Development of an EEG artefact detection algorithm and its application in grading neonatal hypoxic-ischemic encephalopathy

Mark E. O'Sullivan (mark.osullivan@ucc.ie) [a], Gordon Lightbody (g.lightbody@ucc.ie) [a, b], Sean R. Mathieson (sean.mathieson@ucc.ie) [a], William P. Marnane (l.marnane@ucc.ie) [a, b], Geraldine B. Boylan (g.boylan@ucc.ie) [a, c], John M. O'Toole (jotoole@ucc.ie) [a,c]

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

[b] School of Engineering, University College Cork, Cork, Ireland

[c] Department of Pediatrics and Child Health, University College Cork, Cork, Ireland

Abstract

Objective: The primary aim of this study is to develop and evaluate algorithms for neonatal EEG artefact detection. The secondary aim is to subsequently assess its application as a post-processing routine for automated EEG grading of background abnormalities in neonatal hypoxic-ischemic encephalopathy (HIE). *Methods:* A database of neonatal EEG with expertly annotated artefacts was used to train and validate machine learning models to automatically identify EEG epochs containing artefacts. Three approaches using a simple threshold-based digital signal processing (DSP) method, a machine learning method, and a deep learning method were developed and compared. The artefact detection classifier was subsequently assessed as a post-processing tool to assist in the application of automated EEG grading of HIE. A new deep learning model for grading the EEG was developed by training an existing network on a large, multi-centre dataset. The artefact detection algorithm was integrated into the grading algorithm through a post-processing routine. *Results:* Using a database containing 19 hours of EEG from 51 patients with per-channel and per-second annotations of artefacts, a deep learning convolutional neural network solution achieved best performance for artefact detection with an area under the operating characteristic curve (AUC) of 0.84, compared to an AUC of 0.68 and 0.82 for a DSP method and a random-kernel ridge-classifier model, respectively. The automated EEG grading algorithm was trained and tested on 653 hours of EEG from 181 patients, which achieved an accuracy of 82.8% (95% CI: 80.5% to 85.2%). The percentage of detected artefacts in the misclassified epochs was not statistically different (p=0.568) compared to that of correctly classified epochs. Using artefact detection, a small number of epochs were removed from grading, resulting in a minor increase in accuracy for the EEG grading algorithm from 82.6% to 83.6%.

Conclusion: Deep learning methods achieved highest classification performance for neonatal EEG artefact detection, although a ridge classifier using random kernels achieved comparable performance without significant parameter tuning or training time. The inclusion of artefact detection in automated EEG grading does not significantly improve accuracy in our curated dataset, but does allow for a quality measure to be presented alongside the automated EEG grades which may increase user confidence in its real-world application.

Keywords: EEG; convolutional neural network; artefact detection; hypoxic-ischemic encephalopathy; neonatal; machine learning.

1. Introduction

Electroencephalography (EEG) is the gold standard for assessing neonatal neurological function. It is used for tasks such as seizure detection and to determine both the severity of ongoing hypoxic ischaemic injury and outcome thereafter (Garvey, et al., 2021). Interpretation and grading of EEG is a difficult task and must be completed by an electroencephalographer. Computer-based automated analysis of EEG has the potential to provide continuous, objective and scalable solutions to assist in EEG review. However, artefacts in the EEG can cause uncertainty in the automated analysis and reduce algorithm performance.

