Calibration Methods for Automatic Seizure Detection Algorithms

Ana Borovac[1,2], David Hringur Agustsson[1], Tomas Philip Runarsson[1], and Steinn Gudmundsson[1]

[1] Faculty of Ind. Eng., Mech. Eng. and Comput. Sci., University of Iceland, Reykjavik, Iceland
[2] Kvikna Medical ehf., Reykjavik, Iceland
{anb48, dha4, tpr, steinng}@hi.is

Abstract.
Background: Automatic seizure detection algorithms have been in development for years with the aim of making the analysis of long EEG recordings more efficient. To train such detectors, a large amount of EEG data with precise seizure annotations is required. However, due to privacy concerns, and the inherent complexity of EEG signals, obtaining data sets diverse enough to capture all relevant EEG patterns is difficult. The current state-of-the-art seizure classification algorithms are far from perfect and routinely misclassify EEG segments as seizure where there is no seizure activity and vice versa. A seizure detection algorithm that can indicate where its predictions are of low confidence, thereby requiring verification by a human expert, carries substantial real-world value. Modern seizure detectors based on deep neural networks can output probability/confidence estimates alongside seizure/non-seizure classification, but little attention has been given to how accurate these estimates are, in other words, how well the detector is calibrated.
Methods: In this study, we analyzed the calibration of seizure detectors based on a convolutional neural network, that were trained on adult and neonatal EEG data, respectively. Four calibration methods from the literature, temperature scaling, ensemble, dropout, and mixup were evaluated.
Results: We found that the uncalibrated detectors make the vast majority of the predictions with confidence close to one, i.e., they are overconfident and, therefore, the detectors with higher overall accuracy are better calibrated. Our results indicate that all the calibration methods studied here make the detectors less confident in incorrect predictions, a desirable trait, but to a lesser extent, they also result in detectors less confident in correct predictions. The best calibration was obtained with the ensemble and dropout methods. When class labels in the seizure data are highly imbalanced, it is recommended that confidence estimates for individual classes are analyzed separately.

Keywords: calibration, uncertainty, deep neural networks, automatic seizure detection, electroencephalogram

1 Introduction

Seizures are a common neurological emergency, with an estimated prevalence of 1 % [30], that can cause permanent brain damage, and even death, if untreated [45]. Almost half of the neonates affected by seizures face long-term neurodevelopmental disorders [51]. Adults experiencing seizures are at higher risk for psychiatric disorders such as depression and are about 10 times more likely to commit suicide than the general population [22]. Heart dysfunction provoked by seizures can cause sudden death [44]. To improve the lives of people with seizures, prompt detection and appropriate treatment is crucial. Treatment options include anti-epileptic drugs, brain surgery and electrical brain stimulation [27, 28, 38, 41].

The current gold standard for neonatal and adult seizure detection is an electroencephalogram (EEG), a recording of the electrical activity of the brain. EEG signal acquisition is typically done by placing electrodes on the scalp and the voltage difference between pairs of electrodes is recorded. The recordings can span from minutes to days, depending on the clinical indication for EEG monitoring. Due to signal complexity and high variability [20, 34, 53], analysis of EEG recordings requires time and special expertise that is not always available [7].

The goal of automated seizure detection algorithms (SDAs) [35, 43] is to accelerate EEG analysis significantly while preserving the current level of diagnostic accuracy. Such SDAs could enable the widespread use of EEGs, e.g. in intensive care units, without the need for experts to monitor each recording. The development of SDAs that perform as well as human experts faces two main challenges. First, due to patient privacy issues, there is often a limited amount of data available for algorithm training [10]. Second, obtaining precise annotations of seizure onset and offset times is challenging, as human experts may disagree on the presence of seizure events [9, 17]. SDAs trained on relatively small data sets are expected to have difficulties classifying unseen EEG segments accurately. Compounding the problem is the presence of label noise in the data because of ambiguity in the human annotations of the EEG [5]. Combining seizure/non-seizure predictions with confidence estimates would make the detectors more useful in a clinical setting [3, 24]. By doing so, EEG intervals with low-confidence predictions can be flagged for review by a human expert. The end result would be faster analysis without compromising the accuracy of the annotations. For example, a study using an SDA based on a support vector machine (SVM) suggests that by passing 40 % of the data with the least confident predictions to a human expert, an accuracy of 99 % could be achieved [2].

Many modern SDAs are based on deep neural networks (DNNs) [35, 43]. For a given EEG segment, a neural network classifier outputs a value between zero and one. This value can be interpreted as an estimate of the probability that the segment contains a seizure. A value close to zero indicates that the segment is unlikely to contain a seizure and a value close to one indicates that the segment most likely contains a seizure. The value can thus be regarded as the *confidence* the classifier has in the prediction. By thresholding at, say, 0.5, the segment can be classified as a seizure or non-seizure segment and labelled accordingly in the EEG recording. However, the accuracy of these confidence estimates has received limited attention in the context of SDAs [6]. In case the estimates are accurate, a classifier is considered to be well-*calibrated*. In other words, if a classifier outputs a confidence estimate of, e.g., 0.7 for some EEG segments, and it is correct in its prediction for 70 % of these segments, the classifier is well-calibrated. The same should also hold for other confidence levels.

Guo et al. [16] have reported that DNNs trained on image and document classification tasks tend to be overconfident in their predictions, despite achieving high classification accuracy. Based on their empirical results, they suggest several potential causes that result in poorly calibrated DNNs, including increased model capacity, batch normalization, training with small weight decay and using the cross-entropy loss function [54]. Thulasidasan et al. [50] suggest that training with 0/1 annotations negatively influences calibration and is improved by utilizing mixup [56]. Hein et al. [19] showed that DNNs employing the ReLU activation function can be particularly overconfident in their predictions for data far away from the training data. On the other hand, Minderer et al. [31] found that state-of-the-art DNNs for image classification tend to be well-calibrated and suggest that improvements in model accuracy benefit calibration. It should be noted that the image classification data sets employed in the above studies typically feature hundreds of classes whereas seizure detection is normally formulated as a binary classification task. Researchers have proposed various approaches to improve the calibration of DNNs [1, 12]. These methods include post-processing techniques such as isotonic regression [55] and Platt scaling [39], which adjust the output probabilities of the network in order to improve calibration. Methods such as mixup [50, 56] and dropout [11] modify the training process, and in the case of dropout, also the prediction process.

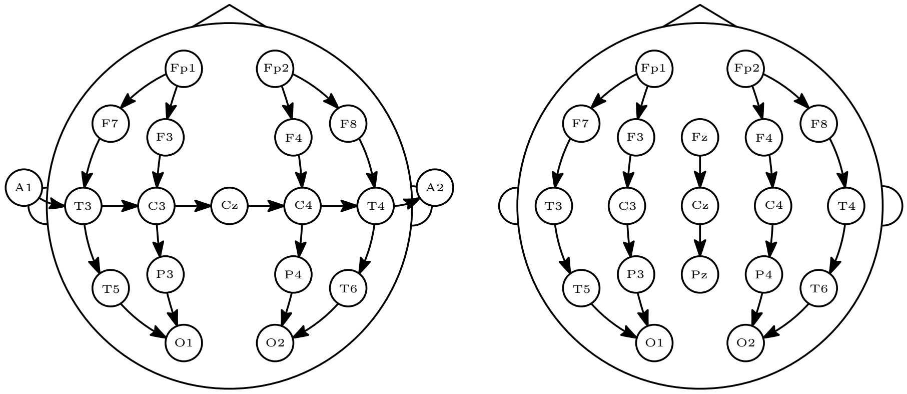
In this work, we extend our previous analysis of SDA calibration [6] by analyzing four different calibration methods that have been found to work well with DNNs, albeit in different settings. We show that neonatal and adult SDAs based on a convolution neural network are overconfident in their predictions and that detectors with higher overall accuracy are better calibrated. All the calibration methods evaluated here, temperature scaling, ensemble, dropout and mixup, make the detector less confident for incorrect predictions. A comparison of the methods is done on two publicly available data sets; one adult and one neonatal data set.

