CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	A deep learning approach to grading neonatal hypoxic-ischemic encephalopathy using ECG spectrograms*
Authors	Rezaei, Kimia;Mathieson, Sean R.;Lightbody, Gordon;Boylan, Geraldine B.;Marnane, William P.
Publication date	2025-12-03
Original Citation	Rezaei, K., Mathieson, S.R., Lightbody, G., Boylan, G.B. and Marnane, W.P. (2025) 'A deep learning approach to grading neonatal hypoxic-ischemic encephalopathy using ecg spectrograms*', 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Copenhagen, Denmark, 14-18, pp. 1–5. https://doi.org/10.1109/EMBC58623.2025.11253390
Type of publication	Conference item
Link to publisher's version	https://ieeexplore.ieee.org/xpl/conhome/11251507/proceeding - 10.1109/EMBC58623.2025.11253390
Rights	© 2025, the Authors. - https://creativecommons.org/licenses/by/4.0/
Download date	2026-11-01 20:12:32
Item downloaded from	https://hdl.handle.net/10468/18436

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

A Deep Learning Approach to Grading Neonatal Hypoxic-Ischemic Encephalopathy Using ECG Spectrograms*

Kimia Rezaei[1,2] Sean R. Mathieson[2] Gordon Lightbody[1,2] Geraldine B. Boylan[2,3] and William P. Marnane[1,2]

***Abstract*— The grade of Hypoxic-Ischemic Encephalopathy (HIE), a condition caused by cerebral oxygen deprivation, can be determined through analysis of Electroencephalogram (EEG). Neonates with moderate to severe HIE can be treated with therapeutic hypothermia, which has prompted the development of an automatic HIE grading system. Electrocardiogram (ECG) signals are easier and more accessible to obtain than the EEG in the Neonatal Unit and may help grade HIE. This research explores two different approaches for grading HIE, leveraging a large ECG dataset from newborns. The classical approach involves calculating heart rate (HR) from ECG signal followed by feature extraction in both time and frequency domains and classification using Random Forest and multilayer perceptron. In contrast, the proposed approach extracts the spectrograms directly from the ECG signal to serve as input for a convolutional neural network. The results indicate that the proposed method achieves a higher AUC while bypassing the time-consuming process of HR calculation and enabling the use of more robust deep learning models.**

***Clinical relevance*— Utilizing newborn ECG signals as an alternative to EEG signals for automatic HIE grading may provide a more accessible method for clinicians to grade encephalopathy and evaluate the need for hypothermia treatment.**

I. Introduction

Hypoxic-Ischemic Encephalopathy (HIE) arises from a perinatal insult leading to oxygen deprivation and inadequate blood flow to the brain. It is one of the most common causes of neonatal mortality and morbidity. HIE may result in brain injury with associated neurological conditions, including epilepsy, visual or hearing impairments, and cerebral palsy [1]. HIE is clinically graded into mild, moderate or severe, shortly after birth. Neonates with moderate to severe HIE can be treated using therapeutic hypothermia (TH) which is the process of body or brain cooling beginning in the first 6-hours following birth. This may mitigate the effects of progressive encephalopathy and improves long term outcomes [2]. Therefore, it is critical to have an early assessment of HIE grade readily accessible in the Neonatal Intensive Care Unit (NICU). Currently, continuous electroencephalogram (cEEG) or amplitude integrated EEG (aEEG) is used to help assess HIE grade. However, classifying HIE using neonatal EEG requires expert interpretation, which may not be available immediately in the Neonatal unit. This has led to studies focusing on automated HIE grading. Besides the problem of specialist availability, introducing neonatal EEG monitoring is challenging due to the high cost of the equipment and the difficulty in interpreting the results in the immediate postnatal period in the NICU [3]. This has prompted the need for alternative methods to help assess those most at risk. The electrocardiogram (ECG) is another potential biomarker which is easily accessible to clinicians and is routinely and continuously monitored in neonates. Additionally, ECG can be recorded from neonates in smaller neonatal units and lower resource settings.

Valuable diagnostic information can be obtained, by analysing and processing of the ECG signal but studies in grading HIE are limited. Existing studies have used the EEG signal [4] [5] or needed to analyse the ECG signal through the extraction of the heart rate (HR) signal [6] [7] [8]. HR variability (HRV) based methods require several preprocessing steps, including QRS complex detection, NN interval extraction and feature computation. Traditional approaches to HRV calculation, such as the Pan-Tompkins algorithm, are computationally complex and require substantial processing power, making them unsuitable for use in low-resource settings or portable devices. Additionally, robust extraction of NN intervals remains a challenging task, further complicating the implementation of these methods for clinical applications [9] [10].