Neonatal EEG approximates a 1/f process, whereby spectral power is inversely proportional to the frequency of the signal, with most of the spectral power contained below 10Hz (Rankine, Stevenson, Mesbah, & Boashash, 2007). Comprehensive pre-processing routines are used to filter and down-sample the EEG signals with the aim of reducing noise and unwanted waveforms (Raurale, et al., 2021) (Moghadam, et al., 2021). However, due to the low amplitude of EEG signals (mostly within ±50μV peak-to-peak), EEG is highly susceptible to artefacts that are within the frequency range of interest and are magnitudes of order larger in amplitude (Sai, Mokhtar, Arof, Cumming, & Iwahashi, 2018). The EEG electrodes and cables can act as antennae capable of coupling with interfering propagating signals. Non-biological sources of artefact can include 50Hz mains noise and interference from other machines in the intensive care unit, such as pumps and ventilators (Alvarez & Rossetti, 2015). Scalp EEG electrodes are receptive to many types of electrophysiological signals, such as the electrocardiogram (ECG), electromyogram (EMG) and electrooculogram (EOG), which can also be recorded by the EEG electrodes and are often much larger in amplitude compared to the EEG. This presents a significant challenge to the personnel interpreting the EEG, as these artefactual waveforms often contain similar characteristics to the EEG (De Vos, et al., 2011). For example, periodic artefacts (ECG, pulse and respiration), can be misinterpreted by non-experts as seizures, and visa-versa.

Prior art has shown improvements in classifications tasks, such as neonatal seizure detection, by removing artefactual EEG (De Vos, et al., 2011) (Khlif, Mesbah, Boashash, & Colditz, 2010). Expert review of a state-of-the-art neonatal seizure detection algorithm showed that over 70% of false detections by the algorithm were due to artefacts, with respiration artefacts being the leading contributor (Mathieson, et al., 2016). Not surprisingly, automated detection and rejection of artefactual EEG signals has been widely researched. Such methods include digital signal processing (De Vos, et al., 2011) (O'Toole & Boylan, 2017) (Nolan, Whelan, & Reilly, 2010), machine learning (Sai, Mokhtar, Arof, Cumming, & Iwahashi, 2018) (Tamburro, Croce, Zappasodi, & Silvia, 2021) (Yasoda, Ponmagal, Bhuvaneshwari, &

Venkatachalam, 2020), and more recently, deep learning techniques (Webb, Kauppila, Roberts, Vanhatalo, & Stevenson, 2021) (Khatwani, et al., 2018). Neonatal EEG recorded in the neonatal intensive care unit is considerably different to adult EEG, and the intensive care unit presents added challenges due to biological and environmental interference. Hence, there is a body of research dedicated to artefact detection in neonatal EEG explicitly (De Vos, et al., 2011) (Khlif, Mesbah, Boashash, & Colditz, 2010) (O'Toole & Boylan, 2017)[14] (Webb, Kauppila, Roberts, Vanhatalo, & Stevenson, 2021). Due to the varying temporal and spectral characteristics of multiple types of EEG artefacts, it is challenging to achieve consistent performance across all artefact types with one classifier. Therefore, many systems focus on classifying individual artefact types or grouping artefacts by similar characteristics (Mannan, Kamran, & Jeong, 2018) (Stevenson, O'Toole, Korotchikova, & Boylan, 2014).

Hypoxic-ischemic encephalopathy (HIE) occurs in 1-3 per 1,000 births in high income countries and contributes significantly to neonatal mortality and morbidity rates (Kurinczuk, White-Koning, & Badawi, 2010) (Lee, et al., 2013). HIE occurs due to a lack of oxygen and blood supply in the brain and is graded clinically as mild, moderate, or severe. Therapeutic hypothermia is used to treat neonates diagnosed with HIE by lowering the core temperature of the neonate for 72 hours within 6 hours of birth (Gluckman, et al., 2005). This treatment has been shown to be effective in improving neonatal outcomes in moderate to severe encephalopathy (Davies, Wassink, Bennet, Gunn, & Davidson, 2019). There has been significant research in recent years in the area of automated EEG interpretation and grading of EEG abnormalities in neonatal HIE to provide diagnostic decision support (Abbasi & Unsworth, 2020) (Matic, et al., 2014) (Ahmed, Temko, Marnane, Lightbody, & Boylan, 2016) (Raurale, et al., 2021) (J, Cheng, & Wu, 2020) (Moghadam, et al., 2021).

