Influence of human-expert labels on a neonatal seizure detector based on a convolutional neural network

Ana Borovac[*]
Faculty of Industrial Engineering, Mechanical Engineering and Computer Science
University of Iceland
Reykjavik, Iceland
anb48@hi.is

Steinn Guðmundsson
Faculty of Industrial Engineering, Mechanical Engineering and Computer Science
University of Iceland
Reykjavik, Iceland
steinng@hi.is

Gardar Thorvardsson
Kvikna Medical ehf.
Reykjavik, Iceland
gardar@kvikna.com

Thomas Philip Runarsson
Faculty of Industrial Engineering, Mechanical Engineering and Computer Science
University of Iceland
Reykjavik, Iceland
tpr@hi.is

Abstract

Neonatal seizures are common among infants and can be detected with an electroencephalogram (EEG). The EEG signals are complex time-series using multiple channels. Human domain experts are often in disagreement when labelling neonatal seizure data. Only few studies will include labels from multiple experts, as annotating hours of EEG recordings is time consuming and expensive. In this study we investigated the differences in performance of a deep-learning-based neonatal seizure detector trained using single expert labelling versus data labelled using the consensus of multiple experts. Results indicate that there are differences even when the experts are in minor disagreement. We find that excluding ambiguously labeled data is important when training a neonatal seizure detector.

1 Introduction

Seizures are common among infants, with a prevalence of 1 – 5 per thousand live births [4]. Since untreated seizures can cause brain damage [1], it is paramount to detect them early. Seizure detection in infants is complicated by the fact that the majority of seizures cannot be observed clinically [2]. The

[*] Kvikna Medical ehf., Reykjavik, Iceland

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

current gold standard for neonatal seizure detection (NSD) is a multi-channel electroencephalogram (EEG) recording with simultaneous video, analyzed by a human expert [14]. The frequency and duration of seizures within an EEG are of clinical interest.

EEGs are time-series that represent the electrical activity of the brain. Neonatal EEG recordings are usually obtained with 4 – 20 electrodes that are placed on the scalp and last from a few hours to days. Analysis of an EEG requires extensive training and is time consuming which hampers widespread use. Automating the procedure is therefore of obvious clinical significance. The measurements have high inter- and intra-patient variability, the EEG is highly dependant on the age of the neonate, its condition [7, 8] and medication [6, 12]. Non-cerebral artifacts such as heartbeat, breathing and infant care frequently contaminate the signal and may mimic seizure activity. Due to the complexity of neonatal EEG signals, human experts are often in disagreement [11], in particular when seizures are short in duration [15].

Even though human experts provide the gold standard neonatal seizure labels, label noise is likely to be present in the training data which can have a negative effect on the performance of a machine learning model [18]. To the best of our knowledge there are only a few studies in the field of neonatal seizure detection addressing label noise by utilizing multiple human-expert labels [11, 13, 15, 17]. In this work we compare five strategies for utilizing labels from multiple human experts in the training of a NSD based on a deep convolutional neural network.

2 Methods

The data set used in the experiments contains segments from 79 neonatal EEG recordings, each approximately 1 hour in length, and accompanying labels from three human experts with 1 sec resolution [16]. The recordings contain 19 channels sampled at 256 Hz that were combined in a longitudinal montage (a frequently used pairwise combination of channels). The segments were split into 16 sec long blocks with 12 sec overlap. The signals were filtered with a 6th order Chebyshev Type 2 band-pass filter with cut-off frequencies of 0.5 Hz and 32 Hz, down-sampled to 32 Hz and standardised so that the mean and standard deviation were zero and one, respectively. Each 16 sec interval was labeled as a seizure or a non-seizure interval per human expert (A, B or C), the majority vote and consensus amongst experts, resulting in five sets of labelings. Ambiguous segments, i.e. segments that were partly labeled as seizure and partly as non-seizure, were excluded. Figure 1 illustrates scoring for a typical EEG segment and the total number of seizure/non-seizure segments is given in table 1. Non-seizure segments were approximately 8 times as many as the seizure segments. The non-seizure segments were therefore randomly sub-sampled to obtain balanced training sets. One network (NSD) was trained for each of the five labelings in table 1.