2 Methods

2.1 Data

The adult EEG data set was obtained from version 2.0.0 of the TUH EEG seizure corpus [18], which consists of recordings with diverse recording set-ups and seizure types. The most frequent seizure type is focal non-specific seizures, but other types, such as generalized non-specific seizures and complex partial seizures are also present. In this study, we utilized a subset of recordings recorded with averaged reference, i.e. average potential of all the electrodes was used as a reference. The acquisition of the signals was done with a version of a NicoletOne EEG system (Natus, USA) and the sampling frequency was between 250 Hz and 1000 Hz. Human experts annotated the recordings using a bipolar temporal central parasagittal montage with 22 channels (Fig. 1), and the same montage was employed in this study. Specifically, channels Fp1-F7, F7-T3, T3-T5, T5-O1, Fp2-F8, F8-T4, T4-T6, T6-O2, T3-C3, C3-Cz, Cz-C4, C4-T4, Fp1-F3, F3-C3, C3-P3, P3-O1, Fp2-F4, F4-C4, C4-P4, P4-O2, A1-T3 and T4-A2 were derived from the recorded signals. The data set contains predefined training, validation and test sets.

The neonatal data set used in this study consists of 79 recordings [47]. Acquisition of the EEG signals was done with NicoletOne EEG system (Natus, USA), using 19 electrodes with the reference electrode located at the midline and the sampling frequency was 256 Hz. To annotate the recordings, three human experts utilized a bipolar longitudinal (double banana) montage with 18 channels. Schematic representation of the montage is given in Fig. 1. For this study, the same set of channels was used, including 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, which were derived from the recorded EEG signals. The data set does not come with predefined training, validation and test set splits. Table 1 shows summary statistics for the two data sets. We note that the data sets are imbalanced, i.e., less than 10 % of the total recording duration corresponds to seizure segments.



(a) Montage for adult EEG recordings.

(b) Montage for neonatal EEG recordings.

Fig. 1: Montages for adult (a) and neonatal (b) EEG recordings. In both cases, the electrodes are positioned according to the 10-20 system. Each arrow denotes an EEG channel that is used as input to an SDA.

Since most adult seizures are present in the frequency range between 3 and 30 Hz [15] and neonatal seizures can be as slow as 0.5 Hz [13], the EEG signals were filtered with a Butterworth band-pass filter with cut-off frequencies 0.5 Hz and 30 Hz. Before filtering, the EEG signals were downsampled to 250 Hz and further downsampled to 62 Hz after filtering, reducing the input size and subsequently model size by approximately factor 4. After filtering and downsampling, each recording was cut into 16 seconds segments.

Table 1: Summary statistics for the adult [18] and neonatal [47] data sets used in this study. A patient "with seizures" has at least one 16 seconds seizure segment. Standard deviations are shown in parentheses.

	Adult data set			Neonatal data set
	Training	Validation	Test	
Number of patients	297	41	41	79
Total duration of recordings [hours]	603.08	372.21	119.98	111.90
Total duration of seizures [hours]	26.41	10.91	7.57	10.91
Fraction of seizure activity [%]	4.38	2.93	6.31	9.75
Average duration of recordings per patient [hours]	2.03 (3.29)	9.08 (17.85)	2.93 (2.07)	1.42 (0.56)
Average duration of seizures per patient with seizures [hours]	0.24 (0.41)	0.34 (0.42)	0.22 (0.31)	0.28 (0.38)
Number of seizure segments	19148	8066	5197	8563
Number of non-seizure segments	127220	79759	24547	20233

To increase the amount of seizure data available for training, an overlap of 12 seconds was used for the seizure segments.

2.2 Seizure Detection Algorithm

The detector takes multi-channel EEG as input and outputs seizure/non-seizure probability estimates. This is accomplished by extracting features from each EEG channel via 11 convolutional layers with 32 filters of size 3×1, followed by batch normalization layers and ReLU activation functions [36]. Average pooling is applied before the fourth, seventh, and tenth convolutional layers. An attention layer is used to combine feature vectors extracted from individual EEG channels into one feature vector [21]. The classification part of the network is a fully connected layer that maps feature vectors of dimension 58 to two outputs (seizure/non-seizure) utilizing a softmax activation function to obtain values in $(0, 1)$ and can be interpreted as class probabilities. With only $29,964$ learnable parameters, the SDA is practically tiny, compared to state-of-the-art networks used in natural language processing and computer vision. A benefit of using such a small network is that it can be deployed on devices with limited computation resources. A detailed description of the SDA is given in [4].

The adult and neonatal detectors were trained by optimizing the negative log-likelihood loss function with the Adam optimizer and a mini-batch size of 256. To address the imbalance between the number of available seizure and non-seizure segments in the training sets, each mini-batch contained 128 seizure and 128 non-seizure segments. One epoch corresponds to a single pass through all available seizure segments and an equal number of randomly selected non-seizure segments. The SDAs were trained for 50 epochs where the initial learning rate of 0.001 was halved every 20 epochs. The number of epochs and the learning rate decay were chosen so that the area under the curve computed on the adult validation set was maximized. The hyper-parameter values used in this experiment are similar to those of previous experiments [5] conducted on the neonatal data set. We observed that the performance of the SDA is insensitive to small changes in hyper-parameter values.

2.3 Calibration methods

Temperature scaling

Platt scaling [39] is a generic method to transform classifier outputs to a probability distribution over classes. It was originally proposed for use with SVM classifiers and has previously been used with an SVM-based neonatal seizure detector to smooth classifier outputs and to aggregate predictions over multiple channels [48]. The method fits a parameterized sigmoid or softmax function to (unscaled) classifier outputs. In [16] a simplified version with one learnable parameter called *temperature scaling* was used to improve the calibration of neural networks trained on image and document data. In case the non-seizure class is denoted with 0 and the seizure class with 1, the calibrated seizure/non-seizure probability estimates are obtained as follows,

$$\hat{p}_j^{(i)} = \frac{\exp\left(z_j^{(i)}/\tau\right)}{\exp\left(z_0^{(i)}/\tau\right) + \exp\left(z_1^{(i)}/\tau\right)}; \quad i = 1, 2, \ldots, N, \ j = 0, 1, \tag{1}$$

where $z_c^{(i)}$ $(c = 0, 1)$ are the unscaled outputs of instance i and $\hat{p}_j^{(i)}$ is the calibrated probability of instance i belonging to class j. The value of the τ parameter is chosen based on a held-out validation set after the network has been trained to avoid overfitting [39]. In case $\tau = 1$ the calibrated probabilities are equal to the softmax outputs and when τ is large the probabilities approach $1/2$. Applying temperature scaling to the unscaled outputs of the network does not change the predicted seizure/non-seizure labels, the only difference is in the probability (confidence) estimates.

