold based on Cohen's κ in the training sample (Duke dataset). Threshold values are provided on a log scale.

a result of this, different electrode annotations lead to different amounts of bipolar derivations considered seizures (e.g., for Fp1 two channels were marked, while for C3 four channels were marked).

Table 5: Attention network performance of DL2 (Class balance) model on the Duke dataset and the Helsinki dataset, as measured by averaged AUC (SD in parentheses). Computation of "Attention AUC" used agreement between each thresholded score and human annotations per channel per epoch; "Attention AUC per epoch" uses agreement of at least one channel score exceeding threshold with human annotation.

Dataset	Attention AUC	Attention AUC per epoch
Duke dataset	0.811 (0.096)	0.927 (0.058)
Helsinki dataset	0.701 (0.107)	0.807 (0.167)

To provide a qualitative measure on how the attention network works, Figure 5 summarizes how weights are distributed in the attention network in seizure/non-seizure samples for one of the patients for the entire recording. We also visualize the output of the network during the beginning and end of a seizure event in Figure 6. For this patient, all seizures were focused on leads O1 and O2, as annotated in the Duke dataset. While the algorithm and rater agreed on the general location of the seizure, there was disagreement on the exact start and end location.

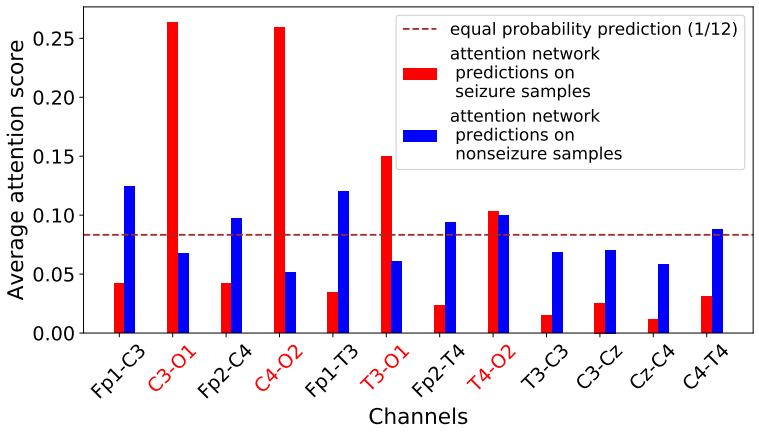

Figure 5: Average attention scores across all samples for one of the patients from Duke dataset. Ticks marked red - ground truth, provided by epileptologist (for all seizures of this patient, epileptologist marked O1 and O2 as electrodes where seizures are visible). Attention AUC for this patient was 0.88.

