# Bridging the source-target mismatch with pseudo labeling for neonatal seizure detection

Aengus Daly
*Department of Mathematics, Munster Technological University - MTU*
*Department of Electrical and Electronic Engineering & INFANT, University College Cork*
Cork, Ireland
Aengus.Daly@mtu.ie

Gordon Lightbody
*Department of Electrical and Electronic Engineering & INFANT, University College Cork*
Cork, Ireland
ORCID: 0000-0002-8155-0826

Andriy Temko
*Department of Electrical and Electronic Engineering, University College Cork*
Cork, Ireland
ORCID: 0000-0001-6548-0971

***Abstract*—The mismatch between training and testing conditions is a known problem in the machine learning community. In this work, we outline a process of how a model which was trained under one set of conditions can be adapted to a new set of conditions by means of pseudo labeling. This is shown for the domain area of neonatal seizure detection. A previously developed deep learning architecture is first trained on a publicly available source dataset. It is then evaluated on another target dataset which was recorded in a different center, with different equipment, and annotated by a different expert. This model is then used to create pseudo labels on a sample of the target dataset, fine-tuned with the created pseudo labels, and re-evaluated on the target dataset. The results show a relative improvement of 13.5% and 28.8% in AUC and the number of seizures detected respectively. Various factors of the pseudo labeling procedure such as the amount of data vs confidence in pseudo labels are analyzed and presented.**

***Keywords—pseudo labeling, training and testing conditions mismatch, EEG, neonatal seizure detection, deep learning***

## I. Introduction

Deep learning models are known to be data hungry. Leveraging all available data while assuring data quality is part of a usual recipe to improve the accuracy in many domain areas [1], [2], [3]. In the context of machine learning challenges, even the right usage of unlabeled data can give a competitive advantage. In numerous recent Kaggle competitions pseudo labeling is a key component of the winning solutions, in particular computer vision [4] and NLP competitions [5].

In pseudo labeling, a model is first constructed by training it on the available labeled data. This teacher model is then utilized to create pseudo labels for the unlabeled database. Pseudo labeling can be used for knowledge distillation where a single powerful teacher model (or more often an ensemble of models) create soft labels (probabilistic output) from which a simpler student model can learn to perform the task at a smaller computational cost while preserving a similar level of performance [6].

Soft labels, i.e. probabilities produced by the pseudo labeling process, can also be hardened (converted to 0-1 for binary classification) to gain extra labeled data which can then be used to fine-tune the original model. Where the teacher and student model are the same, as is the case here this label hardening process creates extra information from which the model can learn. It is worth noting that without label hardening, i.e. by utilizing only soft labels (probabilistic output) the model gains no new information and will not improve. Usually, confidence considerations are incorporated in the process so that only datapoints that are more likely to be correct are used.

Pseudo labeling can also help with the issue of performance reduction due to source-target mismatched conditions. It has been shown that model performance often degrades when it is trained on a source database from one center/location/domain and tested on a target database from a different center/location/domain. In the domain area of biomedical signal processing such as EEG classification this performance reduction is primarily due to source-target differences in recording equipment, local operating procedures, and annotation quality/subjectivity [7]. This source of performance reduction is also seen in other domain areas such as medical imaging applications [8], cross language models in NLP [9] and microbiology [10]. This limits the usefulness of off-the-shelf available models and hinders the use of large annotated databases in many inter-related domain areas.

In this work, we show how the performance of a model trained on a publicly available dataset can be improved with a sample of target data without annotations. Here a pseudo labeling technique is used for the area of neonatal seizure detection, to attenuate the mismatch between source (training) database, and the target (testing) database that come from different centers.

This proposed research answers the following questions:

- Can pseudo labeling help with the source-target mismatch for the problem of neonatal seizure detection?
- What are the main parameters to consider in pseudo labeling?
- What are the means to improve the quality of pseudo labels?