Dropout

Dropout is a simple and widely used regularization technique for improving the generalization of DNNs [46]. The idea behind dropout is to randomly drop nodes from the network with a fixed probability, p. This forces the network to learn more robust features that are not dependent on any single node and reduce overfitting. Dropout is usually only used during training, i.e., the full network is used to obtain predictions. Dropout can also be employed in the prediction phase (Monte Carlo dropout). In this setting, the final seizure/non-seizure prediction is obtained by averaging T softmax outputs. It has been shown empirically that Monte Carlo improves the calibration of DNNs [29] and can be interpreted as approximate Bayesian inference [11]. Dropout with $p = 0.1$ was used for the convolutional and attention layers and $p = 0.5$ for the fully connected layer [11, 46]. The average of $T = 10$ softmax predictions was used for final probability estimates (averaging over a larger number of predictions gave similar results, data not shown).

Deep ensembles

An ensemble of multiple DNNs, referred to as *deep ensemble* in the following, has been shown to give small improvements in classification performance compared to the best individual model in the ensemble [23]. An added benefit of using an ensemble of DNNs is improved calibration. Lakshminarayanan et al. [26] found that an ensemble with only five models trained with the same setup can lead to a noticeable improvement in calibration. Here we used an ensemble of 10 SDAs. Each individual SDA was trained with the same training parameters and the same data. The resulting SDAs were nevertheless not identical since network weights were randomly initialized for each network prior to training. Additionally, the order in which the data was presented to the network was different since the data was randomly shuffled for every epoch. Once the detectors were trained, the final prediction was obtained by averaging the softmax outputs.

Mixup

Mixup is a data-agnostic augmentation method that has been found to improve the generalization of many neural network architectures [56]. It has also been found to improve the model calibration of classifiers for both images and text [50]. Mixup creates augmented training examples by forming linear combinations of feature-target pairs. A new feature-target $(\tilde{x}, \tilde{y})$ is generated as follows,

$$\tilde{x} = \lambda x^{(i)} + (1 - \lambda) x^{(j)}, \tag{2}$$

$$\tilde{y} = \lambda y^{(i)} + (1 - \lambda) y^{(j)}, \tag{3}$$

where $x^{(i)}$ and $x^{(j)}$ are two randomly selected training EEG segments, $y^{(i)}$ and $y^{(j)}$ are corresponding 0/1 (non-seizure/seizure) labels and $\lambda \in [0, 1]$ is a random variable drawn from a Beta distribution with hyper-parameter α. It is important to select an appropriate α to achieve good results, in this study, $\alpha = 0.3$ was used after testing several different values on the adult validation set.

2.4 Evaluation

The adult SDA was assessed using a dedicated test set, while the evaluation of the neonatal SDA was done using leave-one-subject-out cross-validation since the data set lacks a distinct test set.

The SDAs were assessed for their classification performance using the area under the curve (AUC), sensitivity (SE), and specificity (SP). Sensitivity refers to the fraction of correctly classified seizure segments, while specificity refers to the fraction of correctly classified non-seizure segments. The *confidence* of a prediction is the softmax output of the predicted seizure/non-seizure class, i.e., the class with the higher probability estimate. Since the threshold for seizure/non-seizure prediction is set at 0.5, the confidence estimates range between 0.5 and 1.0.

A *reliability diagram* [32] is a visual representation of a classifier's calibration, as shown in figure 2. The diagram shows the fraction of accurately predicted segments, providing an empirical estimation of the true underlying accuracy, against confidence levels. A well-calibrated classifier is indicated by empirical frequencies that align closely with the line of average confidence within a given bin. If there is sufficient data, the average confidence line should approximate the identity line.

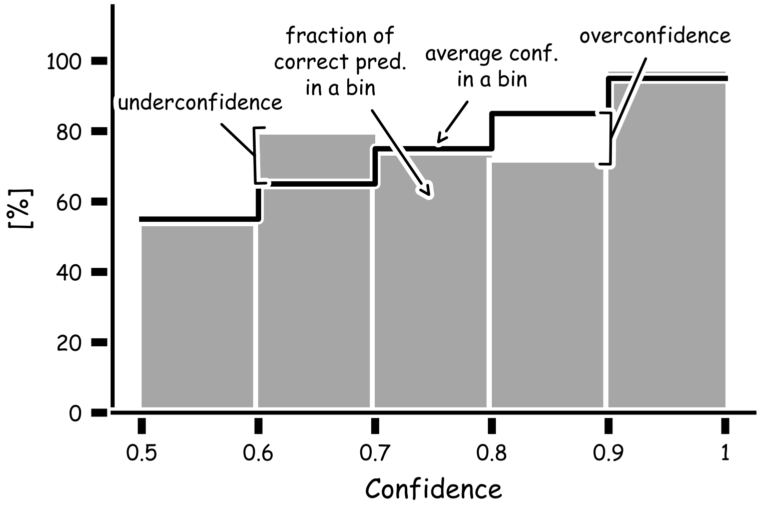


Fig. 2: Reliability diagram. The interval between the lowest (0.5) and highest (1.0) possible confidence values is split into five equally sized bins. All EEG segments are allocated to a bin based on the confidence of their predictions. Grey bars represent the fraction of the correctly predicted segments in a bin. The black curve represents the average confidence in each bin. Differences between the bars and the curve indicate miscalibration, i.e. the SDA is either underconfident or overconfident for the predictions in the bin.

To evaluate the calibration metrics, all N available seizure and non-seizure segments were split into $K = 5$ bins based on the confidence estimate made by the SDA. Bin edges were set such that the interval between the lowest (0.5) and highest (1.0) possible confidence is partitioned into equally sized intervals. The set of segments in bin k is denoted with B_k and $|B_k|$ is the number of segments in bin k. The fraction of correct predictions (*empirical frequency*) in a bin k is denoted with $\mathrm{acc}(B_k)$ and the average confidence estimate in bin k with $\mathrm{conf}(B_k)$.

The *expected calibration error* (ECE) [16],

$$\mathrm{ECE} = \sum_{k=1}^{K} \frac{|B_k|}{N} \, |\mathrm{acc}(B_k) - \mathrm{conf}(B_k)|, \tag{4}$$

measures the difference between predicted confidence and empirical frequency. Bins with more segments weigh more than bins with fewer segments. The closer the metric is to zero, the better calibrated the model is.

In medical applications, classifiers that are not overconfident in the predictions are preferred. Therefore, we include the *overconfidence error* (OE) [50] for calibration evaluation,

$$\mathrm{OE} = \sum_{k=1}^{K} \frac{|B_k|}{N} \, \mathrm{conf}(B_k) \cdot \max(\mathrm{conf}(B_k) - \mathrm{acc}(B_k), 0), \tag{5}$$

A modification of the static calibration error (SCE) [33] is proposed to capture calibration of individual classes (seizure and non-seizure) when the class frequencies differ widely,

$$\mathrm{SCE} = \frac{1}{KC} \sum_{k=1}^{K} \sum_{c=1}^{C} \frac{|B_{c_k}|}{N_c} \, |\mathrm{acc}(B_{c_k}) - \mathrm{conf}(B_{c_k})|. \tag{6}$$

In comparison with the original definition in [33], this definition differs in the weighting factor $|B_{c_k}|/N_c$, where N_c is the number of segments of class c (seizure or non-seizure). Here the weights are proportional to the number of segments in each class and not the total number of segments. As a result, all the classes have the same weight in the overall sum and the imbalanced data issue is addressed. In other words, the static calibration error is the average expected calibration error using segments of just one class.

