Calibration of Automatic Seizure Detection Algorithms

A. Borovac, T. P. Runarsson, G. Thorvardsson and S. Gudmundsson

IEEE SPMB 2022

Seizures

Seizures are common; 1 – 5 / 1000 live births (neonates), 3 – 5 / 100 people (adults).

Seizures can cause brain damage and may be life-threatening.

Gold standard diagnostics: Multi-channel video electroencephalogram (EEG).

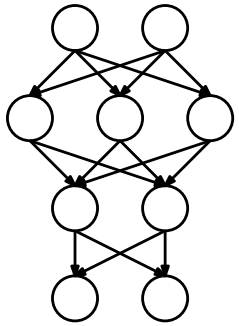
Reviewing hours of EEG is time-consuming and requires special expertise.

→ Automatic EEG seizure detection algorithm

Pisani et al., Incidence of neonatal seizures, perinatal risk factors for epilepsy and mortality after neonatal seizures in the province of Parma, Italy, 2018; Vasudevan et al., Epidemiology and aetiology of neonatal seizures, 2013; Kobau et al., Prevalence of self-reported epilepsy or seizure disorder and its associations with self-reported depression and anxiety: results from the 2004 Healthstyles Survey, 2016; Litt et al., Prediction of epileptic seizures, 2002

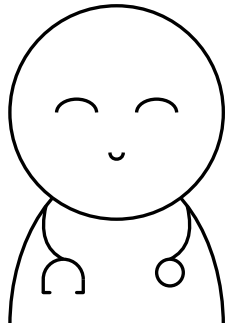
Automatic Seizure Detection Algorithm (SDA)

SDA with human-level performance



Usually trained on small amount of data

Practice

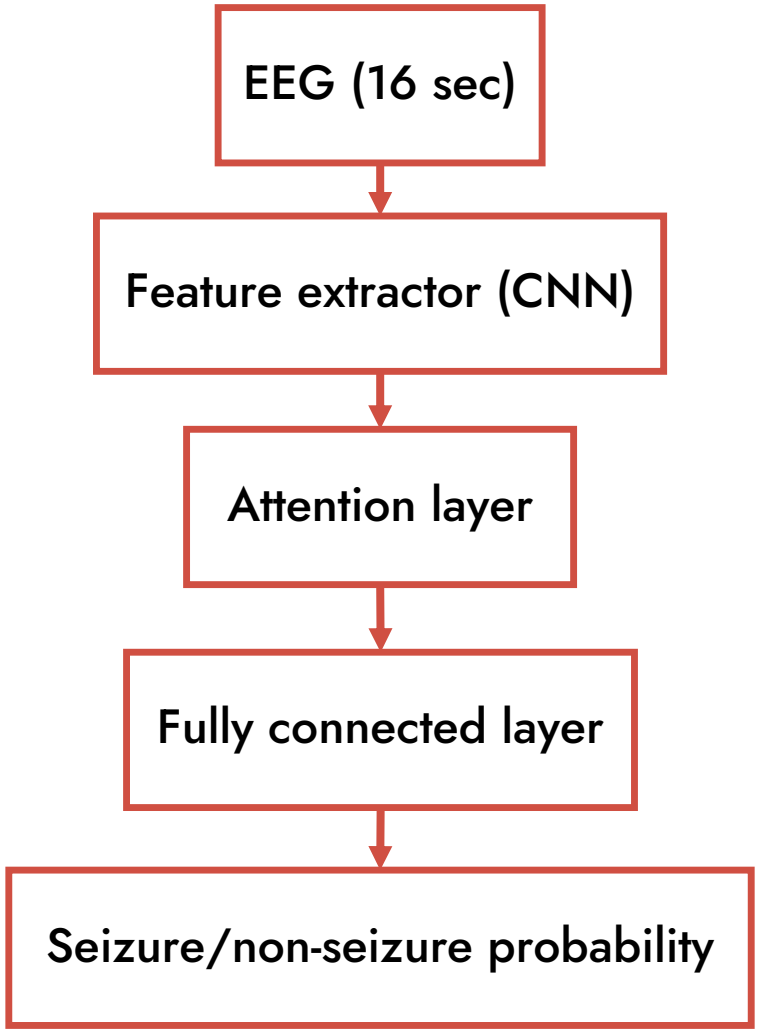


Trust

Can the SDA inform the user when it is not confident in its prediction?

Is the SDA well-*calibrated*, i.e. are probability estimates for seizure/non-seizure predictions close to empirical frequencies?