## II. Materials and Methods

### A. Pseudo labeling technique using hard labels

Pseudo labeling works by using a teacher model trained on an annotated source database to label an unannotated database. Variants of pseudo labeling techniques include loss function adjustments [1], graph-based similarity [3], knowledge distillation[6], and iterative curriculum labeling [11]. In this research, a teacher model to detect neonatal

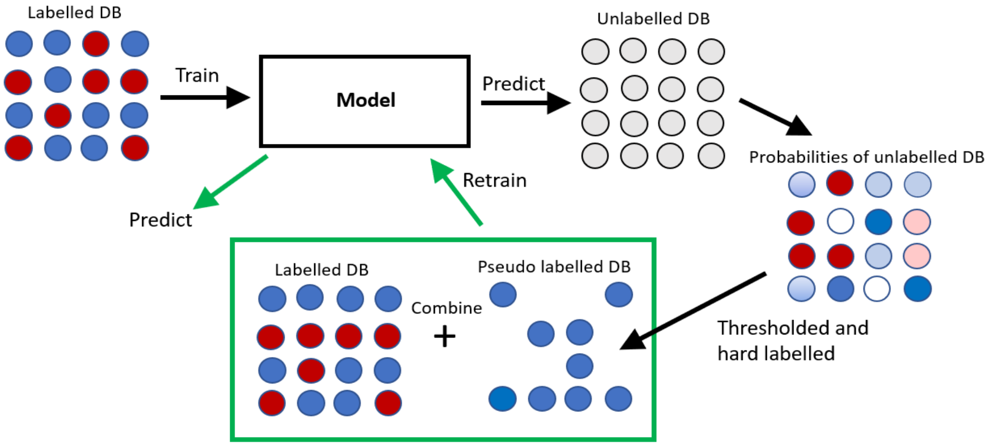

Figure 1. Pseudo labeling technique. A model, trained using the source labeled database (DB), is used to create probabilities for the sample unlabeled DB. The probabilities are thresholded and hard labeled for the non-seizure class to create a pseudo labeled DB that is combined with the source labeled DB to make a new training DB for the model. The model used here is the same for the teacher and the student model.

seizures is created using a publicly available dataset (source conditions) as training data.

This teacher model is evaluated on a large internal dataset (target conditions) of continuous EEG to establish the baseline performance. The teacher model is then used to pseudo label a sample dataset from target conditions which is also publicly available. The source and target conditions are very different (source-target mismatch) as the datasets come from different centers, recorded using different equipment, annotated by different experts etc. Additionally, the source dataset is composed of many short (1h) segments whereas in contrast the target conditions are represented by real-life monitoring scenario of continuous (over 24h) EEG recordings.

The teacher model produces soft pseudo labels for the unlabeled dataset, i.e. probabilistic outputs for seizure/non-seizure. A threshold is applied to these probabilities, in order to divide them for a given class into two categories – high confidence and low confidence decisions. Here the high confidence probabilities for the non-seizure class are converted to hard labels-0, and thus a new labeled (non-seizure) dataset is formed. This new labeled dataset is added to the source conditions dataset to form a bigger training dataset. Retraining with only non-seizure pseudo labeled data was preferred here since the sample dataset does not contain any confident representation of the seizure class of the target condition. The student model is retrained or fine-tuned on this bigger dataset.

It is important to perform hard labeling if the teacher and student models are the same as is the case in this study. Without hard labeling, retraining will have no effect since the output of the model will be equal to the soft labels this same model has produced earlier. The error (the difference between the output and the soft labels) on these new datapoints will be then zero and no learning will occur. Hard labeling is the only process that creates new information for the system from which the student model can learn. It is worth noting that the process can be different when the teacher and student models are not the same, e.g. when a teacher model is represented by an ensemble of powerful heavy models whereas a student model is relatively small. In this case soft labels can be used for the student model and the process, well known as knowledge distillation [6], aims also to reduce computational complexity of the inference by using the smaller student model while preserving the performance of the ensemble.

The effect of the confidence threshold on probabilistic outputs is studied in this work using the performance obtained on the large target database. Fig. 1 shows the schematic of the pseudo labeling process.

### B. *Deep learning model for seizure detection*

Seizures in neonates are seen in EEG signal data by a repetitive pattern that can evolve in amplitude and/or in frequency over time, see Fig. 2 where examples of a seizure and non-seizure are given.

A deep learning model for detection of neonatal seizure is designed and used here for both the teacher and student model. It is trained on a 16s sliding windows of raw multichannel EEG signal with a 1s shift, with no hand-crafted feature engineering used.

