Calibration of Automatic Seizure Detection Algorithms

A. Borovac[1,2], T. P. Runarsson[1], G. Thorvardsson[2] and S. Gudmundsson[1]

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

***Abstract*— An EEG seizure detection algorithm employed in a clinical setting is likely to encounter many EEG segments that are difficult to classify due to the complexity of EEG signals and small data sets frequently used to train seizure detectors. The detectors should therefore be able to notify the clinician when they are uncertain in their predictions and they should also be accurate for confident predictions. This would enable the clinician to focus mainly on the parts of the recording where confidence in predictions is low. Here we analyse the calibration of neonatal and adult seizure detection algorithms based on a convolutional neural network in terms of how well the output seizure/non-seizure probabilities estimate the corresponding empirical frequencies. We found that the detectors turned out to be overconfident, in particular when incorrectly predicting seizure segments as non-seizure segments. The calibration of both detectors, measured in terms of expected calibration error and overconfidence error, was improved noticeably with the use of Monte Carlo dropout. We find that a straightforward application of dropout during training and classification leads to a noticeable improvement in the calibration of EEG seizure detectors based on a convolutional neural network.**

Keywords*— *electroencephalogram, automatic seizure detection, uncertainty, calibration

I. Introduction

Seizures are common in the neonatal period [1], as well as in later stages of life [2]. Neonatal seizures should be detected and treated promptly as they often have an underlying brain injury [3]. In adulthood, the seizures may have a major impact on the quality of life and can be life threatening [4]. The current gold standard of seizure detection is a video electroencephalogram (EEG) observed by a human expert. Since EEG recordings frequently span hours to days, are prone to artefacts [5] and have high inter- and intra-patient variance [6, 7], scoring EEG recordings is time-consuming and requires special expertise that is not always available [8].

To speed up the analysis of EEG and make it more widely available, a significant effort has gone into the development of automated (neonatal) seizure detection algorithms (SDAs) [9, 10]. Designing and training SDAs with human-level performance is difficult for two main reasons. First, there is usually only a small amount of data available for training. Second, EEG signals are complex which makes seizure annotation difficult; even human experts with years of experience are often in disagreement [11, 12]. As a result, it may be expected that automatic classification would be difficult for some of the EEG segments. Algorithms that output confidence levels, in addition to seizure/non-seizure labels, are therefore desirable [13, 14]. EEG segments where confidence in prediction is low can then be passed on to the clinician for review. Furthermore, by directing the attention of the clinician to the parts of the recording where uncertainty is highest, manual scoring becomes more efficient.

Modern SDAs are based on deep neural networks (DNNs) [9, 10]. Output class probabilities may be interpreted as confidence estimates, where probabilities close to one would indicate high confidence and probabilities close to ½ would indicate low confidence in the seizure/non-seizure predictions. To the best of our knowledge, it has not been investigated how accurate such confidence estimates are in this setting. A classifier is considered to be well-*calibrated* if the confidence estimates are close to the empirical frequencies. In [15, 16] confidence estimates for a support vector machine classifier were obtained by a version of Platt scaling [17] and Becher et al. [16] additionally estimated confidence levels with trust scores [18].

Guo et al. [19] claim that DNNs are often poorly calibrated and overconfident in their predictions, despite achieving good classification performance. Hein et al. [20] show that DNNs employing the ReLU activation function can be overconfident in predictions for data far away from the training data. In recent years, various methods have been proposed for improving the calibration of DNNs [21, 22]. They include post-processing methods such as isotonic regression [23], conformal prediction [24] and Platt scaling [17], as well as methods that modify the training process such as mixup [25, 26], modelling probability distributions of class probabilities with Dirichlet distributions [27] and the use of dropout during training *and* prediction [28].

In this work, we analyse the calibration of an SDA based on a convolutional neural network and show that the detector is overconfident in its predictions, in particular for seizure segments. The calibration is improved noticeably, without degrading classification performance, by using a simple dropout strategy. The analysis is done on publicly available neonatal and adult EEG data sets.

II. Methods

Data

The neonatal EEG was taken from a data set with 79 recordings [29] and processed as described in [30]. Briefly, the recordings were recorded with 19 electrodes with a common reference and from these signals, a bipolar longitudinal (double banana) montage with 18 channels was derived. The same montage was used by the human experts annotating the recordings [29]. The recordings were cut into 16 sec long segments with 12 sec overlap [31]. We included only segments where all three human annotators were in agreement [30]. Each signal was filtered with a 6th order Chebyshev Type 2 filter with band-pass 0.5 – 16 Hz and down-sampled from 256 Hz to 32 Hz [31]. This frequency band was selected since the cortical activity of neonates normally lies in this range [32–34]. The signals were then normalised to mean zero and standard deviation one [35–37]. Approximately 10 % of the segments contain seizures.