This study introduces an alternative way of analysing the ECG signal for newborns with different grades of HIE. To the best of our knowledge, this is the first study utilizing ECG spectrograms for HIE grading. In this paper, two different approaches will be investigated using a large dataset:

- The classical method involves extracting the HR signal from the ECG, followed by computing HR features and classifying HIE grades using Random Forest (RF) and multilayer perceptron (MLP).
- The proposed method which directly utilizes the time-frequency information by computing the spectrogram of the ECG and classifying HIE through a 2D convolutional neural network (CNN).

This classification is used to determine whether TH needs to be employed. Therefore, HIE was classified into 2 categories: Category 0, where babies do not require TH, and Category 1, where babies require TH. The classical method and the proposed method were both applied on the same

*This publication has emanated from research conducted with the financial support of Taighde Eireann – Research Ireland under Grant number [19/FFP/6782].

[1]Electrical and Electronic Engineering, School of Engineering, University College Cork, Cork, Ireland

[2]INFANT Research Centre, University College Cork, Cork, Ireland

[3]Pediatrics and Child Health, University College Cork, Cork, Ireland

dataset for the purpose of comparison.

II. METHOD

A. Dataset

In this work, two datasets called ANSeR1 and ANSeR2 were employed [11] [12]. The EEG was recorded from neonates at risk of seizures shortly after birth. A subset of EEGs from infants with HIE were graded by a neurophysiologist as normal, mild, moderate, severe and isoelectric, with each baby having up to five 1-hour epochs of annotated signals. Epochs were extracted from the EEG records at postnatal ages of 6, 12, 24, 36, and 48 hours (if available). Recordings without ECG signal or those highly corrupted by noise were excluded.

ECG 1-hour epochs were labelled into 2 categories. Category 0 includes cases of normal and mild HIE, where babies do not require TH, and Category 1 includes cases of moderate, severe, and isoelectric HIE, where babies require TH. For each baby, there was more data available but unlabelled. In order to exploit these recordings to enhance the model performance, the signals between these intervals were weakly labelled based on the previous and post epoch labels. The weak labelling was only applied on ANSeR2 dataset which was used as training dataset.

In this study, 58 out of 91 neonates from ANSeR1 and 76 out of 90 neonates from ANSeR2 were included, with EEG recordings sampled at a rate of 256 Hz. The ANSeR1 dataset includes 216 1-hour epochs (130 class 0, 86 class 1). The ANSeR2 dataset comprises 263 strongly labelled and 1286 weakly labelled 1-hour epochs (799 class 0, 750 class 1).

B. HIE Grading Using HR Features, RF and MLP

The Pan-Tompkin method was used to identify R-peaks in the ECG signal and extract the NN intervals [13]. Although this method is generally effective, it can still produce inaccuracies or missed peaks during QRS detection. The main reason for this issue is that this method is primarily designed to detect peaks in typical heart rhythms, whereas real-world ECG recordings are often affected by various types of artifacts and noise. During HIE, changes in HR can introduce abnormalities in the ECG signal, further complicating accurate detection[8]. This can impact the accuracy of computed values during the feature extraction phase. An example of peak detection from a noisy ECG signal segment using the Pan-Tompkin method is illustrated in Fig. 1. As shown, during the spikes, some peaks were either not detected or incorrectly identified.

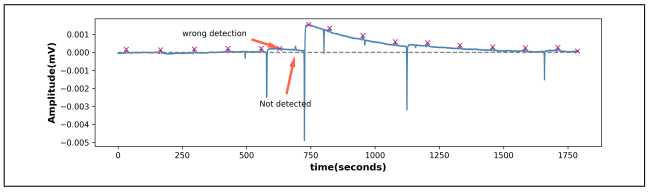


Fig. 1. An example of a noisy ECG signal where Pan-Tompkins missed some peaks and falsely detected others.