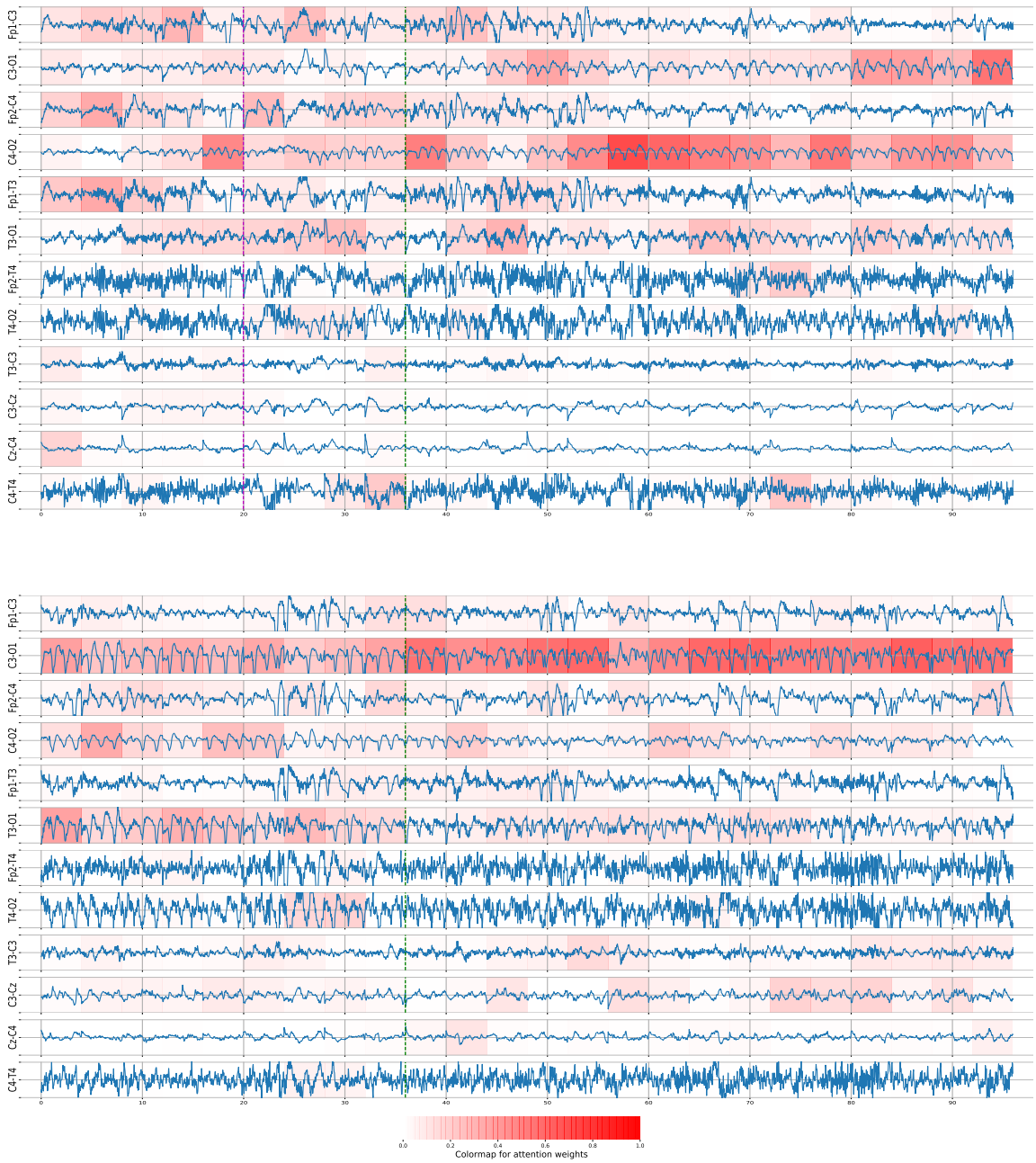

Figure 6: (Top) Beginning of a seizure in a patient from Duke dataset. The green dotted line marks the beginning from the epileptologist. The magenta dotted line marks the beginning decided by network. The colored background intensity corresponds to how much attention weight is given to each channel at each time segment. (Bottom) End of the same seizure. The green dotted line marks the end as labeled by the epileptologist. The network deems the whole segment as a seizure, and most of the weight for its decision is coming from channel C3-O1, which is also deemed the relevant channel by the epileptologist.

4.8. Ablation study

We performed an additional ablation study to determine the impact of the attention layer on overall prediction. We removed the attention layer and performed simple averaging of per-channel features after the distributed feature extractor layer, with the other hyperparameters held constant. The attention layer provided marginal improvement of AUC (0.945 with ablation of attention vs 0.950 for full network). This observation indicates that classification power of our approach comes mostly from the feature extractor selected. Regardless, the utility of the attention layer is useful for communicating the results, as demonstrated in Figure 6.

5. Discussion

5.1. Deep Learning Models and Balancing Strategies

It is well-established that deep learning models suffer under significant label imbalance. Few studies on adult epilepsy explicitly took into account data imbalance (Yuan et al., 2017; Wu et al., 2020), and most of the previous studies on data imbalance in CNN training were focused around CNN for image classification (Johnson and Khoshgoftaar, 2019). In our data, the seizure prevalence per patient is highly variable (from 0.08% to 24.3% in the Duke dataset), so we hypothesized that addressing the data imbalance might be a crucial issue in algorithm performance. However, we found only slight changes in performance due to the varying data balancing strategies. Part of this may be due to the methods evaluated; in our study we explore only data-level methods, comparing no balancing, class balancing (same amount of seizures/non-seizures per batch, also known as 'class-aware sampling' (Shen et al., 2016)), and Patient-Class balancing (same amount of seizures/non-seizures both per batch and per patient in batch). The structure of features in the DL2 model (both per-channel extracted features and weighted average of per-channel features have 48 dimensions) could facilitate a variety of additional data-level approaches (e.g., SMOTE (Chawla et al., 2002)).

We did find highly variable performance with different neural network structures. In our second Deep Learning model (DL2), we proposed an approach that is electrode-number agnostic; that is, it can work on different devices and electrode layouts without retraining the network. In this network, the class balancing approach worked the best. We hypothesize that this is because the Patient-Class balancing over-weighted less common seizures from low seizure-prevalence patients. However, the variability in the models shows that the results on balancing are inconclusive.

Because the post-processing approaches are dependent on the probability estimates, we want to have proper probability estimates. However, these balancing schemes get rid of the class prior and must be corrected to give proper probabilistic estimates. In our case, we have done this by post-scaling of output probabilities (Lawrence et al., 2012; Zhou and Liu, 2006; Buda et al., 2018). This post-scaling could be combined with other calibration approaches (e.g., Platt scaling) to get accurate probabilities.