We also include two calibration metrics which measure the distance of the estimated probability to the target label. Specifically, the Brier score (BS) [8],

$$\mathrm{BS} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{y}^{(i)} - y^{(i)} \right)^2, \tag{7}$$

and negative log-likelihood (NLL) [40],

$$\mathrm{NLL} = -\sum_{i=1}^{N} y^{(i)} \log \hat{y}^{(i)} + \left(1 - y^{(i)}\right) \log \left(1 - \hat{y}^{(i)}\right), \tag{8}$$

where $\hat{y}^{(i)}$ is the softmax output of instance i for class 1 (seizure class) and $y^{(i)}$ is the target label of instance i.

2.5 Implementation

All code was written in Python 3.9. EEG recordings in EDF format were read with the MNE library [14] (version 0.24.1) and pre-processed with SciPy [52] (version 1.8.0). The detectors were developed with PyTorch [37] (version 1.11.0) and an NVIDIA GeForce GTX 1080 Ti graphics card. The code is available at a GitHub repository (github.com/anaborovac/Calibrated-SDA).

3 Results and Discussion

In the following, we refer to an SDA which does not utilize any specific calibration method as *uncalibrated*. We show that the calibration of detectors is highly correlated with overall classification accuracy as most correct and incorrect predictions appear to have high confidence. Using calibration methods does not result in perfectly calibrated SDAs, however, lower confidence in incorrect predictions is obtained with all the methods, temperature scaling, ensemble, dropout and mixup.

3.1 Tuning hyper-parameters

The number of training epochs, learning decay schedule and the α parameter in mixup were optimized by maximizing the AUC on the adult validation set. The neonatal data set is relatively small and it is therefore costly to set aside separate data for validation. Instead of reducing the amount of neonatal data available for training, we decided to simply train the neonatal detector using the same hyper-parameters that we obtained for the adult detector. Fig. 3 shows the negative log-likelihood loss and AUC during training on the adult data sets. The validation loss fluctuates significantly but the AUC is relatively stable after approx. 30 epochs. Longer training and different weight decay schedules gave similar results (data not shown). The fluctuations in the validation loss may be due to a small and imbalanced data set, overfitting or ambiguity in annotations of seizure/non-seizure segments in the data sets.

3.2 Uncalibrated SDAs

To construct the deep ensembles, 10 sets of uncalibrated adult and neonatal SDAs were obtained by starting from random initial weights. We begin by analyzing these detectors individually to gain insight into the variability in the classification and calibration performance of individual classifiers.

Fig. 4 shows that the performance of the adult SDAs is on average slightly better in comparison with the neonatal SDAs. This is not unexpected since the adult training set is significantly larger. The figure also shows that neonatal SDAs have less variability than adult SDAs. This may simply be a consequence of the use of leave-one-subject-out cross-validation on the neonatal data set since averaging over 38 detectors has a smoothing effect. Fig. 4 also shows the expected trade-off between sensitivity and specificity. Detectors with high seizure detection rates incorrectly classify more non-seizure segments as seizures and vice versa. However, since the AUC values are similar for

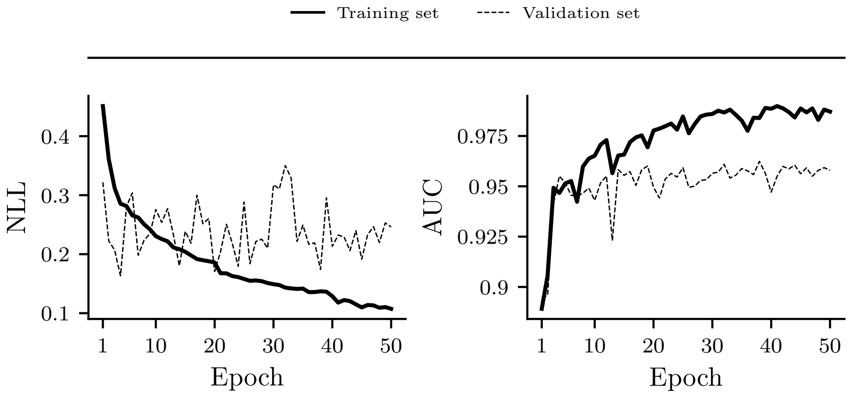


Fig. 3: Left: Negative log-likelihood (NLL) loss on the adult data set during training of an uncalibrated SDA. Right: Corresponding area under the curve (AUC) values.

all adult and neonatal detectors, respectively, the threshold for seizure/non-seizure prediction may be adjusted to achieve the desired classification performance. Adult SDAs exhibit considerable variance in sensitivity. A possible explanation is that the number of seizure segments available for training is much lower than the number of non-seizure segments. This may result in detectors that are not able to accurately capture the relevant

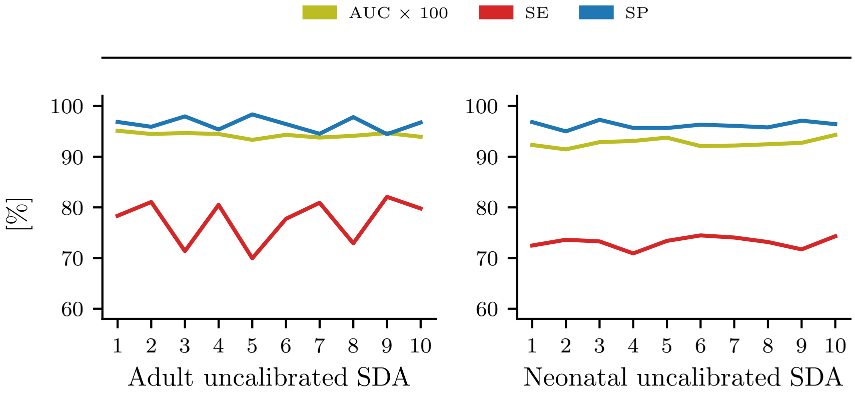


Fig. 4: Area under the curve (AUC), sensitivity (SE) and specificity (SP) for individual uncalibrated adult (left) and neonatal (right) SDAs from the deep ensembles (arbitrary order). Metrics are averaged across all patients which have at least one seizure segment. A separate test set is used to compute metrics for the adult data set while leave-one-subject-out cross-validation is used to compute the metrics for the neonatal data set.

features that differentiate seizures from non-seizure segments. Other factors that could contribute to higher variance are the heterogeneity of the different seizure types and incorrect annotations due to ambiguity in the scoring of EEGs by human experts [9, 17].

Fig. 5 shows the expected and static calibration errors for individual adult and neonatal detectors. In both cases, the expected calibration and overconfidence errors were practically identical (data not shown). This means that the SDAs are overconfident in their predictions, i.e. they are incorrect more frequently than the probabilities returned by the SDAs indicate. This can partly be explained by the use of ReLU activation functions [19] and batch normalization layers [16] in the detectors. The calibration may have been further compromised due to the training of the detectors using binary labels (non-seizure/seizure) [50] and cross-entropy loss function [54].

For the neonatal data set the expected and static calibration errors were very similar since the fraction of seizure segments in left-out neonatal patients is 49 % whereas for the adult test set only 17 % of the segments are seizure segments. Since static calibration error is an average of expected calibration errors calculated separately for seizure and non-seizure segments, it is more suitable for imbalanced data sets (e.g., the adult data set) than the expected calibration error.