The Hampel filter was applied to eliminate outliers from NN interval data, preparing it for further feature extraction. For adults, the average resting HR typically falls between 60-100 BPM, whereas it is notably higher in newborns. Therefore, selecting an epoch length of 4 minutes strikes a balance between an optimal window for HR data analysis and sufficient duration to accurately capture seizure events and other abnormalities [14] [15]. Features were then calculated for each 4-minute segment of the HR signal. An 80% overlap is applied to minimize information loss at the segment boundaries, ensuring a better representation of signal patterns.

Sixteen time and frequency domain features were extracted from each segment [16]. Frequency domain parameters were extracted by estimating the power spectral density (PSD) of NN series using Welch's method. Therefore, the NN sequences need to be resampled at uniform intervals. A spline interpolation technique which involves fitting low-degree polynomials to small subsets of the values was used for this purpose. The four extracted features include Very Low Frequency (VLF) power, Low Frequency (LF) power, High Frequency (HF) power with frequency bands of (0.01-0.04 Hz),(0.04- 0.2 Hz) and (0.2- 2 Hz) respectively, along with the LF/HF ratio. Time domain features include Mean of the NN interval (muNN), the standard deviation of NN Series (SDNN), root mean of squared NNI differences (RMSSD), skewness and kurtosis of NN distribution were extracted. In addition, a geometrical parameter called the triangular index, which represents the ratio between the number of NNI and the maximum of the NNI histogram was calculated. Nonlinear features Approximate Entropy (ApEn) to measure regularity and complexity of a stationary signal, Poincaré parameters SD1, SD2, SD2/SD1 ratio for self-similarity computation [17] and Cardiac Vagal Index (CVI) and the modified Cardiac Sympathetic Index (CSI) [18] were also computed.

Supervised classification was performed using a RF model, with hyperparameters optimized through grid search. The final selected parameters are presented in Table I.

The MLP model built in this study consists of three fully connected layers. The first two layers are followed by batch normalization, rectified linear unit (ReLU) activation, and dropout. The input to the model comprises 16 features, so the first layer is configured with 16 neurons. The number of neurons increases to 64 in the second layer to capture more information from the data. In the third layer, the number of neurons is reduced to 32, and the final layer outputs 2

TABLE I
OPTIMIZED HYPERPARAMETERS OF THE RANDOM FOREST MODEL FOR HIE GRADING, DETERMINED THROUGH GRID SEARCH

parameter	value	parameter	value
number of trees	1000	min samples split	10
max depth	7	min samples leaf	5
max features	4	criterion	gini

neurons, corresponding to the number of classes. The dropout probabilities were set to 0.6 for both layers to enhance regularization. The model's loss was calculated using the cross-entropy loss function. Training was conducted with a batch size of 1,024 over 100 training iteration utilizing AdamW optimization with an initial learning rate (LR) of 10^{-5}. The model had 3,426 parameters and feature weights were initialized using Kaiming method.

C. HIE Grading Using Spectrograms and CNN

This method is proposed to address the limitations and shortcomings of the classical approach. In this method, the time-frequency information of the ECG signal is directly computed using the spectrogram. Spectrograms are one of the common tools of signal representation by computing the change of a nonstationary signal's frequency content over time using consecutive Fourier transforms. The use of ECG spectrograms has also been demonstrated to be effective in classifying the severity of tetanus [19], despite infections present a relatively simpler classification challenge through ECG compared to neurological conditions. Additionally, utilizing spectrograms enables the use of deep learning models like CNNs, which have proven to surpass classical methods in robustness and offer several advantages. CNNs automatically learn features directly from raw data and are specifically designed to handle spatial information While traditional approaches depend on manual feature engineering. CNNs are translation-invariant, meaning they can recognize shifts in the input data and are more resilient to noise.

In the first step, a notch Filter was applied to the ECG signals to remove the 50 Hz noise. Next, a high pass Butterworth filter with a cut off frequency of 0.5 Hz was used to remove baseline drift. Spectrograms were computed for each 4-minute segment of denoised ECG signal. The Hamming window with size of 28 seconds with 1 second shift between consecutive windows was chosen. To emphasize lower frequencies, a logarithmic spectrogram was calculated and trimmed to a maximum frequency of 9 Hz capturing the essential initial harmonics of the signal [20]. This transformation adjusted the spectrogram shape from 3585×213 to a more square-like shape of 253×213. This improves CNN processing by simplifying the application of convolutional and pooling layers, which operate more efficiently with balanced aspect ratios. An example of a 4-minute segment of ECG logarithmic spectrogram is illustrated in Fig. 2. The spectrogram matrix is a 2D array where each element represents the power spectral density of the signal at a given frequency and time. These matrices were then standardized to have a mean of zero and a standard deviation of one, then used as input for the classification.