5.2. Algorithm-rater and Inter-rater Agreement

Our results on agreement in the Duke dataset LOO setting and the Helsinki dataset indicate that high AUC values are not enough for the deep learning seizure detection algorithm

to be immediately transferable to clinical practice. Our algorithm reaches median 0.517 agreement with human rater on the Duke dataset and median 0.59 with consensus of 3 raters on the Helsinki dataset, as compared to 0.807 (IQR 0.540-0.913) of Cohen's κ averaged across 3 pairs of human raters on the Helsinki dataset. We can see that agreement between the algorithm and human raters is worse than agreement between the 3 human raters. It's also evident that variability on cross-dataset prediction for Cohen's κ is much higher. This may be due to different prevalence of seizures in the Helsinki dataset compared to Duke dataset (Stevenson et al., 2015; Vach, 2005) and the generalization error. Tapani et al. (2019) approached the problem of agreement between algorithm and human annotation as agreement between 3 raters (2 humans and an algorithm). They reported that Fleiss' κ (Fleiss and Cohen, 1973) dropped if one human rater is replaced by the algorithm. The need to search for a decision threshold, and to decide the costs of false detections and false negatives (misses), gives an intution that cost-sensitive learning (Ling and Sheng, 2008) may be another approach to address class imbalance, where cost can be either fixed (Wang et al., 2018a) or learned (Khan et al., 2018).

Note that there appears to be some gain possible from personalizing the threshold, meaning that we may need to build strategies to calibrate to individuals. This avenue could be explored through a meta-learning approach.

In addition to the κ metric, metrics that evaluate agreement on a per-event basis could be used to further assess the clinical feasibility of the algorithm. For example, analyzing the positive (seizure) agreements, negative (non-seizure) agreements and disagreements between algorithm and raters, as proposed in Stevenson et al. (2015), done across entire recording or per-hour could be used. These metrics will be addressed in future work.

5.3. Interpretability of the Results

In high-stakes decision-making, many people are rightfully wary of black-box decisions (Rudin, 2019). In our scenario, we view this system as a support tool where any predicted positive could be reviewed more quickly. In such a scenario, it would facilitate chart review to have the system be as descriptive as possible. While our attention-based system does not produce interpretable filters, it can easily highlight relevant channels and time periods for a clinician to review.

As the Attention AUC on at least one channel detected as seizure is 0.927 (SD 0.058) on Duke dataset, we consider that this system could help decrease evaluation time. While the system performance drops down to 0.807 (SD 0.058) when evaluated on the Helsinki dataset, this implies that the system is still robust to true domain shifts and can be increasingly fine-tuned. It is also important to mention that while helping to highlight the relevant channels, attention mechanism does not add to classification power of the model, which can be seen from ablation study.

While other approaches have considered weak labels, the system by O'Shea et al. (2020) was an ensemble of 3 networks with prediction averaged from three outputs. While undoubtedly imroving performance of the model, it significantly constrains the interpretability. Another CNN-based system (Ansari et al., 2019), while using weak labeling, did not provide interpretations of channel importance due to network architecture.

6. Conclusions

In this work we provided an assessment of how different models and balancing methods influence learning in neonatal seizure detection from EEG. We proposed a model that provides a level of importance to each of the channels - a proxy to whether a channel exhibits seizure activity or not. This model is portable to an EEG dataset with an arbitrary amount of channels without need for adjustment or retraining, and can provide decreased checking time for use in a secondary evaluation by a doctor. To our knowledge, we also provided the first assessment of agreement between human raters and deep learning algorithm for detecting neonatal seizures. The system, to date, has shown excellent AUC; however, we do not exactly mimic doctor behaviors towards labeling, and the estimate Cohen's κ values were comparatively low, showing room to further improve the algorithm. Future work will attempt to increase this value by focusing on improved learning strategies, additional data integration, and individualizing to a patient, e.g., by meta-learning.

Acknowledgements

This work was supported by a Children's Miracle Network Hospitals award to D.T., and D.C. and G.S. were supported by the National Institutes of Health under Award Number R01EB026937. The work of D.Yu.I. and G.S. work is partially supported by NIH, NSF, Simons Foundation, Department of Defense, and gifts from Amazon, Google, Microsoft, and Cisco. Support was also provided by the Duke Forge. The Duke PACE (Protected Analytics Computing Environment) system used to compute results is supported by Duke's Clinical and Translational Science Award (UL1TR002553) and by the Duke University Health System.

Authors thank Shelley Rusincovitch for valuable discussions and management of the project, J. Matias Di Martino for helpful conversations, and anonymous reviewers for their thoughtful comments and suggestions. Authors also appreciate the work of Stevenson et al. (2019) and Tapani et al. (2019) for making Helsinki dataset, code for data preprocessing and pre-trained models publicly available.