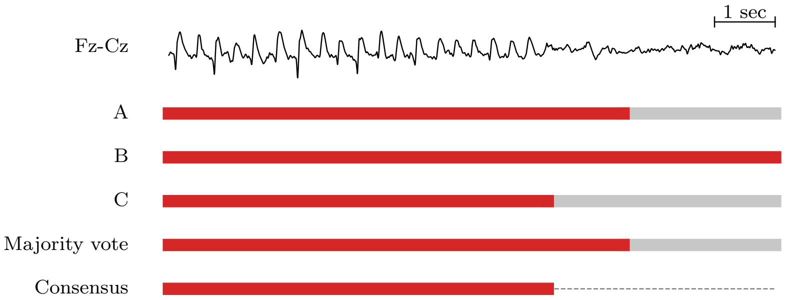

Figure 1: 10 sec EEG segment (channel Fz-Cz), labeling from scorers A, B and C, majority vote and consensus labels. Seizure areas are annotated with red, non-seizure with grey and ignored parts with dashed grey line.

A convolutional neural network proposed by Stevenson et al. [17] was used as a feature extractor. It consists of 10 convolutional layers with 32 filters of size 3 and one convolutional layer with 2 filters of size 3. Each convolutional layer is followed by a batch normalization layer and ReLU activation. Before the fourth, seventh and tenth convolutional layers, average pooling is applied with filters of size 8, 4 and 2 respectively. The stride was set to 3 for all three pooling layers. The feature extractor

Table 1: The total number of seizure and non-seizure segments available for each labeling; human experts (A, B and C), majority vote and consensus labels. The number of seizure and non-seizure segments exclusive to each expert are in parentheses.

Labeling	Seizure	Non-seizure
A	10482 (332)	85075 (619)
B	14170 (2129)	81266 (401)
C	11127 (1043)	83511 (394)
Majority vote	11658	84847
Consensus	8560	78260

is followed by an attention layer [9] and a fully connected layer with two output neurons and softmax activation.

Cross entropy was used as a loss function and the model parameters were optimized using the Adam optimizer with a mini-batch size of 128. The learning rate was set to 0.001 in the beginning and halved every 10 epochs. The model was trained for 40 epochs. Experiments using 30 and 50 epochs gave similar results (data not shown). A fixed number of epochs was used during training due to the prohibitive computational cost of using leave-one-patient-out cross-validation for parameter tuning.

Each of the five models were tested on labelings from experts A, B and C to investigate whether a model trained on labels from a single expert, under- or over-performs models trained on labels from the other experts in any significant way. The models were also tested on the consensus labels. The models were evaluated by leaving one subject out at a time to avoid train-test set overlap. There are 38 patients with at least one 16 sec long consensus seizure segment in the data set [16] and the results report below are based on data from these 38 patients. Cohen's kappa (κ) was used as the performance metric instead of ROC AUC since the test set was highly imbalanced and clinical utility of a NSD does not necessarily follow from a high AUC value [9].

The code used in the experiments was written in Python using PyTorch 1.7.1 and executed on a NVIDIA GeForce GTX 1080 Ti GPU.

3 Results and discussion

The main results are presented in figure 2. The figure shows that all the models performed poorly (i.e. low kappa values) on a small subset of patients. The poor performance is partly caused by the relatively small training set and high inter-patient variability. Some of the recordings have very few seizure or non-seizure segments which means that the performance metric is very sensitive to predictions from these segments.

Experts often disagree on the exact start and end times of seizures. They disagree also on seizures that are shorter than 30 sec in duration [15]. The consensus set excludes these segments, resulting in seizure segments that are in a sense "clean". This appears to be beneficial since the model trained on the consensus labels performs best overall (figure 2). The mean kappa values are between 0.52 and 0.61 for the NSD trained with consensus data.