This paper presents the evaluation of three methods to detect artefacts in neonatal EEG: an ad-hoc threshold-based approach, a linear classifier with random convolutional kernel transforms (known as ROCKET) with set parameters, and a convolutional neural network (CNN) designed specifically for this task. All methods are developed for the binary classification of multiple EEG artefact types, including instantaneous artefacts, such as movement and muscle artefacts, and longer-term periodic artefacts, such as ECG and respiration. To test the application and real-world usability of the artefact detection, a new machine-learning model to grade abnormal activity in the EEG for HIE severity is presented by retraining the architecture used in (Raurale, et al., 2021) on a large, multi-centre dataset. The artefact detection algorithm is subsequently included as a post-processing routine with the automated HIE grading algorithm to evaluate if increased performance can be achieved by integrating artefact detection.

2. Materials

2.1. Dataset

EEG was recorded from term infants from 8 European tertiary neonatal intensive care units (NICUs) across 4 European countries (Ireland, United Kingdom, Netherlands, and Sweden) (ClinicalTrials.gov identifiers NCT02160171 and NCT02431780). 472 neonates had EEG monitoring shortly after birth for a minimum of 6 hours and up to 100 hours (Pavel, et al., 2021). EEGs were recorded using either the NicoletOne ICU Monitor (Natus, WI, USA), Xltek EEG System (Natus), or Neurofax EEG-1200 (Nihon Kohden, Tokyo, Japan). Electrodes were placed according to the 10:20 electrode placement system adapted for neonates using 9 disposable electrodes positioned at F3, F4, C3, C4, Cz, T3, T4, O1 (or P3), and O2 (or P4). The EEG was recorded at a sampling rate of 200, 250, or 256 Hz. All sites obtained written and informed consent before commencing EEG recording.

2.1.1. Annotations for artefact detection

A neurophysiologist reviewed the data and selected 10-minute sections of EEG that contained significant amounts of artefact and another 10-minute section of EEG from the same patient that were generally artefact-free. The duration of these epochs (10-minutes) was chosen as an appropriate trade-off to ensure that more babies could be included rather than less babies with longer epochs. The neurophysiologist then manually annotated the start and end time of each individual artefact, the EEG channel it was present on, and specifically what type of artefact it was. This resulted in a dataset containing 114 10-minute sections of 8-channel EEG from 51 patients with per-channel and per-second annotations. Subsequently, in the generally artefact-free 10-minute sections, the neurophysiologist annotated any short-duration and non-descript artefacts for removal (as they would otherwise be falsely labelled as in the artefact-free EEG section). This reduced the total duration of the dataset to 12.3 hours of multi-channel EEG data. As detailed in Table 1, there was 2.65 hours of specific, per-channel, artefacts annotated and 9.65 hours of artefact-free EEG annotated in the final dataset for algorithm development. Type of artefacts annotated in the dataset were poor electrode contact, movement, muscle, sweat, soother sucking, ECG, respiration, and pulsatile artefacts. Poor electrode contact, movement, and sweat artefacts are typically large in amplitude, while ECG, respiration, pulsatile, and soother sucking artefacts are generally lower in amplitude and are periodic waveforms that can persist throughout EEG recordings.

2.1.2. EEG grading of HIE

EEG was collected from 472 neonates (Pavel, et al., 2021). From this, 266 were confirmed to have a HIE diagnoses only. Seventeen patients were excluded as they did not have EEG within 48 hours of birth and 68 neonates were randomly selected to be held out for a future validation set. From the remaining 181 patients, 1-hour epochs of EEG were pruned from the EEG recordings at 6, 12, 24, 36, and 48 hours post-natal age, where available. Each 1-hour epoch was graded by an expert (SRM and GBB) according to the grading system proposed in (Murray, Boylan, Ryan, & Connolly, 2009), which identifies four individual grades:

- Grade 1 – normal EEG and/or mild abnormalities: continuous background pattern with normal physiological features (e.g. anterior slow waves), or slightly abnormal activity (e.g. mild asymmetry, mild voltage depression, or poorly defined sleep-wake cycling (SWC)).
- Grade 2 – moderate abnormalities: discontinuous activity with inter-burst intervals (IBI) of < 10 seconds, no clear SWC, or clear asymmetry or asynchrony.
- Grade 3 – major abnormalities: discontinuous activity with IBI of 10-60 seconds, severe attenuation of background patterns or no SWC.
- Grade 4 – inactive: background activity of < 10 μV or severe discontinuity with IBI of > 60 seconds.

This resulted in a total of 653, one-hour, 8-channel EEG epochs for training of the EEG-HIE grading algorithm and testing without and with automated artefact detection in the post-processing routine. Examples of EEG-HIE grades are shown in Figure 2 and the distribution of the dataset by EEG grades and by time point (post-natal age) is shown in Table 2.

Label	*No. of 8-second epochs of single-channel EEG with 0% overlap*	*No. of minutes of multi-channel EEG*	*Percentage of dataset*
Clean	34769	579	78.47%
Poor electrode contact	2543	42.6	5.74%
Movement	2160	36	4.88%
Muscle	346	6	0.78%
Sweat	1352	22.8	3.05%
Soother sucking	48	0.6	0.11%
ECG	962	16.2	2.17%
Respiration	1879	31.2	4.24%
Pulsatile	248	4.2	0.56%

Table 1. Description and distribution of artefact annotations.