References

Nicholas S. Abend, Dennis J. Dlugos, Cecil D. Hahn, Lawrence J. Hirsch, and Susan T. Herman. Use of EEG monitoring and management of non-convulsive seizures in critically Ill patients: A survey of neurologists. *Neurocritical Care*, 12(3):382–389, 2010. ISSN 15416933. doi: 10.1007/s12028-010-9337-2.

American Encephalographic Society. Guideline thirteen: Guidelines for standard electrode position nomenclature. *Journal of Clinical Neurophysiology*, 11(1):111–113, 1994.

Amir H. Ansari, Perumpillichira J. Cherian, Alexander Caicedo, Gunnar Naulaers, Maarten De Vos, and Sabine Van Huffel. Neonatal seizure detection using deep convolutional neural networks. *International Journal of Neural Systems*, 29(4):1–20, 2019. ISSN 17936462. doi: 10.1142/S0129065718500119.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. ISSN 18792782. doi: 10.1016/j.neunet.2018.07.011. URL https://doi.org/10.1016/j.neunet.2018.07.011.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(Sept.

28):321–357, 2002. ISSN 10769757. doi: 10.1613/jair.953. URL https://arxiv.org/pdf/1106.1813.pdf%0Ahttp://www.snopes.com/horrors/insects/telamonia.asp.

François Chollet and others. Keras, 2015. URL https://keras.io.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. ISSN 15523888. doi: 10.1177/001316446002000104. URL https://doi.org/10.1177/001316446002000104.

Charles Elkan. The foundations of cost-sensitive learning. *IJCAI International Joint Conference on Artificial Intelligence*, pages 973–978, 2001. ISSN 10450823.

Joseph Fleiss and Jacob Cohen. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613–619, 1973.

Jay Gavvala, Nicholas Abend, Suzette LaRoche, Cecil Hahn, Susan T. Herman, Jan Claassen, Mícheál Macken, Stephan Schuele, and Elizabeth Gerard. Continuous EEG monitoring: a survey of neurophysiologists and neurointensivists. *Epilepsia*, 55(11):1864–1871, 2014. ISSN 15281167. doi: 10.1111/epi.12809.

Hannah C. Glass, David Glidden, Rita J. Jeremy, A. James Barkovich, Donna M. Ferriero, and Steven P. Miller. Clinical neonatal seizures are independently associated with outcome in infants at risk for hypoxic-ischemic brain injury. *Journal of Pediatrics*, 155(3):318–323, 2009. ISSN 00223476. doi: 10.1016/j.jpeds.2009.03.040. URL http://dx.doi.org/10.1016/j.jpeds.2009.03.040.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. *35th International Conference on Machine Learning, ICML 2018*, 5:3376–3391, 2018. URL http://arxiv.org/abs/1802.04712.

Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0192-5. URL https://doi.org/10.1186/s40537-019-0192-5.

Salman H. Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A. Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3573–3587, 2018. ISSN 21622388. doi: 10.1109/TNNLS.2017.2732482.

Liudmila Kharoshankaya, Nathan J Stevenson, Vicki Livingstone, Deirdre M Murray, Brendan P Murphy, Caroline E Ahearne, and Geraldine B Boylan. Seizure burden and neurodevelopmental outcome in neonates with hypoxic–ischemic encephalopathy. *Developmental Medicine and Child Neurology*, 58(12):1242–1248, 2016. ISSN 14698749. doi: 10.1111/dmcn.13215.

Oren Z. Kraus, Jimmy Lei Ba, and Brendan J. Frey. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59, 2016. ISSN 14602059. doi: 10.1093/bioinformatics/btw252.

John Kuratani, Phillip L Pearl, Lucy Sullivan, Rosario Maria S Riel-Romero, Janna Cheek, Mark Stecker, Daniel San-Juan, Olga Selioutski, Saurabh R Sinha, Frank W Drislane, and Tammy N Tsuchida. American Clinical Neurophysiology Society Guideline 5: Minimum Technical Standards for Pediatric Electroencephalography. *Journal of clinical neurophysiology : official publication of the American Electroencephalographic Society*, 33(4):320–323, 8 2016. ISSN 1537-1603 (Electronic). doi: 10.1097/WNP.0000000000000321.

Yin-hsuan Lai, Che-sheng Ho, Nan-chang Chiu, and Chih-fan Tseng. Prognostic factors of developmental outcome in neonatal seizures in term infants. *Pediatrics and Neonatology*, 54(3):166–172, 2013. ISSN 1875-9572. doi: 10.1016/j.pedneo.2013.01.001. URL http://dx.doi.org/10.1016/j.pedneo.2013.01.001.