A high and low pass filter of 0.5Hz and 12.8Hz respectively are used to pre-process the EEG signal,

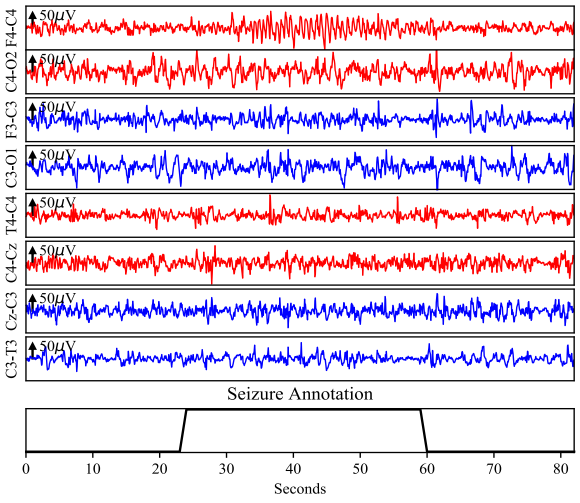

Figure 2. A segment of 80s of multichannel neonatal EEG with a seizure annotation. Source: Adapted from [12]

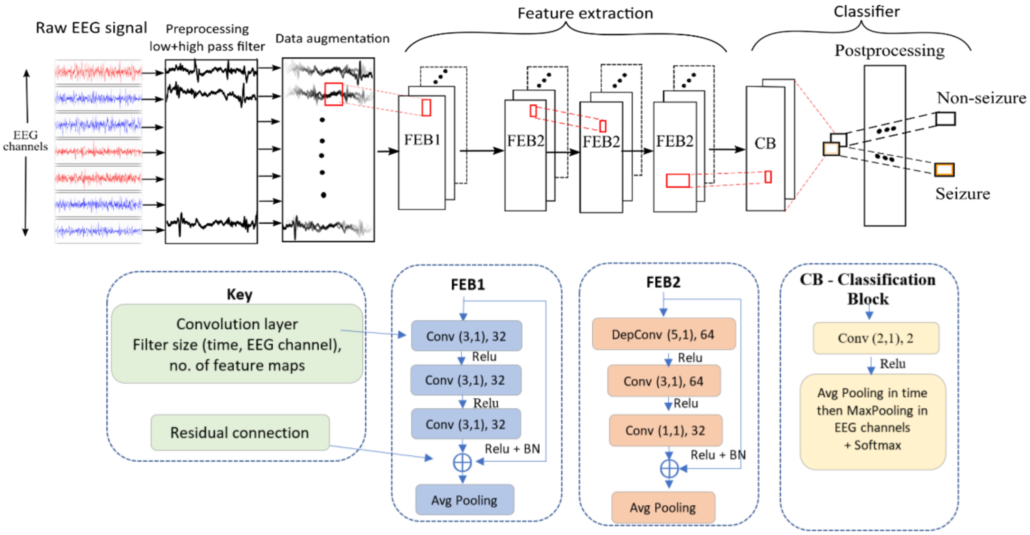

Figure 3. Model architectural details with feature extraction blocks (FEB) and a classification block (CB).

removing noise and unwanted artefacts followed by downsampling to 32Hz. The information content of neonatal seizures is known to be mostly concentrated in the range between 0.5 and 12Hz [13].

This model uses a fully convolutional architecture that allows it to run on any number of EEG channels and any length of the input. In addition, the model does not require strong labels (per channel annotations) and can be trained with weak labels (overall across-channel annotations).

This model training uses the latest architectural advances in training convolutional models [14] such as residual connections [15], data augmentation (e.g. Mixup and random adjustment of the amplitude of the EEG signal), depth-wise convolutions, and bottleneck layers. The total no. of trainable parameters is 45k.

The model consists of four feature extraction blocks (FEB) followed by a classification block. The first FEB has three convolutional layers with a filter of small sample size (3,1), followed by batch normalization and average pooling that downsamples the feature maps.

The subsequent FEBs are three similar blocks that consist of a depth-wise convolutional first layer that extracts further information at the feature maps' level using a larger filter of size (5,1). This is followed by two convolutional layers one with a filter of (3,1) followed by one with a filter of (1,1); these two layers also have an inverted bottle neck design that takes as input an expanded 64 feature maps and then condenses it to 32. Similar to the first FEB this is followed by batch normalization and average pooling.

The effect of the design of these three similar FEBs that use depth-wise convolutions, the convolutions with the varying kernel sizes, and the inverted bottleneck is to extract a variety of representations at different layers of the network thus boosting its ability to extract detailed knowledge of the difference between seizure and non-seizure from the inputted EEG signal.

The fifth block of the network is a classification block that funnels the networks representations into seizure and non-seizure paths using average pooling in time followed by max pooling in EEG channel.