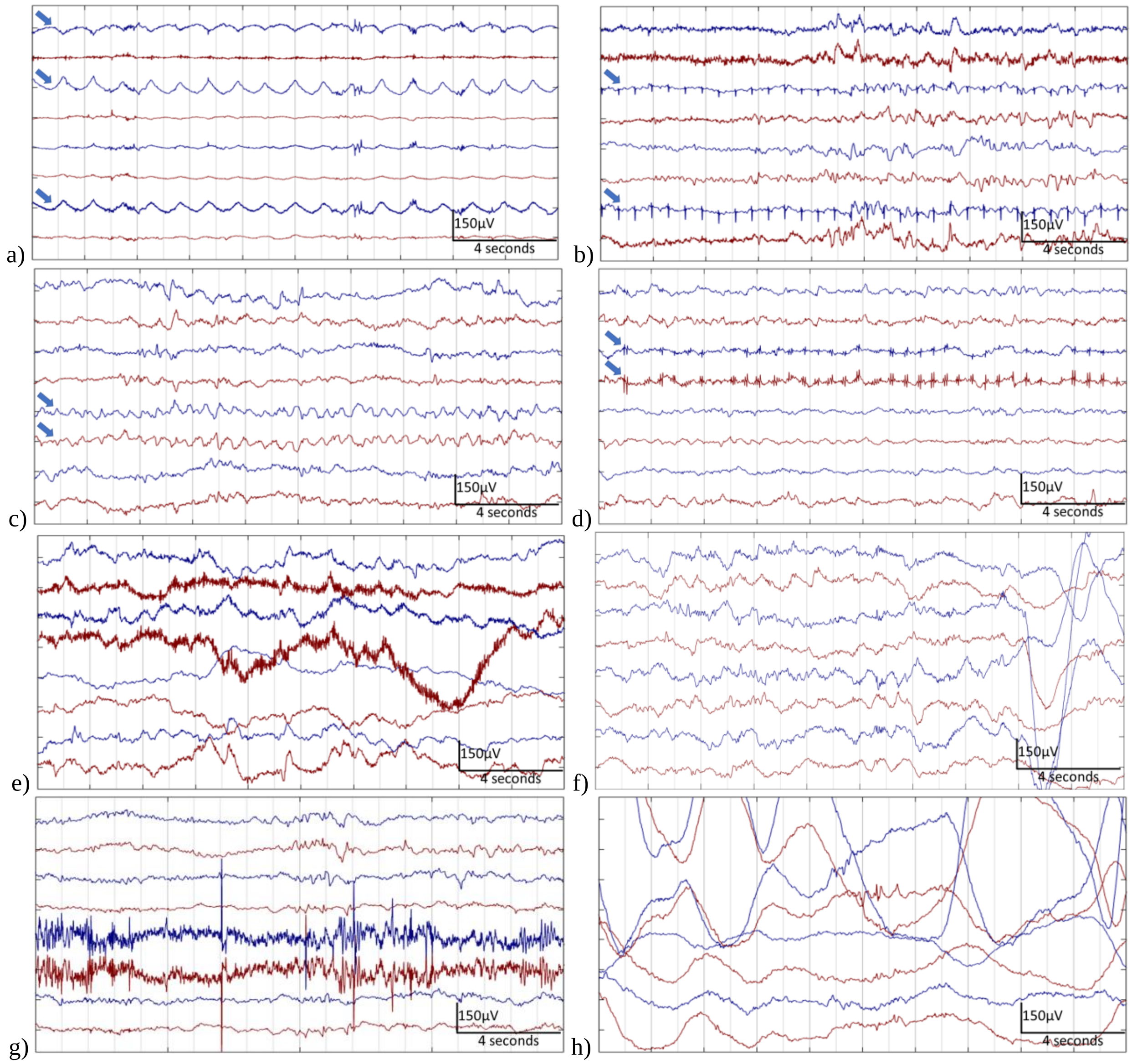

Figure 1. Examples of neonatal EEG artefacts. a) Respiration artefact, b) ECG artefact, c) Pulsatile artefact, d) Soother-sucking artefact, e) Muscle artefact, f) Movement artefact, g) Poor electrode contact artefact, and h) Sweat artefact.

Class	*No. of 1-hour EEG files*
Grade 1	353
Grade 2	166
Grade 3	66
Grade 4	68

Time point (post-natal age)	*No. of 1-hour EEG files*
6 hour	76
12 hour	124
24 hour	156
36 hour	152
48 hour	145

Table 2. a) Distribution of EEG grades and b) Distribution of post-natal age time points