SDA: Architecture



Input

Extracts features from each EEG channel separately

Aggregates any number of feature vectors into one feature vector

Maps features to two outputs

Output

Optimised for neonatal seizure detection.

O'Shea et al., Investigating the impact of CNN depth on neonatal seizure detection performance, 2018; Isaev et al., Attention-Based Network for Weak Labels in Neonatal Seizure Detection, 2020

Training

Neonatal SDA (Helsinki Dataset)	Adult SDA (TUH EEG Seizure Corpus)
79 recordings	297 + 41 + 41 recordings
0.5 – 16 Hz	0.5 – 25 Hz
Leave-one-subject-out cross-validation	Separate test set

Stevenson et al., A dataset of neonatal EEG recordings with seizure annotations, 2019; Harati et al., The TUH EEG CORPUS: A big data resource for automated EEG interpretation, 2014

Performance Evaluation

Classification performance: area under the curve (AUC), sensitivity (SE) and specificity (SP).

Calibration: expected calibration error (ECE) and overconfidence error (OE).

sum over K bins

portion of correctly classified segments in bin k

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

portion of segments in bin k

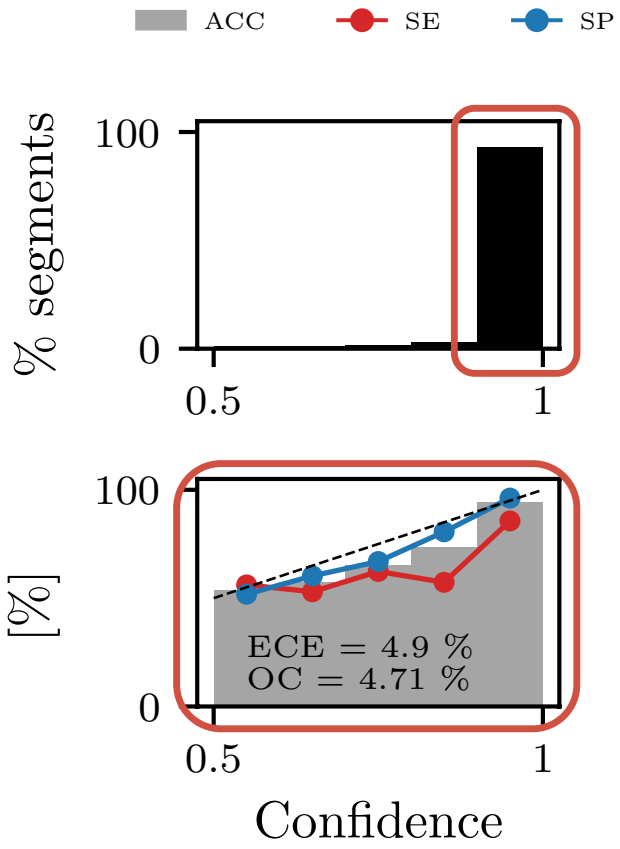
average confidence of segments in bin k

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

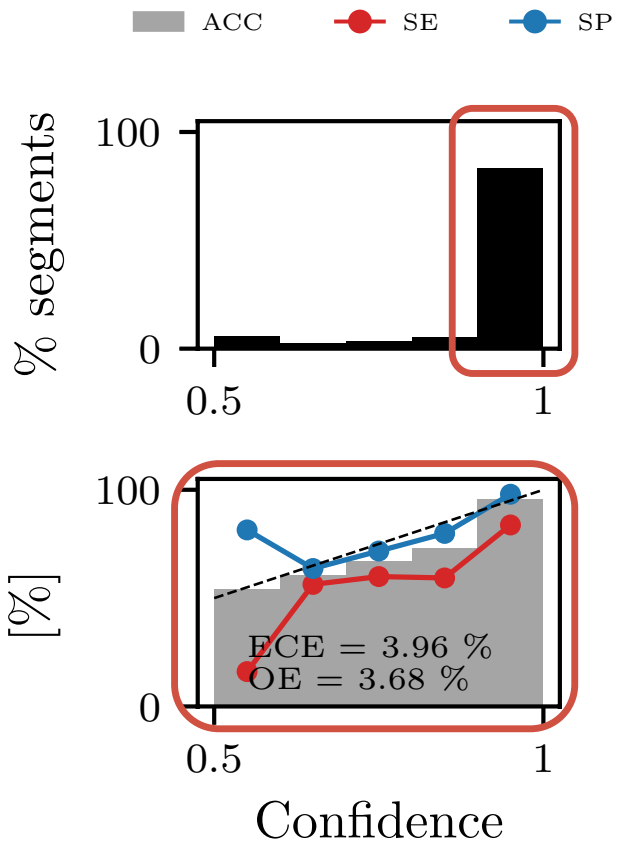
Guo et al., On calibration of modern neural networks, 2017; Thulasidasan et al., On mixup training: Improved calibration and predictive uncertainty for deep neural networks, 2019