The adult EEG was taken from the TUH EEG seizure corpus, version 1.5.4 [38]. The set of recordings with averaged reference was used together with a bipolar temporal central parasagittal montage with 22 channels. The same montage was used by the human experts annotating the recordings. The training, validation and test sets contain recordings from 297, 41 and 41 patients, respectively. The pre-processing was similar to the neonatal data. Labelled signals were included if the manual annotation had a confidence value of one. The signals were cut into 16 sec long segments. There is an overlap of 12 sec for the seizure segments to make the set of seizures larger. The signals were filtered with a 0.5 – 25 Hz band-pass filter [39], down-sampled to 50 Hz and normalised in the same way as before. The fraction of segments containing seizures in the three data sets is 12 %.

Seizure Detection Algorithm

The SDA from [40] was used for both data sets in a binary setting (seizure/non-seizure). The detector is based on a DNN that uses multi-channel EEG as input. The network extracts features from each channel separately with a convolutional neural network [41] and combines the feature vectors into a single feature vector with an attention layer [42]. This is followed by a fully connected layer with two output nodes and a softmax activation function which provides confidence estimates for the classification. Because of the difference in sampling rates, the input size differs between the neonatal and adult data sets, resulting in a different number of features extracted from each channel (24 for the neonatal EEG vs. 44 for the adult EEG). Consequently, the numbers of parameters in the attention and fully connected layers are different. The neonatal detector has 29352 learnable parameters while the adult detector has 29712.

The training of the detector followed [30]. The neonatal (adult) detector was trained for 30 (50) epochs with the Adam optimizer, with an initial learning rate of 0.001 which was then halved every 10 epochs. Mini-batches were of size 128. Since the data sets were highly imbalanced, each mini-batch was balanced, i.e. there were 64 seizure segments and 64 non-seizure segments in each mini-batch. Hence, each epoch contained all the available seizure segments and an equal number of randomly selected non-seizure segments.

Dropout

Dropout is a simple and widely used regularization technique for improving the generalisation of DNNs [43]. With dropout, nodes are omitted at random from the network with fixed probability p during training, together with their connections. This prevents hidden nodes in the network from relying too much on other hidden nodes to correct their mistakes, which in turn reduces overfitting. In the typical setting (*standard dropout*), dropout is only used during training in order to reduce the amount of computations in the testing phase. In *Monte Carlo dropout*, T forward passes are performed with dropout enabled in the prediction phase and the predictions are averaged. It has been observed empirically that this can give a slight improvement in prediction accuracy over simple dropout. Due to the extra computational cost, Monte Carlo dropout is infrequently used for this purpose but it has the additional benefit of providing probability estimates that are better calibrated than those obtained with standard dropout [44]. The connection between Monte Carlo dropout and model uncertainty is provided in [28] where Monte Carlo dropout is interpreted as approximate Bayesian inference in deep Gaussian processes. Dropout with probability $p = 0.1$ was used for all nodes in the convolutional and attention layers (the nodes in the input layer were excluded) but for the nodes in the fully connected layer $p = 0.5$ [28, 43]. The average of $T = 10$ softmax predictions was used to obtain final probability estimates (averaging over a larger number of predictions gave similar results).

Performance Evaluation

The classification performance of the SDAs was evaluated with the area under the curve (AUC), sensitivity (SE) and specificity (SP).

The confidence of a single prediction is defined as the highest softmax output of the detector. For binary classification tasks, the confidence values, therefore, lie between 0.5 and 1. The calibration was evaluated with

the expected calibration error [19],

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

and overconfidence error which gives high weight to confident but wrong predictions, a situation that is of particular concern in medical applications [25],

$$\mathrm{OE} = \sum_{k=1}^{K} \frac{|B_k|}{N} \mathrm{conf}(B_k) \cdot \max(\mathrm{conf}(B_k) - \mathrm{acc}(B_k), 0), \tag{2}$$

where the confidence values have been partitioned into K equally sized bins (here $K = 5$), B_k is the set of segments where the confidence level falls into bin k, $|B_k|$ is the number of segments in bin k, $\mathrm{acc}(B_k)$ is the portion of correctly classified segments in bin k, $\mathrm{conf}(B_k)$ is the average confidence of segments in bin k and N is the total number of segments.