Steve Lawrence, Ian Burns, Andrew Back, Ah Chung Tsoi, and C Lee Giles. Neural Network Classification and Prior Class Probabilities. In Grégoire Montavon, Geneviève B Orr, and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade: Second Edition*, pages 295–309. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-35289-8. doi: 10.1007/978-3-642-35289-8{_}19. URL https://doi.org/10.1007/978-3-642-35289-8_19.

Charles X. Ling and Victor S. Sheng. Cost-sensitive learning and the class imbalance problem. *Encyclopedia of Machine Learning*, pages 231–235, 2008. doi: 10.1.1.15.7095.

S. Mathieson, J. Rennie, V. Livingstone, A. Temko, E. Low, R. M. Pressler, and G. B. Boylan. In-depth performance analysis of an EEG based neonatal seizure detection algorithm. *Clinical Neurophysiology*, 127(5):2246–2256, 2016a. ISSN 18728952. doi: 10.1016/j.clinph.2016.01.026. URL http://dx.doi.org/10.1016/j.clinph.2016.01.026.

Sean R. Mathieson, Nathan J. Stevenson, Evonne Low, William P. Marnane, Janet M. Rennie, Andrey Temko, Gordon Lightbody, and Geraldine B. Boylan. Validation of an automated seizure detection algorithm for term neonates. *Clinical Neurophysiology*, 127(1):156–168, 2016b. ISSN 18728952. doi: 10.1016/j.clinph.2015.04.075. URL http://dx.doi.org/10.1016/j.clinph.2015.04.075.

K B Nash, S. L. Bonifacio, H. C. Glass, J E Sullivan, A. J. Barkovich, D. M. Ferriero, and M. R. Cilio. Video-EEG monitoring in newborns with hypoxic-ischemic encephalopathy treated with hypothermia. *Neurology*, 76(6):556–562, 2 2011. ISSN 1526-632X (Electronic). doi: 10.1212/WNL.0b013e31820af91a.

Alison O'Shea, Gordon Lightbody, Geraldine Boylan, and Andriy Temko. Neonatal Seizure Detection Using Convolutional Neural Networks. *arXiv*, 2017. URL https://arxiv.org/pdf/1709.05849.pdf.

Alison O'Shea, Gordon Lightbody, Geraldine Boylan, and Andriy Temko. Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture. *Neural Networks*, 123:12–25, 2020. ISSN 18792782. doi: 10.1016/j.neunet.2019.11.023. URL https://doi.org/10.1016/j.neunet.2019.11.023.

Eric T. Payne, Xiu Yan Zhao, Helena Frndova, Kristin McBain, Rohit Sharma, James S. Hutchison, and Cecil D. Hahn. Seizure burden is independently associated with short term outcome in critically ill children. *Brain*, 137(5):1429–1438, 2014. ISSN 14602156. doi: 10.1093/brain/awu042.

Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson, and Gianluca Bontempi. Calibrating probability with undersampling for unbalanced classification. *Proceedings - 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015*, pages 159–166, 2015. doi: 10.1109/SSCI.2015.33.

Gabriel M Ronen, David Buckley, Sharon Penney, and David L Streiner. Long-term prognosis in children with neonatal seizures: A population-based study. *Neurology*, 69(19):1816–1822, 2007. ISSN 00283878. doi: 10.1212/01.wnl.0000279335.85797.2c.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL http://dx.doi.org/10.1038/s42256-019-0048-x.

Renée A Shellhaas, Taeun Chang, Tammy Tsuchida, Mark S Scher, James J Riviello, Nicholas S Abend, Sylvie Nguyen, Courtney J Wusthoff, and Robert R Clancy. The American Clinical Neurophysiology Society's Guideline on Continuous Electroencephalography Monitoring in Neonates. *Journal of clinical neurophysiology : official publication of the American Electroencephalographic Society*, 28(6):611–617, 12 2011. ISSN 1537-1603 (Electronic). doi: 10.1097/WNP.0b013e31823e96d7.

Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9911 LNCS:467–482, 2016. ISSN 16113349. doi: 10.1007/978-3-319-46478-7{_}29.

N. J. Stevenson, K. Tapani, L. Lauronen, and S. Vanhatalo. A dataset of neonatal EEG recordings with seizure annotations. *Scientific Data*, 6:1–8, 2019. ISSN 20524463. doi: 10.1038/sdata.2019.39. URL http://dx.doi.org/10.1038/sdata.2019.39.

Nathan J. Stevenson, Robert R. Clancy, Sampsa Vanhatalo, Ingmar Rosén, Janet M. Rennie, and Geraldine B. Boylan. Interobserver agreement for neonatal seizure detection using multichannel EEG. *Annals of Clinical and Translational Neurology*, 2(11):1002–1011, 2015. ISSN 23289503. doi: 10.1002/acn3.249.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 07-12-June:1–9, 2015. ISSN 10636919. doi: 10.1109/CVPR.2015.7298594.