The input tensor has a shape of $1 \times 253 \times 213$ representing the number of channels, height, and width, respectively. The simplicity of the input data allowed the design of a shallow yet effective and time-efficient CNN. The proposed CNN architecture consists of 4 blocks for feature extraction each consisting of a convolutional layer followed by a batch normalization, ReLU and max pooling. The filter size of the first convolutional layer is 5×5. The max pooling and batch normalization layers were applied afterward to reduce the number of parameters to train and help to limit overfitting. The max pooling layer decreases the spatial dimensions of the matrix, but the receptive field, which is the size of the region in the input that produces the feature, increases [21]. Therefore, the filter size was reduced to 3×3 in the second and third convolutional layer to capture more local information from the input data. Consequently, the filter size of the fourth convolutional layer was further reduced to 2×2. The final layer of this model is a global-averaging layer, producing one sample for all channels. Afterwards, classification was performed by 2 fully connected layers. The overview of the proposed CNN is illustrated in Fig. 3.

Cross entropy loss function was used and the network trained using AdamW optimization. Weights initialization was employed using Kaiming method. The model was trained with a batch size of 1,024 over 100 training iterations to optimize the loss and a starting LR of 10^{-6} was used but scheduled to be increased during training. Additionally a weight decay value of 10^{-4} was considered. The CNN only required 21,192 trainable parameters. All experimental procedures are executed using PyTorch 2.3 on Tesla V100-SXM2-16GB GPU.

D. Training and testing procedure for HIE grading

In all methods, the classifiers were trained on the ANSeR2 dataset and tested on the totally unseen data which was ANSeR1 dataset. During the training process for loss optimization of CNN and MLP models, the training data (ANSeR2) was randomly split into training (80%) and validation (20%) sets. Training was stopped when the validation loss began to increase, indicating the onset of overfitting. The model corresponding to the highest validation AUC was selected and subsequently evaluated on the entirely unseen ANSeR1 dataset, comprising unedited clinical data from multiple hospitals, to assess its generalization performance. The predicted probabilities of all test samples in each 1-hour epoch were averaged, and this mean probability determined the HIE grade for each baby. The classification performance on the test data was assessed using the area under the

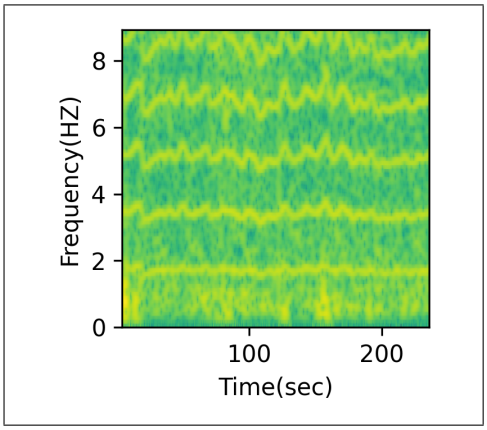


Fig. 2. Logarithmic spectrogram of a 4-minutes ECG segment (253×213)

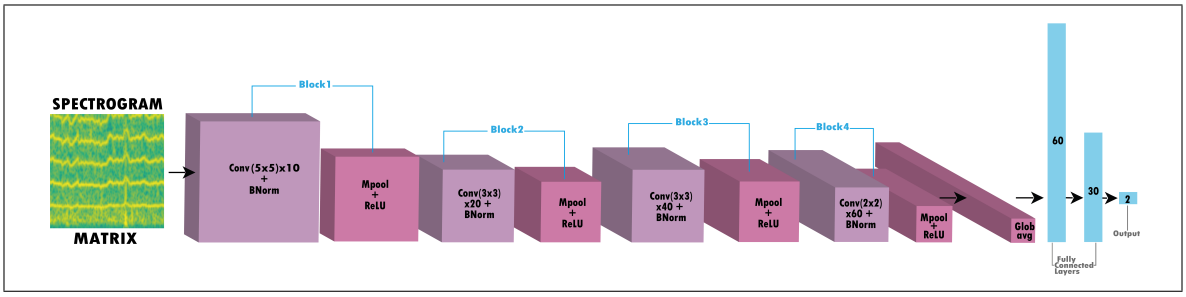


Fig. 3. Proposed CNN structure. The input is a 2D spectrogram matrix derived from a 4-minute ECG segment. The model consists of four blocks, each comprising a convolutional layer (Conv), batch normalization (BNorm), and max pooling (MPool). A global average layer (Glob-ave) is applied at the end to aggregate the extracted features, followed by a fully connected layer with two linear layers for classification.

receiver-operating curve (AUC) metric and confusion matrices.