Leave-one-subject-out cross-validation was used for the evaluation of the neonatal SDA. The adult SDA was evaluated on a separate test set.

III. Results and Discussion

Detectors employing Monte Carlo dropout are referred to as *calibrated* in the following and they are compared to detectors that were trained without using any dropout during training and prediction (*not calibrated*).

Table 1 shows the performance of the neonatal and adult SDAs on the two data sets, averaged across patients with seizures. Inter-patient variability is quite high, in particular for the sensitivity metric. The variability of the specificity metric is lower for the adult SDA. This indicates that there are some patients that are difficult to classify in both data sets and for those, it would be preferred to obtain uncertain predictions rather than certain incorrect predictions. Such property of a detector may in the future also increase the trust of the clinicians using the system [13, 14].

Table 1. Mean area under the curve (AUC), sensitivity (SE) and specificity (SP) across the patients with at least one seizure segment. The standard deviations are shown in parentheses.

Neonatal SDA	**AUC**	**SE [%]**	**SP [%]**
Uncalibrated	0.90 (0.15)	76.02 (29.91)	93.80 (14.54)
Calibrated	0.93 (0.11)	78.39 (28.65)	95.24 (10.09)
Adult SDA			
Unalibrated	0.90 (0.14)	66.47 (32.15)	95.70 (4.85)
Calibrated	0.89 (0.16)	70.27 (32.29)	93.63 (7.05)

While the SDA architecture was designed for neonatal EEG it nevertheless gives fairly good results on the adult data set. For comparison, the best DNN architecture (out of 15 tested) reported in [45] has an AUC of 0.92, sensitivity 83 % and specificity 85 % on the TUH data set, with the caveat that [45] used a slightly different testing procedure. Detectors with high specificity (e.g., above 90 %) are often preferred in the online clinical setting to avoid frequent disruption due to false detections.

Table 1 shows that the classification performance (in terms of AUC, sensitivity and specificity) of the calibrated neonatal SDA is marginally better than for the uncalibrated detector, while the adult uncalibrated and calibrated SDAs perform similarly. This is in line with previous studies which applied Monte Carlo dropout for the classification of (medical) images [46–48].

Even though the performance of the uncalibrated and calibrated SDAs are similar in terms of the average AUC, sensitivity and specificity metrics, they can differ considerably in predictions on individual recordings. Figure 1 shows predictions for a single neonatal recording. Since the prediction confidence corresponds to the highest softmax output, the confident seizure predictions have a seizure probability close to one and confident non-seizure predictions have a seizure probability close to zero. The detector without calibration is confident in false seizure predictions for a big portion of the recording, but the predictions of the calibrated SDA which have high seizure probability are in agreement with the three human annotators that labelled the data set. Three out of seven consensus seizures are clearly detected and two additional seizures can quickly be identified by inspecting the areas with high uncertainty (figure 1), i.e. with seizure probability around 0.5.

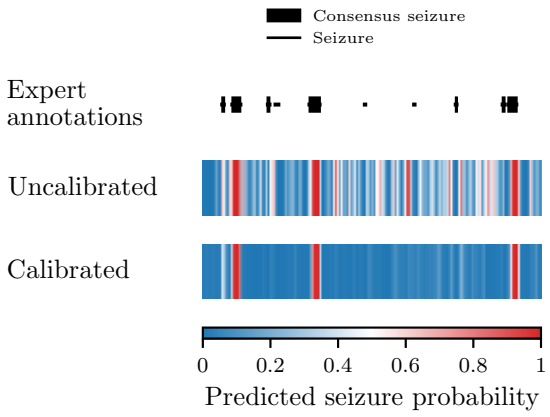
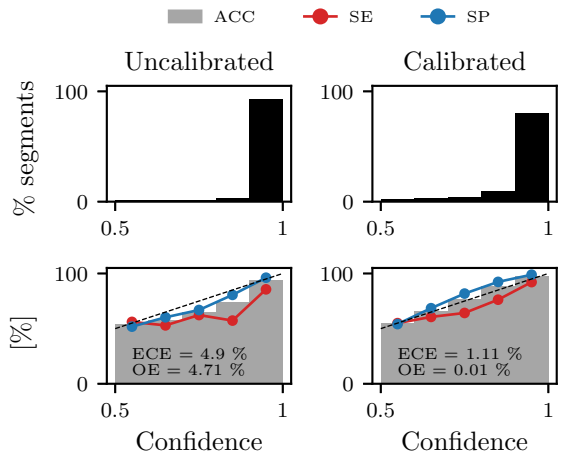
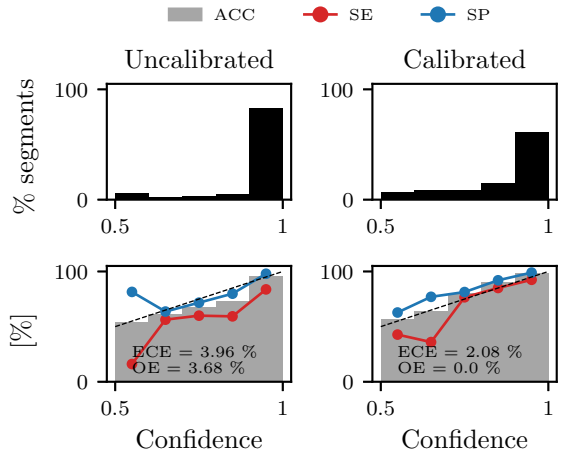