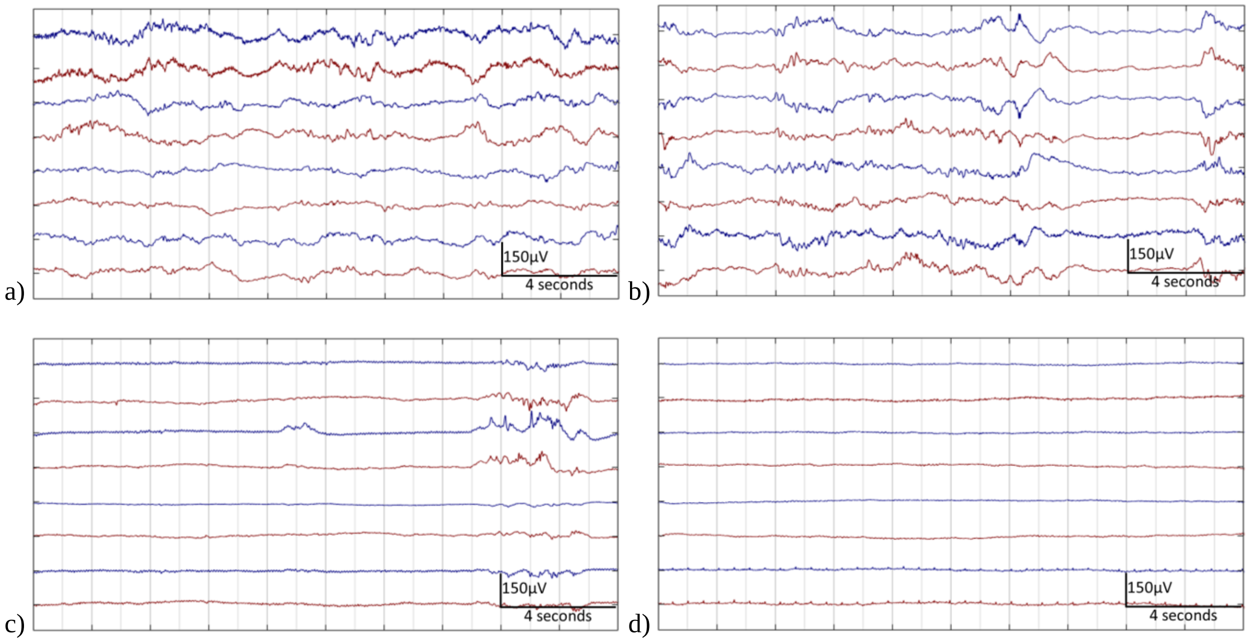

Figure 2. EEG grades a) 1 b) 2, c) 3, and d) 4.

2.2. Prior art and methods used

2.2.1. NEURAL

A threshold-based DSP algorithm was developed as part of the pre-processing routine to remove artefacts from neonatal EEG. This approach, as part of the neonatal EEG feature set in Matlab package (NEURAL) [16], was proposed as a simple and standard method for artefact removal and has been widely used across various applications, including seizure detection, sleep classification and HIE grading (Nandy, et al., 2019) (Abdur Rahman, et al., 2019) (Raurale, et al., 2021).

In this work, the DSP artefact rejection algorithm is used to establish baseline performance for artefact rejection without machine learning. The DSP method identifies artefacts based on amplitude of the EEG, odd-one-out channels (poor electrode contact), and low power channels (associated with electrode coupling). The rule-based procedure is used as follows:

i. ***High- and low-amplitude artefact using the bipolar montage:***

   a. Band-pass filter the signal between 0.5–10Hz

   b. Calculate envelope of signal using $|z(n)| = |x(n) + jH[x(n)]|$, where H is the Hilbert Transform.

   c. High-amplitude: if |z(n)| > 300 μV, then apply a 10-second collar and label as artefact. This label is applied across all channels in the same hemisphere.

   d. Low-amplitude: if |z(n)| > 100 μV, apply a 3-second collar. The label is only applied to the current channel.

   e. Repeat for all channels .

ii. ***Poor electrode artefact using the referential montage:***

   a. Band-pass filter the signal between 0.5–20Hz.

   b. Calculate the correlation coefficient among all channels ($N_{chan} \times N_{chan}$). For each channel, compute the mean correlation from this channel to all other channels. If this is less than 0.05, then label this channel as artefact.

iii. ***Electrode-coupling artefact in bipolar montage:***

   a. Band-pass filter the signal between 0.5–20Hz.

   b. Calculate the power in channel *c* using: $p_c = \frac{1}{N}\sum_{n=0}^{N-1}|x(n)|^2$, where *N* is the length of EEG *x(n)*

   c. Compute threshold as: $thres = median([p_1, p_2, \ldots, p_{Nchan}])\ /\ 4$

   d. Label the channel c as artefact if $p_c < thres$

   e. Repeat step d. for all channels in the hemisphere and repeat for the other hemisphere.

The high–low amplitude detector is a modified version of the NEURAL method (O'Toole & Boylan, 2017), developed to better quantify movement artefact in the EEG (Raurale, et al., 2021). The order is important: first the high amplitude artefacts are identified and removed before the low-amplitude artefacts.