III. RESULT

Classifiers using HR features as input achieved AUCs of 79.18% for RF and 79.59% for MLP, with the deep learning based MLP marginally outperforming RF. For the proposed method using the spectrogram as input to the CNN, an AUC of 80.83% was obtained, demonstrating its higher performance. Fig. 4 illustrates the ROC curves and AUC values for all models over the unseen test set. While the AUC values are relatively close, the CNN with spectrogram input achieves the highest performance.

The performance of the proposed method varies with spectrogram length, shape, and window size. To optimize performance, different spectrogram parameters were tested with the CNN, as detailed in Table II. Parameters were selected with an emphasis on maintaining square-shaped spectrograms for balanced and effective results. The best performance was achieved using a 4-minute spectrogram with a 28-second window size.

Fig. 5 illustrates the classification results of all models across the 5 clinical HIE grades. The charts display the correct and incorrect binary classification of 1-hour ECG epochs, with the epochs grouped according to their clinical grades. A notable observation is the frequent incorrect classification of mild HIE cases as class 1, which could lead to inappropriate treatment with TH. After re-evaluation by an expert neurophysiologist, 17 1-hour epochs were identified as borderline cases between mild and moderate grades. Due to the complexity of HIE signals and the inherent difficulty in assigning objective labels, these cases could not be clearly annotated and were therefore excluded. The confusion matrices for all models shown in Fig. 6, demonstrate that the number of incorrect classifications was significantly reduced after removing the borderline cases. Furthermore, the CNN accuracy increased from 71.3% to 77.39%.

TABLE II

COMPARISON OF CNN HIE GRADING AUC RESULTS FOR SPECTROGRAMS WITH DIFFERENT PARAMETERS. SEGMENT LENGTH AND WINDOW SIZE ARE GIVEN IN SECONDS(S); IMAGE SIZE IS IN PIXELS(P); MAXIMUM FREQUENCY (FREQ) IS IN HERTZ (HZ); AND AUC VALUES ARE REPORTED AS PERCENTAGES.

Segment length (S)	Image size (P)	Window size (S)	Maximum Freq (Hz)	AUC (%)
240	253×213	28	9	80.83
120	113×105	16	7	79.86
300	251×251	50	5	78.81

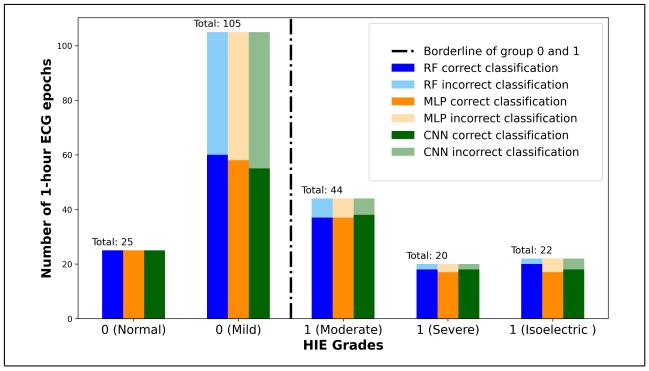


Fig. 5. Models correct and incorrect binary classification of 1-hour ECG epochs, with the epochs grouped according to their clinical grades.

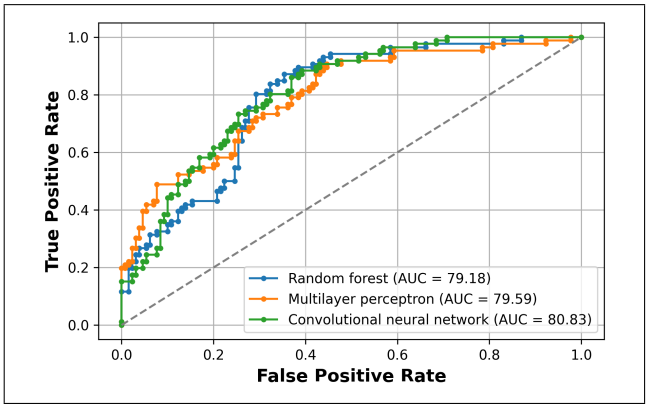


Fig. 4. ROC curves and corresponding AUC values for the Random Forest and Multilayer Perceptron models using extracted features, and the Convolutional Neural Network using spectrograms. Results are based on the 1-hour epoch from an unseen test set of patients.