Karoliina T. Tapani, Sampsa Vanhatalo, and Nathan J. Stevenson. Time-varying EEG correlations improve automated neonatal seizure detection. *International Journal of Neural Systems*, 29(4), 2019. ISSN 17936462. doi: 10.1142/S0129065718500302.

A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. Boylan. EEG-based neonatal seizure detection with Support Vector Machines. *Clinical Neurophysiology*, 122(3):464–473, 2011a. ISSN 13882457. doi: 10.1016/j.clinph.2010.06.034. URL http://dx.doi.org/10.1016/j.clinph.2010.06.034.

A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. B. Boylan. Performance assessment for EEG-based neonatal seizure detectors. *Clinical Neurophysiology*, 122(3):474–482, 2011b. ISSN 13882457. doi: 10.1016/j.clinph.2010.06.035. URL http://dx.doi.org/10.1016/j.clinph.2010.06.035.

Andriy Temko and Gordon Lightbody. Detecting neonatal seizures with computer algorithms. *Journal of Clinical Neurophysiology*, 33(5):394–402, 2016. ISSN 15371603. doi: 10.1097/WNP.0000000000000295.

Andriy Temko, Geraldine Boylan, William Marnane, and Gordon Lightbody. Robust neonatal EEG seizure detection through adaptive background modeling. *International Journal of Neural Systems*, 23(4):5–8, 2013. ISSN 01290657. doi: 10.1142/S0129065713500184.

Andriy Temko, William Marnane, Geraldine Boylan, and Gordon Lightbody. Clinical implementation of a neonatal seizure detection algorithm. *Decision Support Systems*, 70:86–96, 2015. ISSN 01679236. doi: 10.1016/j.dss.2014.12.006. URL http://dx.doi.org/10.1016/j.dss.2014.12.006.

Werner Vach. The dependence of Cohen's kappa on the prevalence does not matter. *Journal of Clinical Epidemiology*, 58(7):655–661, 2005. ISSN 08954356. doi: 10.1016/j.jclinepi.2004.02.021.

Haishuai Wang, Zhicheng Cui, Yixin Chen, Michael Avidan, Arbi Ben Abdallah, and Alexander Kronzer. Predicting hospital readmission via cost-sensitive deep learning. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(6):1968–1978, 2018a. ISSN 15579964. doi: 10.1109/TCBB.2018.2827029.

Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018b. ISSN 0031-3203. doi: 10.1016/j.patcog.2017.08.026. URL http://dx.doi.org/10.1016/j.patcog.2017.08.026.

S O Wietstock, S L Bonifacio, J E Sullivan, K B Nash, and H C Glass. Continuous video electroencephalographic (EEG) monitoring for electrographic seizure diagnosis in neonates. *Journal of Child Neurology*, 31(3):328–332, 2016. ISSN 17088283. doi: 10.1177/0883073815592224.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80, 1945. ISSN 00994987. doi: 10.2307/3001968. URL http://www.jstor.org/stable/3001968.

Jimmy Ming-Tai Wu, Meng-Hsiun Tsai, Chia-Te Hsu, Hsien-Chung Huang, and Hsiang-Chun Chen. Intelligent signal classifier for brain epileptic EEG based on decision tree, multilayer perceptron and over-sampling approach. In Kohei Arai and Rahul Bhatia, editors, *Advances in Information and Communication*, pages 11–24, Cham, 2020. Springer International Publishing. ISBN 978-3-030-12385-7.

Qi Yuan, Weidong Zhou, Liren Zhang, Fan Zhang, Fangzhou Xu, Yan Leng, Dongmei Wei, and Meina Chen. Epileptic seizure detection based on imbalanced classification and wavelet packet transform. *Seizure*, 50:99–108, 2017. ISSN 15322688. doi: 10.1016/j.seizure.2017.05.018. URL http://dx.doi.org/10.1016/j.seizure.2017.05.018.

Zhi Hua Zhou and Xu Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1): 63–77, 2006. ISSN 10414347. doi: 10.1109/TKDE.2006.17.

Appendix A.