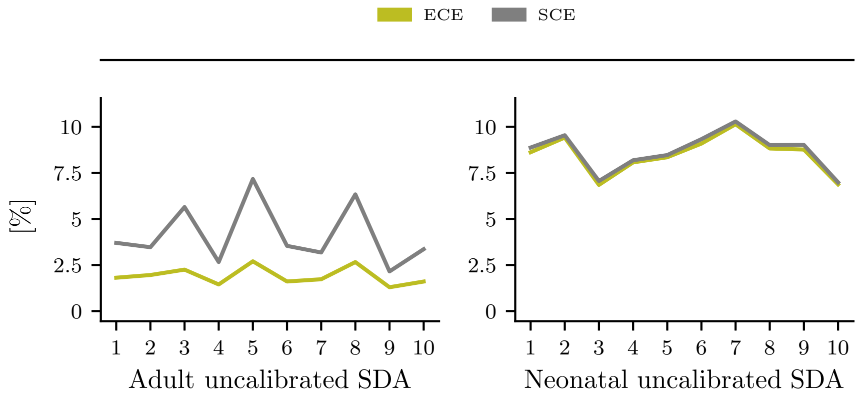


Fig. 5: Expected calibration error (ECE) and static calibration error (SCE) for individual uncalibrated adult (left) and neonatal (right) SDAs from the deep ensembles. The metrics are calculated based on all available segments in the test set for adult SDAs and left-out patients for neonatal SDAs.

Due to the imbalance in the data sets we decided to investigate the calibration of seizure and non-seizure classes individually. Fig. 6 shows that lower sensitivity/specificity results in higher expected calibration error, i.e. in worse calibration. This is a consequence of most (above 83 %) segments being predicted with confidence greater than 0.9. From equation (4) it follows that the bin with segments predicted with the highest confidence affects the expected calibration error the most. Therefore, if the overall accuracy is very high and also close to the overall confidence, the detector is well-calibrated. The figure shows that the expected calibration error is higher for the seizure segments than for the non-seizure segments. However, there are more non-seizure segments in the adult test set and therefore calibration on the non-seizure segments weighs more in the computation of the expected calibration error.

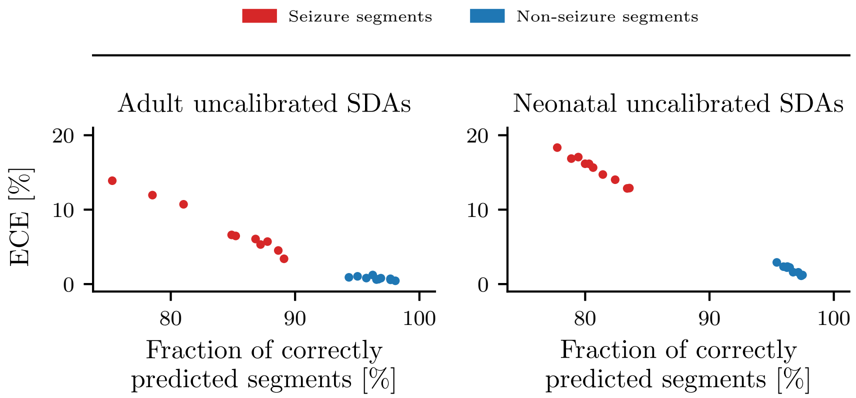


Fig. 6: Expected calibration error (ECE) calculated using seizure (red) and non-seizure (blue) segments. Each red and blue point represents one uncalibrated adult (left) and neonatal (right) SDA in the deep ensembles. The metrics are calculated from all available segments in the test set for the adult SDAs and from the left-out patients for neonatal SDAs.

3.3 Calibrated SDAs

The AUC was averaged over all SDAs in an ensemble and the individual classifier with AUC closest to the ensemble average was selected as a representative uncalibrated SDA in the following. Table 2 shows how different calibration methods affect the performance of the adult and neonatal SDAs. The classification performance of the adult SDAs using temperature scaling is identical to the uncalibrated classifier since the scaling procedure only affects confidence estimates and not the predicted class. The method is therefore not listed separately in the table. Temperature scaling was not applied to the neonatal SDAs since there was no dedicated validation set available for tuning the temperature parameter τ.

The performance metrics in Table 2 have fairly large variance across patients for all the SDAs, the sensitivity metric in particular. A likely explanation is that there are far fewer seizure segments (13 %) in the training set, in comparison to non-seizure segments (Table 1). Although each mini-batch is balanced during training, the seizure class has fewer

Table 2: Patient-based classification metrics for uncalibrated and calibrated adult and neonatal SDAs. Metrics are averaged across patients with at least one 16 seconds long seizure segment. For uncalibrated detectors, the range of values for all detectors in the ensemble is reported. Standard deviations are shown in parentheses.

	Uncalibrated	Calibrated		
		Ensemble	Dropout	Mixup
Adult SDA				
Area under the curve	0.94 (0.11)	0.95 (0.11)	0.95 (0.11)	0.96 (0.09)
Sensitivity [%]	77.74 (26.51)	79.72 (25.94)	79.57 (26.26)	77.77 (23.62)
Specificity [%]	96.44 (5.44)	97.51 (4.07)	97.0 (4.14)	96.45 (5.03)
Neonatal SDA				
Area under the curve	0.93 (0.12)	0.95 (0.10)	0.92 (0.14)	0.92 (0.13)
Sensitivity [%]	71.71 (30.77)	74.11 (27.84)	71.67 (28.68)	68.94 (31.48)
Specificity [%]	97.11 (7.04)	97.28 (7.95)	94.07 (12.77)	96.91 (5.98)

examples defining it. Furthermore, the metrics are only based on 31 adult and 38 neonatal patients, respectively, and differences in performance for 2 – 3 patients can end up having a significant effect on the overall mean and standard deviation. In both data sets, there are approx. three patients that the SDAs consistently had problems classifying and result in sensitivity values below 50 % and/or specificity below 90 %. The heterogeneity of the different seizure types and ambiguity in EEG signals are likely to contribute to the variability as well. From Table 2 we conclude that calibration methods do not outperform uncalibrated detectors, but they also do not noticeably degrade classification performance in terms of the area under the curve, sensitivity and specificity. This is in line with previous studies which applied ensembling, dropout and mixup for image classification [23, 25, 49, 50, 57].

The calibration metrics in Table 3 were computed by averaging over all available test segments, instead of first computing the corresponding metrics over the patients and then averaging. The reason is that the number of segments behind each patient varies widely. For some patients, there are fewer than 100 segments and this causes difficulties when computing metrics based on confidence bins. Overall, large improvements in calibration were not observed for either data set. However, we observe that all the calibration methods reduce the overconfidence error, a highly desired feature in medical applications. An overconfidence error close to zero for the adult data set implies that the calibrated SDAs are mainly underconfident in their predictions since the expected calibration errors are non-zero. The neonatal SDAs are overconfident, after employing the calibration methods, but the level of overconfidence error has decreased.

Table 3: Segment-based calibration metrics for uncalibrated and calibrated adult and neonatal SDAs. The metrics were calculated on all available segments in a test set for the adult SDAs and on the left-out patients for the neonatal SDAs.