Figure 1. An example of predictions made by the uncalibrated and calibrated neonatal SDAs for a 56 min long neonatal recording. The recording contains seven seizures where all three human experts were in agreement and three additional seizures were labelled by at least one of the experts.

The calibration of the SDAs is further analyzed in figure 2. The uncalibrated neonatal (adult) detector predicts about 93 % (83 %) of the examples with confidence close to one. The reliability diagrams for this



(a) Neonatal SDA



(b) Adult SDA

Figure 2. Neonatal (a) and adult (b) SDAs without calibration (left) and with calibration (right). Confidence histograms (black) show the fraction of predictions with a given confidence value and reliability diagrams (grey) show the expected accuracy as a function of confidence value. Deviations from the dashed lines represent miscalibration. Accuracy (ACC), sensitivity (SE), specificity (SP), expected calibration error (ECE) and overconfidence error (OE).

case show that the confidence levels do not reflect the true accuracy, as indicated by deviation from the dashed lines. The deviation is clearly lower when calibration is applied and this is also reflected in the expected calibration error. In addition, the uncalibrated detectors are overconfident in their predictions, seizure predictions in particular, which results in a high overconfidence error. The error drops to 0.01 % and 0.0 % for the calibrated neonatal and adult detectors, respectively. Low overconfidence error would consequently allow the user to trust predictions with a high (e.g., > 0.9) confidence level. In other words, highly confident predictions are almost always correct and the risk of false detection or a missed detection is low.

Figure 3 illustrates the relationship between classifier performance and mean confidence levels for individual patients in the adult data set. The calibrated SDA

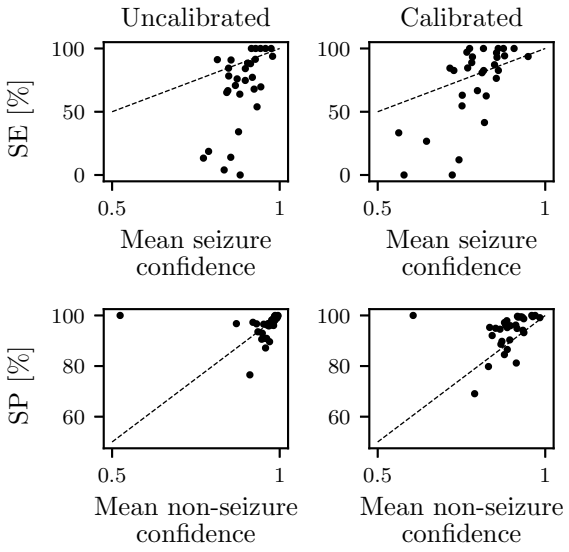


Figure 3. Each dot represents a patient from the adult data set. Mean seizure (non-seizure) confidence is the average confidence level of segments predicted as seizures (non-seizures). These two values are estimates for sensitivity (SE) and specificity (SP).

estimates for sensitivity and specificity are closer to the true values (dashed lines) and more importantly, the confidence estimates for the difficult examples are much lower than for the patients on which the SDA performs almost perfectly. Similar observations are made also on the neonatal data (data not shown).

IV. Conclusion

In this work, we have shown that an SDA based on a DNN architecture optimised for neonatal seizure detection, can be retrained on adult EEG data to provide a reasonably accurate classifier for adult EEG. However, despite good classification performance, neonatal and adult detectors were overconfident in the predictions which may reduce user trust [13, 14]. Our results demonstrate that dropout [28] improves calibration, in particular for the seizure segments. A well-calibrated detector can notify the user when it is not confident in its predictions and leave the decision to the user. This allows the user to focus quickly on the parts of the recording where the automatic detection is uncertain.