A.1. Duke dataset Annotation summary

Patient ID	Amount of seizures	Total hours	Total seizure hours	Seizure rate[3]
PT1	20	48.00	0.49	0.0102
PT101	141	93.19	3.20	0.0343
PT113	47	113.79	0.49	0.0043
PT114	45	95.71	1.30	0.0136
PT121	79	97.57	1.04	0.0107
PT130	6	104.09	0.54	0.0052
PT149	8	24.00	0.14	0.0057
PT151	2	24.00	0.10	0.0041
PT16	47	72.00	0.77	0.0107
PT2	38	48.00	0.70	0.0146
PT29	11	24.00	1.26	0.0525
PT32	2	101.22	0.24	0.0024
PT34	30	102.39	0.97	0.0094
PT37	3	85.87	0.46	0.0053
PT39	4	115.46	0.40	0.0034
PT4	333	48.00	11.70	0.2438
PT43	5	107.18	0.14	0.0013
PT47	89	116.93	3.76	0.0321
PT54	41	48.00	0.83	0.0172
PT60	151	48.00	7.16	0.1493
PT62	5	24.00	0.08	0.0035
PT67	15	24.00	0.60	0.0248
PT73	11	103.51	1.72	0.0166
PT77	58	24.00	0.98	0.0409
PT79	116	48.00	2.15	0.0448
PT8	64	48.00	0.53	0.0109
PT84	181	77.21	4.57	0.0592
PT91	10	107.74	0.10	0.0009
PT94	29	103.37	1.09	0.0105
PT95	148	128.10	1.96	0.0153
PT96	39	114.69	1.36	0.0119
Total	1778	2320.00	50.82	–
Mean	57.35	74.84	1.64	0.0281
Std	71.95	35.31	2.41	0.0491

A.2. Feature Extractor Architecture

Feature extractor was inspired by Inception network (Szegedy et al., 2015), allowing to extract features at multiple scales. We used an architecture with two Inception blocks, shown in Figure S1. Feature extractor had 8,514 trainable parameters in total.

3. Seizure rate =(Total amount of seizure seconds in recording)/(Total amount of seconds in recording)

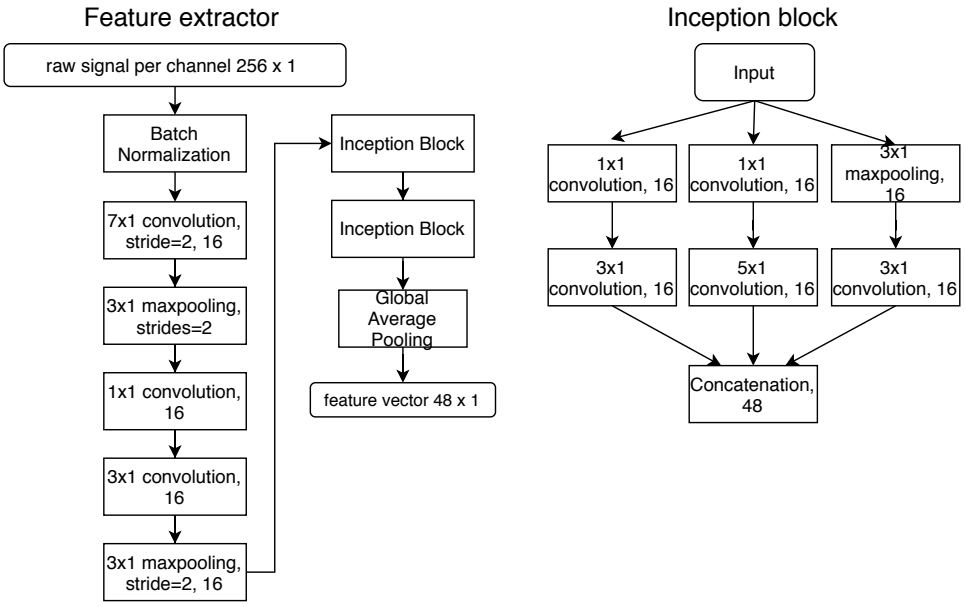

Figure S1: Schema of the feature extractor used in deep learning models. The last number in each block is the amount of filters. Strides are equal to one if not stated otherwise.

A.3. Post-processing: Probability Reweighting

Our derivations for adjustments of the classifier output probability given the prevalence of positive class follow the lines of (Pozzolo et al., 2015; Elkan, 2001). We can model our sampling strategy as follows: Let s be a Bernoulli variable defining whether an epoch is taken into the training sample or not, y a label taking two values (1 for seizure, 0 for non-seizure), X an epoch. Then

$$(s|y=i) \sim Bernoulli(\beta_i), i = \{0,1\}$$

Also, let $p(y=i) = \pi_i$.

$p(s|X,y) = p(s|y)$ since only label is important to make a decision whether an epoch is taken into the training subset for class balanced training. The strategy where we take equal amount of seizures and non-seizures into the training sample can be defined as

$$\beta_0\pi_0 = \beta_1\pi_1$$

, or

$$\frac{\beta_0}{\beta_1} = \frac{\pi_1}{\pi_0} = \beta$$

.