	Uncalibrated	Calibrated			
		Temp. scaling	Ensemble	Dropout	Mixup
Adult SDA					
Expected cal. error [%]	1.61	2.58	0.62	1.58	4.05
Overconfidence error [%]	1.55	0.0	0.03	0.0	0.0
Static calibration error [%]	3.54	3.21	2.02	2.31	3.65
Brier score	0.04	0.04	0.03	0.04	0.04
Negative log-likelihood	0.17	0.17	0.13	0.14	0.16
Neonatal SDA					
Expected cal. error [%]	8.76	-	4.72	4.80	1.76
Overconfidence error [%]	8.76	-	4.72	4.80	1.51
Static calibration error [%]	9.01	-	5.40	5.27	7.62
Brier score	0.10	-	0.08	0.09	0.11
Negative log-likelihood	0.59	-	0.29	0.32	0.35

For the neonatal SDAs, a large difference between expected and static calibration errors was not expected since the data set is balanced. For mixup, however, the two metrics differed considerably (Table 3). This indicates that the detector is overconfident for segments of one class and underconfident for the other. When the seizure and non-seizure components of the metrics are analyzed separately (Fig. 8), it appears that the SDA with mixup is overconfident in predicting seizure segments and not confident enough when predicting non-seizure segments.

The predicted confidence values are analysed in more detail in Fig. 9 where the confidence estimates of correct and incorrect predictions are analyzed separately. Diagrams with similar patterns are obtained when only seizure or non-seizure segments are used (data not shown). The uncalibrated SDAs are confident in the predictions, both correct and incorrect, and most of them have confidence close to one. As much as it is desired that the SDAs are confident in their correct predictions, it is also important that incorrect predictions are made with lower confidence, making it possible to inform the user that some parts of EEG are difficult to classify. In this case, binary seizure/non-seizure predictions can be accompanied by confidence values as illustrated in Fig. 7. For all the calibration methods, the number of incorrect predictions in the most confident bins is

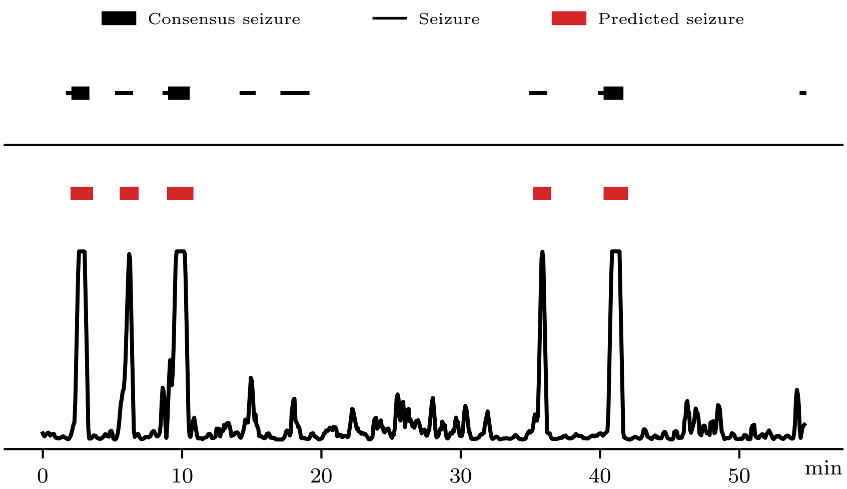


Fig. 7: Predictions for a short neonatal EEG recording (< 1 hour), obtained with an SDA employing dropout. Black blocks represent consensus seizures where the three human experts scoring the recording were in agreement. Black lines represent seizures annotated by at least one of the three experts. Red blocks represent seizure predictions and the corresponding probability estimates (confidence values) are denoted with a black curve.

clearly reduced, which is what we want, but the SDAs are less confident in their correct predictions compared to uncalibrated SDAs. However, in the latter case, the reduction is fairly small and mainly the bin with the second confidence increases in size.

Mixup results in an SDA with the lowest average confidence among the calibration methods studied here and with the lowest number of segments in the most confident bin. This also means that the largest number of incorrect predictions are, as preferred, predicted with confidence close to 0.5. In the clinical setting, this would imply that more segments would be passed for an additional review done by a human expert, but only a few incorrectly classified would be missed. This is especially noticeable for neonatal SDA. Note that the hyper-parameters were tuned on the adult validation set and different results could be obtained if they were to be tuned on neonatal data.

4 Conclusion

In line with a previous study [6], we find that uncalibrated SDAs tend to be overconfident in their predictions and the probability corresponding to an incorrect prediction gives little indication that the prediction is wrong. Since most predictions are made with confidence close to one, a more accurate detector is also better calibrated.

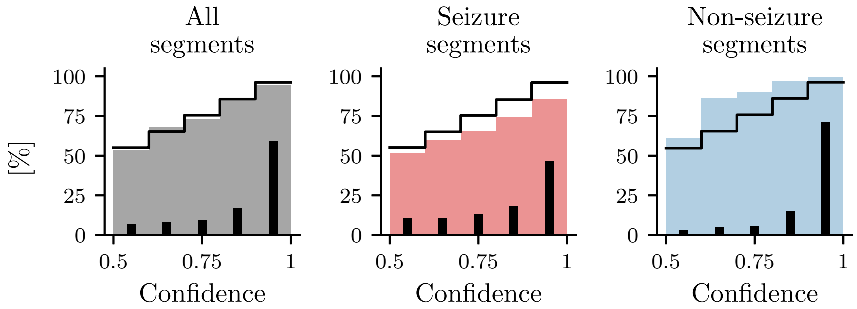


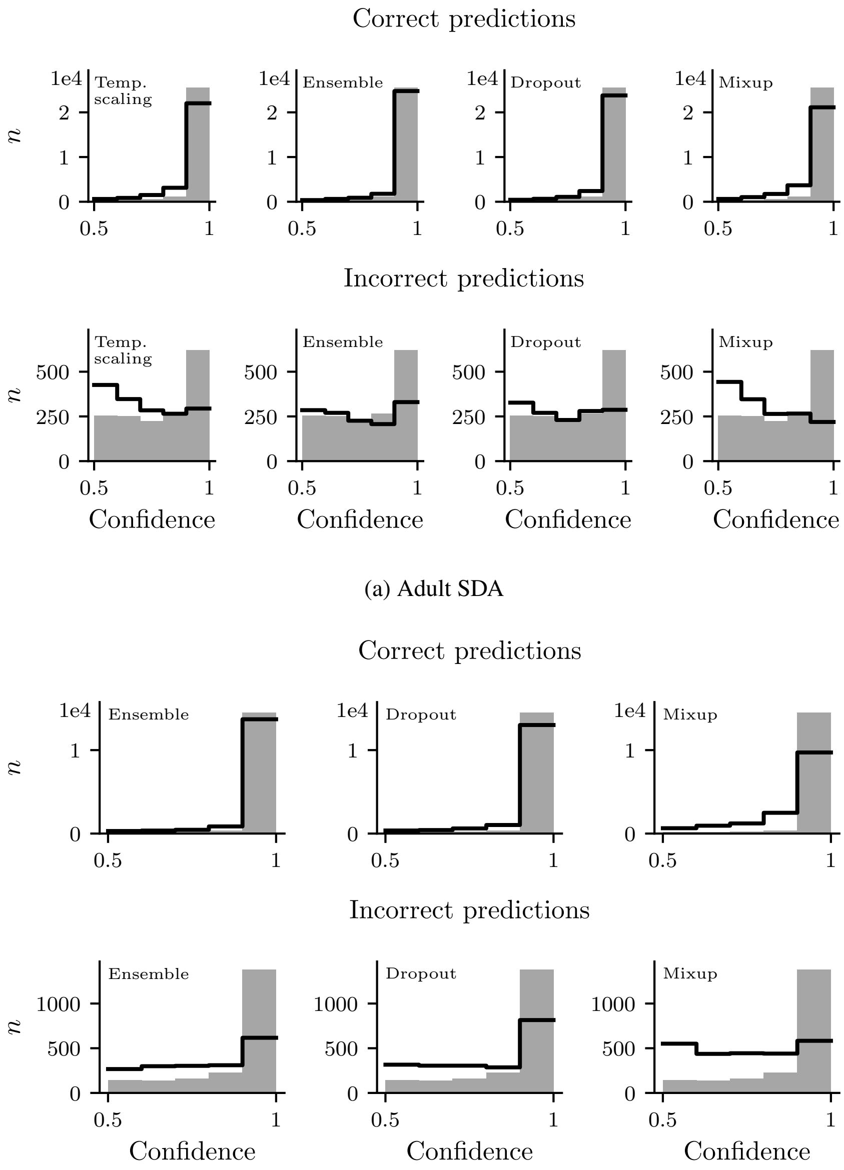
Fig. 8: Reliability diagrams for a neonatal SDA trained with mixup. The black step function indicates the average confidence of segments in each bin. Coloured bars indicate the fraction of correctly predicted segments in a bin. Black bars indicate the fraction of segments in a bin. The detector is overconfident for seizure segments and underconfident for non-seizure segments leading to an expected calibration error of 1.76 and a static calibration error of 7.62.