Overconfident Predictions

Neonatal SDA



Adult SDA

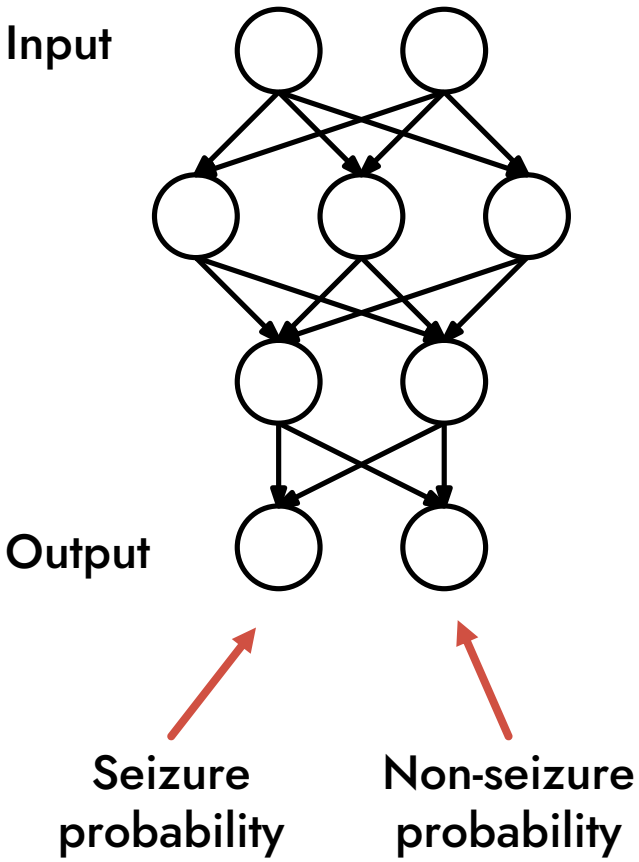


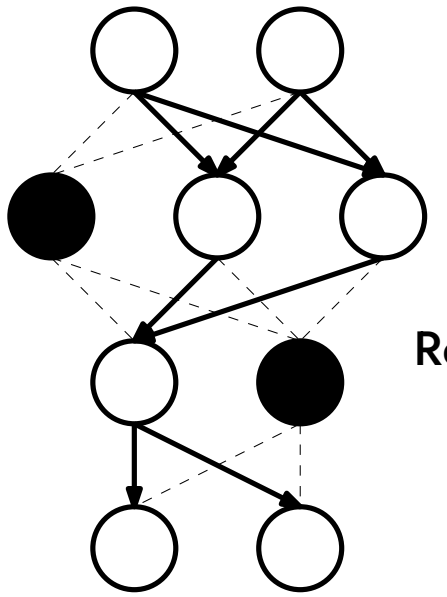
(N)SDA is confident in almost all predictions.

(N)SDA is overconfident in the predictions.

Guo et al., On calibration of modern neural networks, 2017

Dropout





Remove node with probability p

Dropout used during training and prediction.

Gal et al., Dropout as a bayesian approximation: Representing model uncertainty in deep learning, 2016