As suggested in [46, 47] Monte Carlo dropout may not perform well in case of a distribution shift. In our case, the shift can be attributed to the different age groups, recording equipment and protocols. Therefore, further research is needed to investigate the influence of a distribution shift on the calibration of an SDA.

Dropout may also be combined with mixup training [25] and post-processing schemes such as Platt scaling [17] in the future. More work is also needed to find out how to present the output of calibrated detectors in intuitive and informative ways in the clinical setting.

ACKNOWLEDGEMENTS

This project receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 813483.

REFERENCES

[1] F. Pisani, C. Facini, E. Bianchi, G. Giussani, B. Piccolo, and E. Beghi, "Incidence of neonatal seizures, perinatal risk factors for epilepsy and mortality after neonatal seizures in the province of Parma, Italy," *Epilepsia*, vol. 59, no. 9, pp. 1764–1773, 2018.

[2] R. Kobau, F. Gilliam, and D. J. Thurman, "Prevalence of self-reported epilepsy or seizure disorder and its associations with self-reported depression and anxiety: results from the 2004 Healthstyles Survey," *Epilepsia*, vol. 47, no. 11, pp. 1915–1921, 2006.

[3] C. Vasudevan and M. Levene, "Epidemiology and aetiology of neonatal seizures," *Seminars in Fetal and Neonatal Medicine*, vol. 18, no. 4. Elsevier, 2013, pp. 185–191.

[4] B. Litt and J. Echauz, "Prediction of epileptic seizures," *The Lancet Neurology*, vol. 1, no. 1, pp. 22–30, 2002.

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

[6] R. A. Hrachovy and E. M. Mizrahi, *Atlas of neonatal electroencephalography*. Springer Publishing Company, 2015.

[7] S. Noachtar and J. Rémi, "The role of EEG in epilepsy: a critical review," *Epilepsy & Behavior*, vol. 15, no. 1, pp. 22–33, 2009.

[8] G. Boylan, L. Burgoyne, C. Moore, B. O'Flaherty, and J. Rennie, "An international survey of EEG use in the neonatal intensive care unit," *Acta paediatrica*, vol. 99, no. 8, pp. 1150–1155, 2010.

[9] B. Olmi, L. Frassineti, A. Lanata, and C. Manfredi, "Automatic Detection of Epileptic Seizures in Neonatal Intensive Care Units Through EEG, ECG and Video Recordings: A Survey," *IEEE Access*, vol. 9, pp. 138 174–138 191, 2021.

[10] S. Saminu, G. Xu, Z. Shuai, I. Abd El Kader, A. H. Jabire, Y. K. Ahmed, I. A. Karaye, and I. S. Ahmad, "A recent investigation on detection and classification of epileptic seizure techniques using EEG signal," *Brain Sciences*, vol. 11, no. 5, p. 668, 2021.

[11] A. Dereymaeker, A. H. Ansari, K. Jansen, P. J. Cherian, J. Vervisch, P. Govaert, L. De Wispelaere, C. Dielman, V. Matic, A. C. Dorado *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*, vol. 128, no. 9, pp. 1737–1745, 2017.

[12] J. Halford, D. Shiau, J. Desrochers, B. Kolls, B. Dean, C. Waters, N. Azar, K. Haas, E. Kutluay, G. Martz *et al.*, "Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG recordings," *Clinical Neurophysiology*, vol. 126, no. 9, pp. 1661–1669, 2015.

[13] E. Begoli, T. Bhattacharya, and D. Kusnezov, "The need for uncertainty quantification in machine-assisted medical decision making," *Nature Machine Intelligence*, vol. 1, no. 1, pp. 20–23, 2019.

[14] B. Kompa, J. Snoek, and A. L. Beam, "Second opinion needed: communicating uncertainty in medical machine learning," *NPJ Digital Medicine*, vol. 4, no. 1, pp. 1–6, 2021.

[15] A. Temko, E. Thomas, W. Marnane, G. Lightbody, and G. Boylan, "EEG-based neonatal seizure detection with support vector machines," *Clinical Neurophysiology*, vol. 122, no. 3, pp. 464–473, 2011.