The four calibration methods included in this study did not degrade classification performance, i.e. their sensitivity, specificity and AUC values were similar to the uncalibrated SDAs. A slight improvement among the classification metrics for the adult SDA utilizing ensembling, dropout or mixup, was observed. These methods can be regarded as regularisation techniques that reduce model overfitting and improve generalization which in turn can explain increased detector accuracy.

With some additional computational costs we found a modest improvement in calibration, with the ensemble method giving the largest improvement, followed by dropout. Mixup gave mixed results for the adult SDA but did better on the neonatal data. Temperature scaling gave little improvement.

In [6] we found that dropout gave a noticeable improvement in expected calibration error for both adult and neonatal, SDAs. A possible explanation for this discrepancy is that here we are starting from a more accurate uncalibrated classifier than in our earlier work which then tends to be better calibrated [31]. The pre-processing and evaluation steps used here are slightly different from previous studies which also contributes to the difference. In this study, the EEG data was not normalized prior to feeding it to the network, and the mini-batch size was larger. This resulted in slightly more accurate uncalibrated SDAs than before. Additionally, here the non-seizure segments do not overlap and consequently, seizure segments represent a bigger portion of the test data. Since calibration on these segments tends to be worse than on the non-seizure segments, the overall expected calibration error is higher. Analyzing the calibration of each class is therefore advised in case of class imbalance.

All the calibrated detectors were noticeably less confident in incorrect predictions compared to uncalibrated detectors. EEG segments with low confidence values can then be



(a) Adult SDA

(b) Neonatal SDA

Fig. 9: Gray histograms represent the number of correctly and incorrectly classified segments of an arbitrarily chosen uncalibrated adult and neonatal SDA. The step functions represent the number of correctly and incorrectly classified segments.

passed to a human expert for manual review and eventual correction. This is a desirable property of an SDA if the main objective is to develop a detector that is as accurate as possible. However, in order for the SDA to be useful in clinical practice and make reviewing more time-efficient, the expert should not be given the majority of the acquired data for review. To limit the amount of data that requires human expertise, the confidence in correct predictions should be close to one and these EEG segments would therefore not be passed on to the expert. This pattern was observed for the ensemble and dropout, for temperature scaling and mixup the drop in confidence for correct predictions was larger and unfavourable.

Further work is needed to evaluate if such a setup makes EEG scoring more efficient in the clinical environment. Introducing confidence estimates alongside binary seizure/non-seizure predictions to the EEG monitors could however confound the interpretation of the user. A study on how to best present the confidence estimates is therefore needed. Another possibility would be to present to the user only a few selected EEG segments from which the detector would learn and subsequently improve. The segments could e.g., be chosen with an active learning approach [42].

Bibliography

[1] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U.R., et al.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Information Fusion **76**, 243–297 (2021)

[2] Becker, T., Vandecasteele, K., Chatzichristos, C., Van Paesschen, W., Valkenborg, D., Van Huffel, S., De Vos, M.: Classification with a deferral option and low-trust filtering for automated seizure detection. Sensors **21**(4), 1046 (2021)

[3] Begoli, E., Bhattacharya, T., Kusnezov, D.: The need for uncertainty quantification in machine-assisted medical decision making. Nature Machine Intelligence **1**(1), 20–23 (2019)

[4] Borovac, A., Gudmundsson, S., Thorvardsson, G., Moghadam, S.M., Nevalainen, P., Stevenson, N., Vanhatalo, S., Runarsson, T.P.: Ensemble learning using individual neonatal data for seizure detection. IEEE journal of translational engineering in health and medicine **10**, 1–11 (2022)

[5] Borovac, A., Guðmundsson, S., Thorvardsson, G., Runarsson, T.P.: Influence of human-expert labels on a neonatal seizure detector based on a convolutional neural network. In: The NeurIPS 2021 Data-Centric AI Workshop (2021)

[6] Borovac, A., Runarsson, T.P., Thorvardsson, G., Gudmundsson, S.: Calibration of Automatic Seizure Detection Algorithms. In: 2022 IEEE Signal Processing in Medicine and Biology Symposium (SPMB). pp. 1–6. IEEE (2022)

[7] Boylan, G., Burgoyne, L., Moore, C., O'Flaherty, B., Rennie, J.: An international survey of EEG use in the neonatal intensive care unit. Acta paediatrica **99**(8), 1150–1155 (2010)

[8] Brier, G.W., et al.: Verification of forecasts expressed in terms of probability. Monthly weather review **78**(1), 1–3 (1950)

[9] Dereymaeker, A., Ansari, A.H., Jansen, K., Cherian, P.J., Vervisch, J., Govaert, P., De Wispelaere, L., Dielman, C., Matic, V., Dorado, A.C., et al.: Interrater agreement in visual scoring of neonatal seizures based on majority voting on a web-based system: the neoguard eeg database. Clinical Neurophysiology **128**(9), 1737–1745 (2017)

[10] Eicher, J., Bild, R., Spengler, H., Kuhn, K.A., Prasser, F.: A comprehensive tool for creating and evaluating privacy-preserving biomedical prediction models. BMC Medical Informatics and Decision Making **20**(1), 1–14 (2020)

[11] Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)

[12] Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al.: A survey of uncertainty in deep neural networks. arXiv preprint arXiv:2107.03342 (2021)

[13] Gotman, J.: Automatic detection of epileptic seizures. Handbook of clinical neurophysiology **3**, 155–165 (2003)

[14] Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., et al.: MEG and EEG data analysis with MNE-Python. Frontiers in neuroscience p. 267 (2013)

[15] Grewal, S., Gotman, J.: An automatic warning system for epileptic seizures recorded on intracerebral EEGs. Clinical neurophysiology **116**(10), 2460–2472 (2005)

[16] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)

[17] Halford, J., Shiau, D., Desrochers, J., Kolls, B., Dean, B., Waters, C., Azar, N., Haas, K., Kutluay, E., Martz, G., et al.: Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG recordings. Clinical Neurophysiology **126**(9), 1661–1669 (2015)