The NSD trained with labels from expert B performs worst, irrespective of the test set. Table 1 shows that this expert labeled 27 % - 35 % more segments as seizures than experts A and C. Some of these additional seizure segments are confusing the classifier, leading to an increased number of false seizure predictions. This led to higher sensitivity and lower specificity (table 2).

Training on labels from expert A resulted in a model that performed the best, out of the three models trained on labels from a single expert. Expert A annotated the least number of exclusive segments (table 1) and agreed with at least one of the other two experts for most parts of the EEG recordings.

4 Conclusion

The experiments show that NSD performance can depend strongly on the expert responsible for scoring the EEG, as the results for expert B clearly show. The results from expert B also show

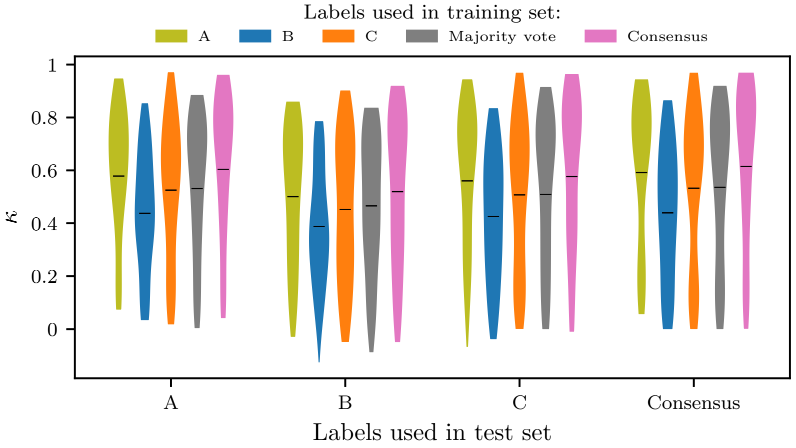

Figure 2: Comparison of Cohen's kappa (κ) values of models trained using different labels illustrated by the different colours. Results are compared with different test labels. Solid lines denote the mean values.

Table 2: Mean sensitivity and specificity values for different training/test labels.

	Test labels							
	Sensitivity [%]				Specificity [%]			
Training labels	A	B	C	Consensus	A	B	C	Consensus
A	76.77	67.55	75.93	80.51	89.96	91.22	90.90	92.28
B	79.12	71.41	77.38	80.74	81.93	82.73	82.87	83.85
C	75.94	66.85	73.69	78.30	88.04	88.76	88.92	90.08
Majority vote	78.68	70.19	76.36	80.80	86.91	87.96	87.89	89.16
Consensus	75.15	66.19	73.51	78.47	91.62	92.61	92.37	93.68

significant differences compared to the model using the majority vote in the training set. Improvement in classifier performance due to using majority vote of multiple domain experts has previously been observed in a study on prostate cancer classification [10].

When labels from multiple experts are available, using consensus labels can reduce label noise and improve the overall accuracy of the NSD. This is in agreement with previous findings on other types of data [18]. It further indicates that if the data labels are close to being noise-free, a clinically relevant NSD can be obtained even when the training set is relatively small. For comparison, kappa values calculated between the human experts over the entire data set were in the range 0.63 to 0.73.

Models trained on labels from a single expert did not result in models that captured the criteria the experts used to identify seizure segments. Explanations include the model architecture not capturing all the information an expert uses to determine the absense/presence of seizures. When scoring an EEG, experts frequently inspect segments that occur earlier or later in the recording. This behaviour is not captured by the convolutional network used here. Another explanation could be inattentional blindness [3]. However, there does not exist an absolute truth in EEG recordings, comparable to biopsies in skin cancer detection [5] and mistakes can not be easily confirmed.

To conclude, when using labels from one human expert it must be kept in mind that the labels are subjective to the expert and the performance of a model is highly dependent on the expert labelling the data. Therefore, when training a NSD it is important to reduce the label noise by excluding segments with ambiguous labels.

Acknowledgments and Disclosure of Funding

Authors thank Nathan J Stevenson for sharing the architecture of the neural network feature extractor and the pre-processing of data. This project receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 813483.