Residual connections [15] at each FEB are used that enable the network to train efficiently and utilize all the layers of the network. Residual connections add the input mapping of the block to the output mapping of the block, thus requiring the block to learn this difference/residual mapping during training. Thus the block at a minimum learns the input mapping and also is more capable of learning more efficiently from this starting residual mapping position. Architectural details are given in Fig. 3.

The model weights were randomly initialized using glorot uniform and training was run three separate times, achieving three weight-variants of the same model. For inference, the average probabilistic output of the three models was used. The Rectified Adam (RAdam) optimizer, a variant of Adam, was used to adjust the weights during training.

Three postprocessing steps to reduce the number of false alarms are applied as in [16], i.e. a moving average smoothing filter of ~1 min on the probabilistic outputs, per patient adjustment by averaging the previous ~ 10 minutes of non-seizure EEG activity, and lastly all seizures detected were extended on both sides by a collar of 30 seconds.

### *C. Databases*

A publicly available database [17] is used as the source database in this study. This database consists of short 1-2h excerpts from 79 neonates who were admitted to the NICU of Helsinki University Hospital. This database comprises of 112 hours of EEG recordings with 11 hours of seizure activity from 342 seizure events. Eighteen channels of EEG were recorded at 256Hz.

The target conditions are represented by an internal database of continuous multichannel EEG, comprising data from 78 neonates with 23 experiencing seizure events recorded at NICU, Cork University Hospital. This database consists of 4,570 hours of unedited EEG recordings, with 57.7 hours of seizure activity from 1,704 seizure events. It has an 8-channel bipolar montage and was recorded at 256Hz or 200Hz.

A publicly available sample dataset [18] that represents target conditions is used. It consists of EEG recordings from 53 neonates who were diagnosed with hypoxic-ischemic encephalopathy totaling 169 hours.. This database was released for the purpose of background EEG grading and does not have seizure annotations. This database could be used as representation of the non-seizure (background EEG) class straightaway, however it was noted by the authors in [18] that a small number of seizure events might be present in the database. For this reason, it is safer to extract only non-seizure representations from this database. A pseudo labeling technique was used to extract these representations. The soft pseudo labels (probabilities) obtained with the teacher model were thresholded with percentile thresholds and the chosen percentage of the database was then hard labeled all as non-seizure and combined with the source database to form a new training database from which the student model can learn. This attenuated the mismatch between source and target conditions. As the source database has originally 18-channel bipolar montage, a similar montage was created for the sample database using the 9 referential EEG channels available so that both public databases can be combined for efficient training.

## III. Results

### A. *Performance metrics*

Area under the Receiver Operator Characteristic Curve (AUC) is used as the primary performance metric. This calculates the area under the curve when the sensitivity of the binary classifier is plotted against the specificity (or 1-specificity) at various classification threshold levels, where sensitivity is the percentage of seizure epochs correctly labeled as seizure by the model; specificity is the percentage of non-seizure epochs correctly labeled as non-seizure by the model. While AUC is a commonly used metric to compare the performances of binary classification algorithms, the AUC90 metric gives the area under the curve where Specificity>0.9. AUC90 is a representative metric of the level of false alarms which is crucial in a clinical setting and is also reported here.

While AUC/AUC90 are informative evaluation metrics that jointly measures the percentage of seizures epochs detected and the percentage of non-seizures epochs detected across a wide range of operating levels it is also beneficial to report the event metrics. These measure seizure events, namely the percentage of actual seizures detected and the percentage of seizures incorrectly detected per hour, i.e. a false alarm rate. Event metrics show the performance across various operating points and similar to the AUC90 allow comparisons across the resulting clinically appropriate low false alarm rate levels.

The relative percentage increase is also given for comparison purposes, e.g. an AUC improved from 0.95 to 0.96 represents a 20% relative percentage increase in AUC, (0.96 – 0.95)/(1 – 0.95).

### B. *Performance results*

TABLE I. Comparison of the results of adding pseudo labeled data on the target dataset.

| | **AUC\|AUC90 (%)** | **Relative increase over baseline in %** |
|---|---|---|
| Baseline (without pseudo labels or sample DB) | 96.3\|72.3 | 0.0\|0.0 |
| Baseline with pseudo labels +98.5% of sample DB | 96.8\|78.3 | 13.5\|21.7 |
| Baseline without pseudo labels + 100% of sample DB | 96.4\|77.6 | 2.7\|19.1 |

Table I shows the performance of the 3 models, baseline trained on the source database and evaluated on the target database, baseline plus 98.5% of the sample database defined by pseudo labeling, and baseline plus 100% of the sample database where all data is added as non-seizure. It can be seen that the model that utilizes the pseudo labels outperforms the models that were trained on the source database alone or simple combination of both datasets, with the AUC increasing from 96.3% to 96.8%, a 13.5% relative percentage increase.