IV. CONCLUSION

In this work, the experimental results demonstrate the potential of using ECG spectrograms and deep learning models

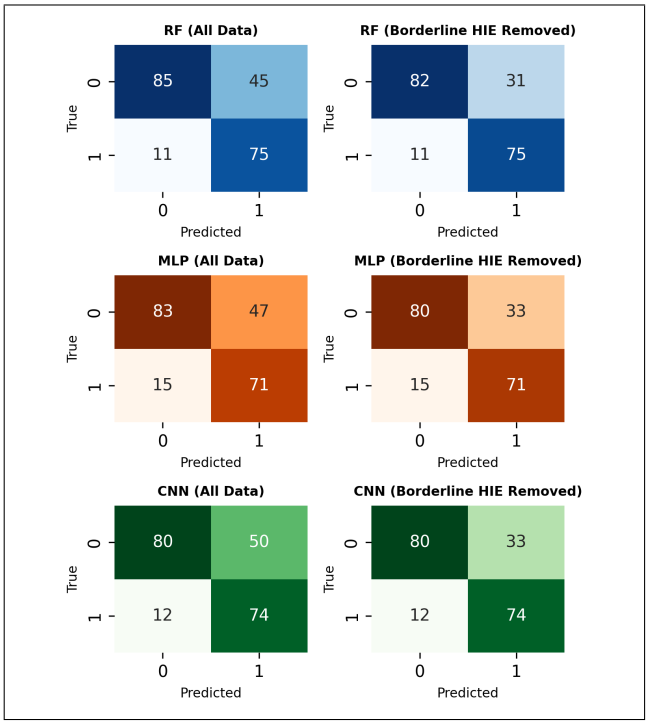


Fig. 6. Models confusion matrices based on all data and after removing the border line HIE 1-hour epochs.

for HIE grading, although the results are marginally better than traditional HRV-based methods. However, this has the advantage of removing the process of NN interval extraction from ECG data which is normally employed using the Pan-Tompkins method, which is time-consuming, not completely accurate, and requires extensive data cleaning. Additionally, using spectrograms allows for the leverage of state-of-the-art robust deep learning models. Moreover, an AUC of 80.83% highlights the significant potential for further exploration in automatic HIE grading, though additional investigation is needed. This result advances the ability to detect HIE grades, reducing reliance on constant expert availability and enabling use in lower-resource hospitals. Using the easily obtained ECG signal may help clinicians to objectively grade the severity of encephalopathy which may in turn help assess suitability for therapeutic hypothermia. Limitations of this study include variability of inter-rater label agreement particularly on borderline cases. This study did not take account of the effect of TH and medication.

References

[1] P. Greco *et al.*, "Pathophysiology of hypoxic–ischemic encephalopathy a review of the past and a view on the future," *Acta Neurologica Belgica*, vol. 120, pp. 277–288, 2020.

[2] G. Wassink, J. O. Davidson, S. K. Dhillon, K. Zhou, L. Bennet, M. Thoresen, and A. J. Gunn, "Therapeutic hypothermia in neonatal hypoxic-ischemic encephalopathy," *Current Neurology and Neuroscience Reports*, vol. 19, pp. 1–10, 2019.

[3] D. Neubauer, D. Osredkar, D. Paro-Panjan, A. Skofljanec, and M. Derganc, "Recording conventional and amplitude-integrated eeg in neonatal intensive care unit," *European Journal of Paediatric Neurology*, vol. 15, no. 5, pp. 405–416, 2011.

[4] S. A. Raurale, G. B. Boylan, S. R. Mathieson, W. P. Marnane, G. Lightbody, and J. M. O'Toole, "Grading hypoxic-ischemic encephalopathy in neonatal eeg with convolutional neural networks and quadratic time–frequency distributions," *Journal of Neural Engineering*, vol. 18, no. 4, p. 046007, 2021.