[16] T. Becker, K. Vandecasteele, C. Chatzichristos, W. Van Paesschen, D. Valkenborg, S. Van Huffel, and M. De Vos, "Classification with a deferral option and low-trust filtering for automated seizure detection," *Sensors*, vol. 21, no. 4, p. 1046, 2021.

[17] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999.

[18] H. Jiang, B. Kim, M. Guan, and M. Gupta, "To trust or not to trust a classifier," *Advances in neural information processing systems*, vol. 31, 2018.

[19] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," *International conference on machine learning*. PMLR, 2017, pp. 1321–1330.

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

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

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

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

[24] G. Shafer and V. Vovk, "A Tutorial on Conformal Prediction." *Journal of Machine Learning Research*, vol. 9, no. 3, 2008.

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

[26] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[27] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in neural information processing systems*, vol. 31, 2018.

[28] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.

[29] N. J. Stevenson, K. Tapani, L. Lauronen, and S. Vanhatalo, "A dataset of neonatal EEG recordings with seizure annotations," *Scientific data*, vol. 6, p. 190039, 2019.

[30] A. Borovac, S. Guðmundsson, G. Thorvardsson, and T. P. Runarsson, "Influence of human-expert labels on a neonatal

seizure detector based on a convolutional neural network," *The NeurIPS 2021 Data-Centric AI Workshop*, 2021.

[31] N. J. Stevenson, K. Tapani, and S. Vanhatalo, "Hybrid neonatal EEG seizure detection algorithms achieve the benchmark of visual interpretation of the human expert," *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2019, pp. 5991–5994.

[32] J. J. Alix, A. Ponnusamy, E. Pilling, and A. R. Hart, "An introduction to neonatal EEG," *Paediatrics and Child Health*, vol. 27, no. 3, pp. 135–142, 2017.

[33] M. Kitayama, H. Otsubo, S. Parvez, A. Lodha, E. Ying, B. Parvez, R. Ishii, Y. Mizuno-Matsumoto, R. A. Zoroofi, and O. C. Snead III, "Wavelet analysis for neonatal electroencephalographic seizures," *Pediatric neurology*, vol. 29, no. 4, pp. 326–333, 2003.

[34] A. M. Husain, "Review of neonatal EEG," *American journal of electroneurodiagnostic technology*, vol. 45, no. 1, pp. 12–35, 2005.

[35] G. Xu, T. Ren, Y. Chen, and W. Che, "A One-Dimensional CNN-LSTM Model for Epileptic Seizure Recognition Using EEG Signal Analysis," *Frontiers in Neuroscience*, vol. 14, p. 578126, 2020.

[36] H. Mukhtar, S. M. Qaisar, and A. Zaguia, "Deep convolutional neural network regularization for alcoholism detection using EEG signals," *Sensors*, vol. 21, no. 16, p. 5456, 2021.

[37] A. Shoeibi, D. Sadeghi, P. Moridian, N. Ghassemi, J. Heras, R. Alizadehsani, A. Khadem, Y. Kong, S. Nahavandi, Y.-D. Zhang *et al.*, "Automatic diagnosis of schizophrenia in EEG signals using CNN-LSTM models," *Frontiers in Neuroinformatics*, vol. 15, 2021.

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

[39] J. Gotman, J. Ives, and P. Gloor, "Frequency content of EEG and EMG at seizure onset: possibility of removal of EMG artefact by digital filtering," *Electroencephalography and clinical neurophysiology*, vol. 52, no. 6, pp. 626–639, 1981.

[40] A. Borovac, S. Gudmundsson, G. Thorvardsson, S. M. Moghadam, P. Nevalainen, N. Stevenson, S. Vanhatalo, and T. P. Runarsson, "Ensemble learning using individual neonatal data for seizure detection," *arXiv preprint arXiv:2204.07043*, 2022.

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

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

[43] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[44] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, "Leveraging uncertainty information from deep neural networks for disease detection," *Scientific reports*, vol. 7, no. 1, pp. 1–14, 2017.

[45] K. Lee, H. Jeong, S. Kim, D. Yang, H.-C. Kang, and E. Choi, "Real-Time Seizure Detection using EEG: A Comprehensive Comparison of Recent Approaches under a Realistic Setting," *arXiv preprint arXiv:2201.08780*, 2022.

[46] R. Krishnan and O. Tickoo, "Improving model calibration with accuracy versus uncertainty optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 18 237–18 248, 2020.

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

[48] Z. Zhang, A. V. Dalca, and M. R. Sabuncu, "Confidence calibration for convolutional neural networks using structured dropout," *arXiv preprint arXiv:1906.09551*, 2019.