The event metrics were calculated for each of the individual 23 nenonates that experience seizures from the target database, which in total comprises of 57.7 hours of seizure activity from 1,704 seizure events. These 23 individual event metrics were then averaged to give a representative event metric in a clinical setting, see Fig 4. The model that employs the pseudo labels surpasses the model without pseudo labels at all operating points for the event metrics. Taking a suitable clinical operating point of 0.2 false detections per hour, the model employing the pseudo labels detect 57.8% of seizure events while the model without pseudo labels detects 40.7%, a relative percentage increase of 28.8% in seizures detected.

To evaluate the impact of the quality of pseudo labels on the performance, a stronger (better performing) teacher model was obtained by training on another large continuous internal database from target conditions. This database consists of 72 neonates, with 18 experiencing seizure, 77.7 hours of annotated seizure, 889 hours of continuous EEG. The same pseudo labeling technique was used where a certain threshold on the probability of non-seizure was chosen and all signal data greater than this threshold was annotated as non-seizure. This pseudo labeled dataset was added to the source dataset and the model was retrained. It was observed that with a stronger teacher, a similar level of performance could be obtained with the addition of only 33% of the sample dataset, reaching an AUC of 96.8%.

## IV. Discussion

This research shows that the source-target mismatch can be overcome by pseudo labeling. An unannotated sample database from one center is pseudo labeled and then

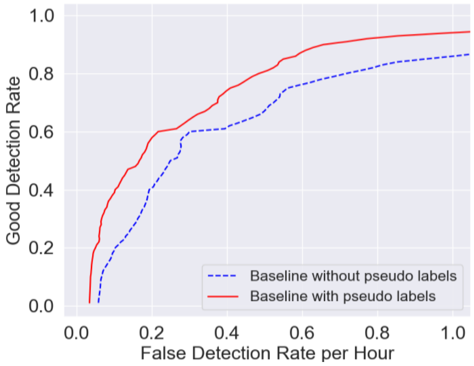

Figure 4. Event metrics comparison for the baseline model without pseudo labels and with pseudo labels.

combined with an annotated database from another center to train and fine-tune a deep learning model. The AUC relative performance increases by 13.5%. It can be seen from Table I, that the usage of all sample data without pseudo labels, i.e. assuming that all the data within is fully non-seizure data improves the performance already. This shows the robustness of the developed model. However, when the sample dataset is subjected to pseudo labeling and a small portion of the data (1.5%) with the highest probability of seizure is discarded from consideration, the performance increases even further, which is reflected in AUC and further in both AUC90 and the event metrics which represent the clinically important operating points. With pseudo labeling, there is always a trade-off between the amount of data that can be taken and the quality of the associated labels. It appears that by discarding 1.5% of the data, most seizures and seizure-like looking activities are removed and the quality of the additional dataset is increased.

An alternative is to utilize a stronger teacher. With the better quality pseudo labels coming from the stronger model, a similar level of performance was reached with only 33% of the data. This quality of the pseudo labels is influenced by the quality and quantity of the data used to train the stronger teacher. The stronger teacher is trained on a large database from the target conditions whereas the weaker teacher is trained on a smaller database from the source conditions. Thus less pseudo labels from the stronger teacher are required to reduce the source-target mismatch as they are created using target data and so are of higher quality. Regarding quantity the stronger teacher was built using a training database that is 7 times larger in terms of annotated seizure and 8 times larger in total EEG recording when compared to the training database for the weaker teacher model. As the percentage of pseudo labeled data added to the training database is increased the confidence in the annotations decreases as the likelihood of annotating seizure signal data as non-seizure increases. Thus the smaller the percentage of pseudo label data added to the training the better.

Fine-tuning the model using only the pseudo labeled database is common [1] but here it is not appropriate as there are no seizure events in the new pseudo labeled database. Instead the model is fully retrained from scratch using the combined pseudo labeled and source labeled databases. As the source labeled database is quite small we had the luxury of retraining from scratch.

Larger performance increases can be expected if both classes of interest are more comprehensively represented in the sample database.

## V. Conclusion

This study outlined the procedure that enables a model trained on a publicly available dataset to adapt to new testing conditions. It has been shown that by using the pseudo labeling method on a relatively small unannotated sample of target conditions the model performance can be marginally improved.