[5] S. Yu, W. P. Marnane, G. B. Boylan, and G. Lightbody, "Neonatal hypoxic-ischemic encephalopathy grading from multi-channel eeg time-series data using a fully convolutional neural network," *Technologies*, vol. 11, no. 6, 2023.

[6] R. Ahmed, A. Temko, W. P. Marnane, G. Boylan, and G. Lightbody, "Classification of hypoxic-ischemic encephalopathy using long term heart rate variability based features," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 2355–2358.

[7] V. Matic, P. J. Cherian, D. Widjaja, K. Jansen, G. Naulaers, S. Van Huffel, and M. De Vos, "Heart rate variability in newborns with hypoxic brain injury," in *Oxygen Transport to Tissue XXXV*. Springer, 2013, pp. 43–48.

[8] R. M. Goulding, N. J. Stevenson, D. M. Murray, V. Livingstone, P. M. Filan, and G. B. Boylan, "Heart rate variability in hypoxic ischemic encephalopathy: correlation with eeg grade and 2-y neurodevelopmental outcome," *Pediatric research*, vol. 77, no. 5, pp. 681–687, 2015.

[9] M. Bolanos, H. Nazeran, and E. Haltiwanger, "Comparison of heart rate variability signal features derived from electrocardiography and photoplethysmography in healthy individuals," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2006, pp. 4289–4294.

[10] R. Gutierrez-Rivas, J. J. Garcia, W. P. Marnane, and A. Hernandez, "Novel real-time low-complexity qrs complex detector based on adaptive thresholding," *IEEE Sensors Journal*, vol. 15, no. 10, pp. 6036–6043, 2015.

[11] A. M. Pavel *et al.*, "A machine-learning algorithm for neonatal seizure recognition: a multicentre, randomised, controlled trial," *The Lancet Child and Adolescent Health*, vol. 4, no. 10, pp. 740–749, 2020.

[12] J. M. Rennie *et al.*, "Characterisation of neonatal seizures and their treatment using continuous eeg monitoring: a multicentre experience," *Archives of Disease in Childhood-Fetal and Neonatal Edition*, vol. 104, no. 5, pp. F493–F501, 2019.

[13] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 3, pp. 230–236, 1985.

[14] O. Doyle, A. Temko, W. Marnane, G. Lightbody, and G. Boylan, "Heart rate based automatic seizure detection in the newborn," *Medical engineering and physics*, vol. 32, no. 8, pp. 829–839, 2010.

[15] G. Volpes, C. Bara, A. Busacca, S. Stivala, M. Javorka, L. Faes, and R. Pernice, "Feasibility of ultra-short-term analysis of heart rate and systolic arterial pressure variability at rest and during stress via time-domain and entropy-based measures," *Sensors*, vol. 22, no. 23, p. 9149, 2022.

[16] K. Rezaei, K. Yu, S. R. Mathieson, A. Flynn, G. Lightbody, G. Boylan, and W. P. Marnane, "Assessing the effectiveness of heart rate variability as a diagnostic tool for brain injuries in infants," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024, pp. 1–4.

[17] M. Toichi, T. Sugiura, T. Murai, and A. Sengoku, "A new method of assessing cardiac autonomic function and its comparison with spectral analysis and coefficient of variation of r–r interval," *Journal of the autonomic nervous system*, vol. 62, no. 1-2, pp. 79–84, 1997.

[18] J. Jeppesen, S. Beniczky, P. Johansen, P. Sidenius, and A. Fuglsang-Frederiksen, "Detection of epileptic seizures with a modified heart rate variability algorithm based on lorenz plot," *Seizure*, vol. 24, pp. 1–7, 2015.

[19] P. Lu *et al.*, "Improving classification of tetanus severity for patients in low-middle income countries wearing ecg sensors by using a cnn-transformer network," *IEEE Transactions on Biomedical Engineering*, vol. 70, no. 4, pp. 1340–1350, 2022.

[20] L. Cabrera-Quiros, G. Varisco, Z. Zhan, X. Long, P. Andriessen, E. J. Cottaar, and C. van Pul, "Estimation of heart rate directly from ecg spectrogram in neonate intensive care units," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2020, pp. 320–323.

[21] A. Araujo, W. Norris, and J. Sim, "Computing receptive fields of convolutional neural networks. distill, 2019," 2019.