2.2.2. Random Convolutional Kernel Transform

The RandOm Convolutional Kernel Transform (ROCKET) algorithm was developed as a general purpose algorithm for classifying time series data without requiring domain knowledge of the specific task (Dempster, Petitjean, & Webb, 2020). ROCKET consists of a single layer of 10,000 convolutional kernels coupled with a linear classifier. The basic parameters of a kernel are its length, weight, bias, dilation and padding. ROCKET randomly selects the parameter values for each kernel. Length is randomly selected from the set {7, 9, 11} with equal probability, weights are sampled from a normal distribution $N(0,1)$, bias is sampled from a uniform distribution $U(-1,1)$, dilation is sampled on an exponential scale, and when each kernel is generated, the decision is made whether or not to apply padding (Dempster, Petitjean, & Webb, 2020). Each kernel $\omega(n)$ with a dilation (d) is convolved with the input time series (segment *x(n)* of length L) as below. ROCKET computes two features from each convolution, the maximum value of the convolution and the proportion of positive values (PPV).

$$x(n) * w(n) \; = \; \sum_{m=0}^{L-1} x(n \; + \; md)w(n)$$

$$PPV = \frac{1}{N}\sum_{m=0}^{N-1}[(x(n) * w(n))_m + b \; > 0]$$

Where b is a bias term and $[\cdot]$ represents Iverson brackets in the PPV expression (i.e. 1 if expression is true; 0 otherwise). A linear classifier, such as logistic or ridge regression, is subsequently used to combine the 20,000 kernel features and classify the data. ROCKET has been shown to achieve state-of-the-art accuracy with minimal training time or computational expense compared to traditional deep learning methods (Dempster, Petitjean, & Webb, 2020). The real appeal of the method is that unlike conventional machine (including deep) learning methods, no feature engineering or neural-network architecture design is required. The method can be applied to any signal type and a recent study has shown its utility for neonatal EEG classification (Lundy & O'Toole, 2021). In the context of this work, ROCKET was used to 1) establish a baseline performance for a machine learning approach and 2) to quickly and simply evaluate pre-processing routines for enhanced performance.

2.2.3. Deep learning architectures

During initial development work, several well-established deep learning architectures were implemented and tested for potential models for artefact detection. Within the EEG domain, there are many open-source libraries available for EEG classification, which contain commonly used architectures, such as ResNet and DenseNet (Shuyue, n.d.). A deep residual network was developed by others for the task of neonatal EEG artefact detection (Webb, Kauppila, Roberts, Vanhatalo, & Stevenson, 2021). On a test set, this residual network achieved an area under the characteristics operating curve (AUC) of 0.872 (Webb, Kauppila, Roberts, Vanhatalo, & Stevenson, 2021). A neonatal seizure detection method achieved state-of-the-art results using a 1-dimensional fully convolutional CNN (O'Shea, Lightbody, Boylan, & Temko, 2020). In a fully convolutional network, convolutional layers are used to both extract features and perform classification. They generally require fewer parameters and allow weight sharing, which also minimises the need for dropout layers (Dai, Li, He, & Sun, 2016). In addition, temporal position is maintained in fully convolutional networks (Song, Choi, Han, & Kim, 2018). After initial testing with different architectures, a fully convolutional network was selected as it had the best performance for our data set.

2.2.4. EEG grading algorithm

An automated EEG algorithm to grade severity of injury associated with HIE was previously developed in (Raurale, et al., 2021). This algorithm combines a quadratic time–frequency distribution with a CNN to predict the EEG grade. The CNN classifies 5-minute segments of single-channel EEG. Segment length was chosen after evaluation with different lengths (Raurale, Boylan, Lightbody, & O'Toole, 2020). An EEG grade over the multi-channel, 1-hour EEG file is then predicted using majority voting over the multiple 5-minute segments. The CNN was trained on a dataset containing 54 hours of EEG, achieving a leave-one-out (LOO) cross validation accuracy of 88.9%. When tested on an unseen dataset containing 338 hours of data, it achieved 69.5% accuracy (Raurale, et al., 2021).