References

[1] Stella T Björkman, Stephanie M Miller, Stephen E Rose, Christopher Burke, and Paul B Colditz. Seizures are associated with brain injury severity in a neonatal model of hypoxia–ischemia. *Neuroscience*, 166(1): 157–167, 2010.

[2] Geraldine B Boylan, Nathan J Stevenson, and Sampsa Vanhatalo. Monitoring neonatal seizures. In *Seminars in Fetal and Neonatal Medicine*, volume 18, pages 202–208. Elsevier, 2013.

[3] Trafton Drew, Melissa L-H Võ, and Jeremy M Wolfe. The invisible gorilla strikes again: Sustained inattentional blindness in expert observers. *Psychological science*, 24(9):1848–1853, 2013.

[4] Hannah C Glass, Courtney J Wusthoff, Renée A Shellhaas, Tammy N Tsuchida, Sonia Lomeli Bonifacio, Malaika Cordeiro, Joseph Sullivan, Nicholas S Abend, and Taeun Chang. Risk factors for EEG seizures in neonates treated with hypothermia: a multicenter cohort study. *Neurology*, 82(14):1239–1244, 2014.

[5] Achim Hekler, Jakob N Kather, Eva Krieghoff-Henning, Jochen S Utikal, Friedegund Meier, Frank F Gellrich, Julius Upmeier zu Belzen, Lars French, Justin G Schlager, Kamran Ghoreschi, et al. Effects of label noise on deep learning-based skin cancer classification. *Frontiers in Medicine*, 7:177, 2020.

[6] Gregory L Holmes and Faye Korteling. Drug effects on the human EEG. *American Journal of EEG Technology*, 33(1):1–26, 1993.

[7] Richard A Hrachovy and Eli M Mizrahi. *Atlas of neonatal electroencephalography*. Springer Publishing Company, 2015.

[8] Aatif M Husain. Review of neonatal EEG. *American journal of electroneurodiagnostic technology*, 45(1): 12–35, 2005.

[9] Dmitry Yu Isaev, Dmitry Tchapyjnikov, C Michael Cotten, David Tanaka, Natalia Martinez, Martin Bertran, Guillermo Sapiro, and David Carlson. Attention-based network for weak labels in neonatal seizure detection. *Proceedings of machine learning research*, 126:479, 2020.

[10] Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020.

[11] Aileen Malone, C Anthony Ryan, Anthony Fitzgerald, Louise Burgoyne, Sean Connolly, and Geraldine B Boylan. Interobserver agreement in neonatal seizure identification. *Epilepsia*, 50(9):2097–2101, 2009.

[12] Rawad Obeid and Tammy N Tsuchida. Treatment effects on neonatal EEG. *Journal of Clinical Neurophysiology*, 33(5):376–381, 2016.

[13] 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.

[14] Ronit M Pressler, Maria Roberta Cilio, Eli M Mizrahi, Solomon L Moshé, Magda L Nunes, Perrine Plouin, Sampsa Vanhatalo, Elissa Yozawitz, Linda S de Vries, Kollencheri Puthenveettil Vinayan, et al. The ilae classification of seizures and the epilepsies: Modification for seizures in the neonate. position paper by the ilae task force on neonatal seizures. *Epilepsia*, 62(3):615–628, 2021.

[15] 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.

[16] Nathan J Stevenson, Karoliina Tapani, Leena Lauronen, and Sampsa Vanhatalo. A dataset of neonatal EEG recordings with seizure annotations. *Scientific data*, 6:190039, 2019.

[17] Nathan J Stevenson, Karoliina Tapani, and Sampsa Vanhatalo. Hybrid neonatal EEG seizure detection algorithms achieve the benchmark of visual interpretation of the human expert. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5991–5994. IEEE, 2019.

[18] Xingquan Zhu and Xindong Wu. Class noise vs. attribute noise: A quantitative study. *Artificial intelligence review*, 22(3):177–210, 2004.