The outlined techniques can lead to more widespread usage of publicly available models and allows the developer to benefit from datasets that represent other testing conditions to lead to more generalizable cross-center models.

## Acknowledgment

The authors wish to thank all the clinicians involved in this research, Prof. Boylan and Dr. Mathieson of INFANT, University College Cork, Ireland.

## References

[1] D.-H. Lee, “Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks,” in *Workshop on challenges in representation learning, ICML*, 2013, p. 896.

[2] X. Zhu and Z. Ghahramani, “Learning from labeled and unlabeled data with label propagation,” Sep. 2002. Accessed: Feb. 15, 2023. [Online].https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=8a6a114d699824b678325766be195b0e7b564705.

.[3] A. Iscen, G. Tolias, Y. Avrithis, and O. Chum, “Label Propagation for Deep Semi-supervised Learning.” arXiv, Apr. 09, 2019. doi: 10.48550/arXiv.1904.04717.

[4] “HuBMAP + HPA - Hacking the Human Body.” https://kaggle.com/competitions/hubmap-organ-segmentation (accessed Mar. 06, 2023).

[5] “Feedback Prize - English Language Learning.” https://kaggle.com/competitions/feedback-prize-english-language-learning (accessed Mar. 06, 2023).

[6] G. Hinton, O. Vinyals, and J. Dean, “Distilling the Knowledge in a Neural Network.” arXiv, Mar. 09, 2015. Accessed: Mar. 06, 2023. [Online]. Available: http://arxiv.org/abs/1503.02531

[7] N. J. Stevenson, R. R. Clancy, S. Vanhatalo, I. Rosén, J. M. Rennie, and G. B. Boylan, “Interobserver agreement for neonatal seizure detection using multichannel EEG,” *Annals of Clinical and Translational Neurology*, vol. 2, no. 11, pp. 1002–1011, 2015, doi: 10.1002/acn3.249.

[8] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, “Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study,” *PLoS Med*, vol. 15, no. 11, p. e1002683, Nov. 2018, doi: 10.1371/journal.pmed.1002683.

[9] M. Artetxe and H. Schwenk, “Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond,” *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 597–610, Sep. 2019, doi: 10.1162/tacl_a_00288.

[10] C. V. Weis, C. R. Jutzeler, and K. Borgwardt, “Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review,” *Clinical Microbiology and Infection*, vol. 26, no. 10, pp. 1310–1317, Oct. 2020, doi: 10.1016/j.cmi.2020.03.014.

[11] P. Cascante-Bonilla, F. Tan, Y. Qi, and V. Ordonez, “Curriculum Labeling: Revisiting Pseudo-Labeling for Semi-Supervised Learning.” arXiv, Dec. 10, 2020. doi: 10.48550/arXiv.2001.06001.

[12] A. O’Shea, G. Lightbody, G. Boylan, and A. Temko, “Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture,” *Neural Networks*, vol. 123, pp. 12–25, Nov. 2019, doi: 10.1016/j.neunet.2019.11.023.

[13] M. Kitayama *et al.*, “Wavelet analysis for neonatal electroencephalographic seizures,” *Pediatric Neurology*, vol. 29, no. 4, pp. 326–333, Oct. 2003, doi: 10.1016/S0887-8994(03)00277-7.

[14] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 11966–11976. doi: 10.1109/CVPR52688.2022.01167.

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *arXiv:1512.03385 [cs]*, Dec. 2015, Accessed: Feb. 05, 2020. [Online]. Available: http://arxiv.org/abs/1512.03385

[16] A. Temko, G. Boylan, W. Marnane, and G. Lightbody, “Robust Neonatal EEG Seizure Detection through Adaptive Background Modeling,” *Int. J. Neur. Syst.*, vol. 23, no. 04, p. 1350018, Aug. 2013, doi: 10.1142/S0129065713500184.

[17] N. J. Stevenson, K. Tapani, L. Lauronen, and S. Vanhatalo, “A dataset of neonatal EEG recordings with seizure annotations,” *Scientific Data*, vol. 6, no. 1, Art. no. 1, Mar. 2019, doi: 10.1038/sdata.2019.39.

[18] J. M. O’Toole *et al.*, “Neonatal EEG graded for severity of background abnormalities in hypoxic-ischaemic encephalopathy.” arXiv, Jun. 09, 2022. Accessed: Jul. 28, 2022. [Online]. Available: http://arxiv.org/abs/2206.04420