[18] Harati, A., Lopez, S., Obeid, I., Picone, J., Jacobson, M., Tobochnik, S.: The TUH EEG CORPUS: A big data resource for automated EEG interpretation. In: 2014 IEEE signal processing in medicine and biology symposium (SPMB). pp. 1–5. IEEE (2014)

[19] Hein, M., Andriushchenko, M., Bitterwolf, J.: Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 41–50 (2019)

[20] Hrachovy, R.A., Mizrahi, E.M.: Atlas of neonatal electroencephalography. Springer Publishing Company (2015)

[21] Isaev, D.Y., Tchapyjnikov, D., Cotten, C.M., Tanaka, D., Martinez, N., Bertran, M., Sapiro, G., Carlson, D.: Attention-based network for weak labels in neonatal seizure detection. Proceedings of machine learning research **126**, 479 (2020)

[22] Jones, J.E., Hermann, B.P., Barry, J.J., Gilliam, F.G., Kanner, A.M., Meador, K.J.: Rates and risk factors for suicide, suicidal ideation, and suicide attempts in chronic epilepsy. Epilepsy & Behavior **4**, 31–38 (2003)

[23] Ju, C., Bibaut, A., van der Laan, M.: The relative performance of ensemble methods with deep convolutional neural networks for image classification. Journal of Applied Statistics **45**(15), 2800–2818 (2018)

[24] Kompa, B., Snoek, J., Beam, A.L.: Second opinion needed: communicating uncertainty in medical machine learning. NPJ Digital Medicine **4**(1), 1–6 (2021)

[25] Krishnan, R., Tickoo, O.: Improving model calibration with accuracy versus uncertainty optimization. Advances in Neural Information Processing Systems **33**, 18237–18248 (2020)

[26] Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)

[27] Lamberink, H.J., Otte, W.M., Bluemcke, I., Braun, K.P., Aichholzer, M., Amorim, I., Aparicio, J., Aronica, E., Arzimanoglou, A., Barba, C., et al.: Seizure outcome and use of antiepileptic drugs after epilepsy surgery according to histopathological diagnosis: a retrospective multicentre cohort study. The Lancet Neurology **19**(9), 748–757 (2020)

[28] Le, V.T., Abdi, H.H., Sánchez, P.J., Yossef, L., Reagan, P.B., Slaughter, L.A., Firestine, A., Slaughter, J.L.: Neonatal antiepileptic medication treatment patterns: a decade of change. American journal of perinatology **38**(05), 469–476 (2021)

[29] Leibig, C., Allken, V., Ayhan, M.S., Berens, P., Wahl, S.: Leveraging uncertainty information from deep neural networks for disease detection. Scientific reports **7**(1), 1–14 (2017)

[30] Litt, B., Echauz, J.: Prediction of epileptic seizures. The Lancet Neurology **1**(1), 22–30 (2002)

[31] Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., Lucic, M.: Revisiting the calibration of modern neural networks. Advances in Neural Information Processing Systems **34**, 15682–15694 (2021)

[32] Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning. pp. 625–632 (2005)

[33] Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring Calibration in Deep Learning. In: CVPR Workshops. vol. 2 (2019)

[34] Noachtar, S., Rémi, J.: The role of EEG in epilepsy: a critical review. Epilepsy & Behavior **15**(1), 22–33 (2009)

[35] Olmi, B., Frassineti, L., Lanata, A., Manfredi, C.: Automatic Detection of Epileptic Seizures in Neonatal Intensive Care Units Through EEG, ECG and Video Recordings: A Survey. IEEE Access **9**, 138174–138191 (2021)

[36] O'Shea, A., Lightbody, G., Boylan, G., Temko, A.: Investigating the impact of CNN depth on neonatal seizure detection performance. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 5862–5865. IEEE (2018)

[37] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)

[38] Perucca, E., Brodie, M.J., Kwan, P., Tomson, T.: 30 years of second-generation antiseizure medications: impact and future perspectives. The Lancet Neurology **19**(6), 544–556 (2020)

[39] Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)

[40] Quinonero-Candela, J., Rasmussen, C.E., Sinz, F., Bousquet, O., Schölkopf, B.: Evaluating predictive uncertainty challenge. In: Machine Learning Challenges Workshop. pp. 1–27. Springer (2006)

[41] Razavi, B., Rao, V.R., Lin, C., Bujarski, K.A., Patra, S.E., Burdette, D.E., Geller, E.B., Brown, M.G.M., Johnson, E.A., Drees, C., et al.: Real-world experience with direct brain-responsive neurostimulation for focal onset seizures. Epilepsia **61**(8), 1749–1757 (2020)

[42] Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X.: A survey of deep active learning. ACM computing surveys (CSUR) **54**(9), 1–40 (2021)

[43] Saminu, S., Xu, G., Shuai, Z., Abd El Kader, I., Jabire, A.H., Ahmed, Y.K., Karaye, I.A., Ahmad, I.S.: A recent investigation on detection and classification of epileptic seizure techniques using EEG signal. Brain Sciences **11**(5), 668 (2021)

[44] Schuele, S.U.: Effects of seizures on cardiac function. Journal of clinical neurophysiology **26**(5), 302–308 (2009)

[45] Scott, R.C.: What are the effects of prolonged seizures in the brain? Epileptic Disorders **16**(s1), S6–S11 (2014)

[46] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research **15**(1), 1929–1958 (2014)

[47] Stevenson, N.J., Tapani, K., Lauronen, L., Vanhatalo, S.: A dataset of neonatal EEG recordings with seizure annotations. Scientific data **6**, 190039 (2019)

[48] Temko, A., Thomas, E., Marnane, W., Lightbody, G., Boylan, G.: EEG-based neonatal seizure detection with support vector machines. Clinical Neurophysiology **122**(3), 464–473 (2011)

[49] Thagaard, J., Hauberg, S., Vegt, B.v.d., Ebstrup, T., Hansen, J.D., Dahl, A.B.: Can you trust predictive uncertainty under real dataset shifts in digital pathology? In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 824–833. Springer (2020)

[50] Thulasidasan, S., Chennupati, G., Bilmes, J.A., Bhattacharya, T., Michalak, S.: On mixup training: Improved calibration and predictive uncertainty for deep neural networks. Advances in Neural Information Processing Systems **32** (2019)

[51] Uria-Avellanal, C., Marlow, N., Rennie, J.M.: Outcome following neonatal seizures. In: Seminars in Fetal and Neonatal Medicine. vol. 18, pp. 224–232. Elsevier (2013)

[52] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al.: SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature methods **17**(3), 261–272 (2020)

[53] Webb, L., Kauppila, M., Roberts, J.A., Vanhatalo, S., Stevenson, N.J.: Automated detection of artefacts in neonatal EEG with residual neural networks. Computer Methods and Programs in Biomedicine **208**, 106194 (2021)

[54] Wei, H., Xie, R., Cheng, H., Feng, L., An, B., Li, Y.: Mitigating neural network overconfidence with logit normalization. In: International Conference on Machine Learning. pp. 23631–23644. PMLR (2022)

[55] Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 694–699 (2002)

[56] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)

[57] Zhang, Z., Dalca, A.V., Sabuncu, M.R.: Confidence calibration for convolutional neural networks using structured dropout. arXiv preprint arXiv:1906.09551 (2019)