If $p(y=1|x,s=1)$ is the output of a classifier trained on the balanced set, then by Bayes Theorem we can write the following:

$$p(y=1|X,s=1) = \frac{p(s=1|y=1)p(y=1|X)}{p(s=1|y=1)p(y=1|X) + p(s=1|y=0)p(y=0|X)}, \quad (1)$$

where $p(y=1|x)$ is the probability of seizure in the original unbalanced model, which we are looking for. Let us denote $p = p(y=1|X), p_s = p(y=1|X, s=1)$, so we can rewrite Eq. (1) as

$$p_s = \frac{\beta_1 p}{\beta_1 p + \beta_0(1-p)} = \frac{p}{p + \beta(1-p)}$$

.

Rewriting the formula we get

$$p = \frac{\beta p_s}{\beta p_s + 1 - p_s}. \tag{2}$$

The last formula gives us an adjustment of probability that should be done on the output of our algorithm.

For later post-processing steps, we need to choose the threshold. According to Bayes decision theory, if we deem classification cost of correct examples as 0, the threshold

$$\tau = \frac{l_{1,0}}{l_{1,0} + l_{0,1}}$$

.

Without prior knowledge, we select $l_{1,0} = \pi_1$ and $l_{0,1} = \pi_0$ as the threshold (Pozzolo et al., 2015), so the threshold becomes $\tau = \pi_1$. This corresponds to an operating point of 0.5 of balanced classifier.

A.4. Post-processing: Transforming the Outputs to Improve Robustness

Post-processing was done following (Tapani et al., 2019), and took adjusted output probability per epoch as an input (see Appendix A.3). Since annotations were done by human rater on a per-second basis, post-processing had an upsampling step (converting per-epoch probabilities to per-second probabilities). Since epochs were 8-seconds long with 4-seconds overlap, each epoch prediction was transformed into 4 seconds prediction in upsampling.

The post-processing steps were as follows: a) median filtering of 3 consecutive epochs prediction probabilities; b) upsampling per-epoch predictions back to per-second resolution; c) removing all predictions labeled as 'seizure' which were less than 10 seconds long; d) "collaring" (extending each seizure prediction by 8 seconds in both directions). When computing AUC, steps a) and b) were performed before applying the decision threshold and making a binary 0/1 decision, and steps c) and d) were performed each time after applying the decision threshold.

A.5. Significance Tests of Difference in Performance of DL Models

To assess the difference in performance, as measured by AUC on leave-one patient out cross-validation, we applied Wilcoxon paired signed-rank tests each pair of models on each balancing approach. Table S1 provides the results.

A.6. Attention-MIL Measures per Patient

In this section we provide per-patient attention-MIL performance summary on Duke dataset (Table S2).

Model (balancing)	DL1 (None)	DL1 (Class)	DL1 (Patient-Class)	DL2 (None)	DL2 (Class)	DL2 (Patient-Class)
DL1 (None)	-	0.07454	0.00816	0.17014	0.01025	0.18919
DL1 (Class)	-	-	0.13132	0.86000	0.00070	0.00209
DL1 (Patient-Class)	-	-	-	0.58321	0.00002	0.00017
DL2 (None)	-	-	-	-	0.00004	0.00036
DL2 (Class)	-	-	-	-	-	0.00449
DL2 (Patient-Class)	-	-	-	-	-	-

Table S1: p-values of Wilcoxon signed-rank tests between leave-one-patient out AUCs on each level of balancing of each deep learning model

Patient	Attention AUC	Attention AUC per epoch
PT1	0.844	0.762
PT101	0.851	0.927
PT113	0.850	0.952
PT114	0.844	0.981
PT121	0.765	0.840
PT130	0.903	0.999
PT149	0.668	0.986
PT151	0.784	0.996
PT16	0.993	0.811
PT2	0.857	0.919
PT29	0.543	0.918
PT32	0.971	0.933
PT34	0.861	0.887
PT37	0.693	0.999
PT39	0.869	0.982
PT4	0.758	0.983
PT43	0.813	0.935
PT47	0.719	0.970
PT54	0.735	0.871
PT60	0.826	0.945
PT62	0.863	0.905
PT67	0.752	0.953
PT73	0.853	0.903
PT77	0.750	0.862
PT79	0.828	0.932
PT8	0.838	0.897
PT84	0.814	0.979
PT91	0.974	0.997
PT94	0.643	0.886
PT95	0.833	0.891
PT96	0.851	0.943
Mean	0.811	0.927
Std	0.096	0.058

Table S2: Attention AUC scores. Computation of 'Attention AUC' used agreement between each thresholded score and human annotations per channel per epoch; 'Attention AUC per epoch' uses agreement of at least one channel score exceeding threshold with human annotation.