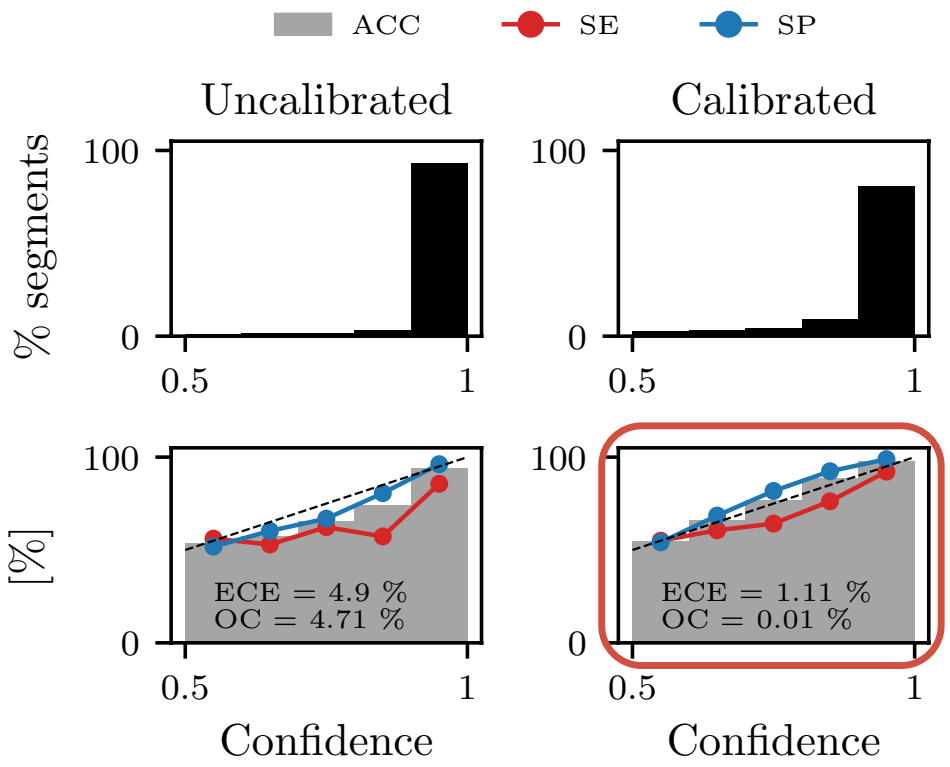
Results

	Neonatal SDA			Adult SDA		
	AUC	SE [%]	SP [%]	AUC	SE [%]	SP [%]
Uncalibrated	0.90	76	94	0.90	66	96
Calibrated	0.93	78	95	0.89	70	94

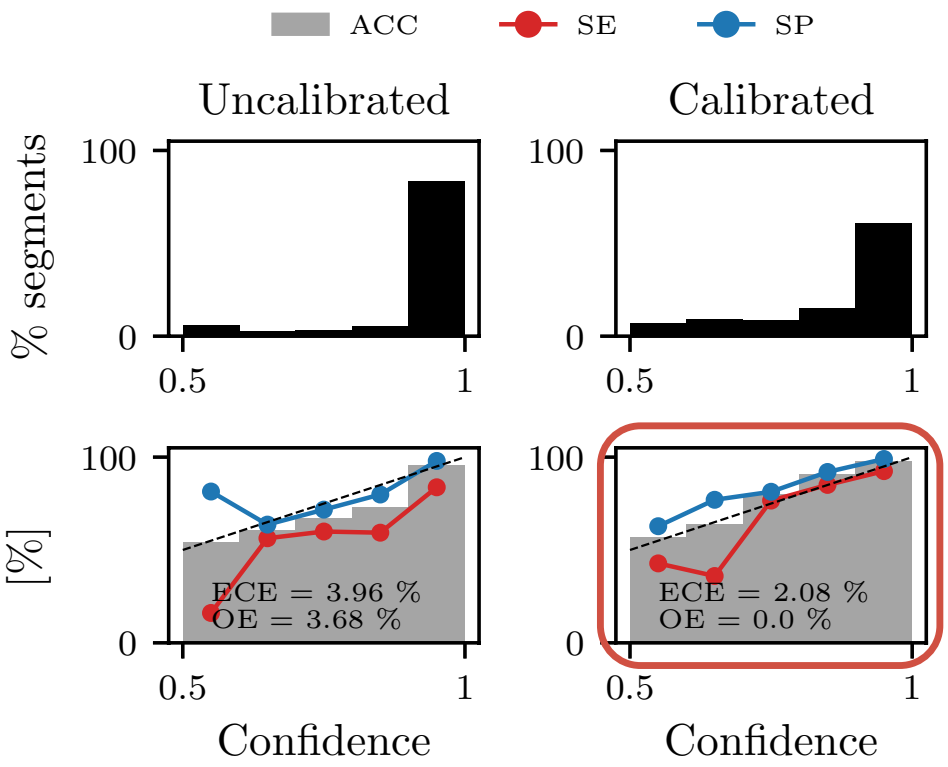
There are no big differences between performance metrics.

Results

Neonatal SDA



Adult SDA



Calibration is noticeably improved.

Can the SDA inform the user when it is not confident in its prediction?

Straightforward application of dropout during training and prediction leads to a noticeable improvement in the calibration of (N)SDA.