In this work, a new CNN model for this EEG grading algorithm is developed by training this architecture on the 653-hour database outlined in Table 2. Subsequently, the artefact detection algorithm developed in this work is used in a post-processing routine to omit artefactual segments. Grading performance is compared with and without artefact detection in post-processing.

3. Methods

The metrics used to assess the performance of the classifiers for both artefact detection and HIE grading include AUC, F1 score, Matthews Correlation Coefficient (MCC), sensitivity, and specificity. The artefact detection classifiers were trained to provide binary classification (artefact or not) but metrics are also reported for its performance on each individual type of artefact. A Dell PowerEdge C4130 with four Nvidia V100 graphics processing units (GPUs) was used in this work. The GPUs were used for training the CNNs.

3.1. NEURAL implementation

The NEURAL artefact detection algorithm was used to establish a baseline performance for artefact detection using digital signal processing methods and engineered thresholds alone. As there was no training involved, the entire artefact dataset was used in testing. The annotated labels versus predicted labels were used to compute performance.

3.2. ROCKET and pre-processing routine optimisation

The overwhelming majority of spectral power in neonatal EEG is below 30 Hz. Low frequency components, such as <0.5 Hz, can dominate the EEG trace making it more difficult to analyse the other components. Therefore, it is common practice to implement a high-pass filter to remove low-frequency activity and DC shifts. A low-pass filter is also implemented to avoid aliasing before down-sampling the EEG. In addition, a suitable epoch length must be selected before analysis. The optimal values for these parameters are dependent on the task at hand. The different types of artefacts annotated in the EEG data vary significantly with respect to their duration, amplitude and frequency content. A wide range of filter cut-off frequencies and epoch lengths were evaluated using the ROCKET algorithm to evaluate the performance of different combinations of pre-processing parameters. This included high-pass cut-off frequencies ∈ {0.5Hz, 1Hz, and 2Hz}, low-pass cut-off frequencies ∈ {10Hz, 20Hz, 30Hz, and 40Hz}, epoch lengths ∈ {2s, 4s, 6s, and 8s}, and sampling-rates ∈ {25Hz, 50Hz, 100Hz, and 200Hz}. Sampling rates and low-pass filter cut off frequencies were paired to preserve the Nyquist–Shannon sampling theorem.

A ridge regression classifier was used with a 10-point log array from 0.001 to 1000 as the regularisation parameter, as recommended in (Dempster, Petitjean, & Webb, 2020). A 5-fold CV routine was used, whereby for each fold 80% of the data was used to train the model and 20% was used to test the model. During training on 80% of data, the classes were balanced by randomly removing epochs from the artefact-free class. During testing on 20% of data, the classes were left as unbalanced to represent the real balance of classes in clinical practice. The best performing pre-processing parameters were subsequently tested in a subject-independent LOO routine.

3.3. CNN and hyper-parameter optimisation

After establishing baseline performances using NEURAL and ROCKET, a fully convolutional CNN-based architecture, as used in (O'Shea, Lightbody, Boylan, & Temko, 2020), was used as the reference architecture for initial development, as shown in Figure 3. For the CNN work, the artefact dataset was segmented into 8-second epochs with 0% overlap as described in Table 1. A pre-processing routine was used to filter the data and down-sample to 64Hz. The architecture includes three blocks, where each block consists of 1 or more 1-D convolutional layers with the ReLU activation function, followed by batch normalisation and average pooling. The final block includes a convolutional layer with two filters followed by pooling and a softmax layer producing the probability-like outputs of artefact and non-artefact. There are a seemingly infinite combination of possible hyper-parameters for the CNN model.

However, a selection of possible combinations of hyper-parameters were evaluated in a single fold test, whereby the models were trained on 80% of the data and tested on 20%. The training data was balanced before training. The hyper-parameters optimised using this routine were the parameters for the first three blocks: number of convolutional layers in each block (nLayer), number of filters (nFilt), kernel size (kSize), stride length (stride), pooling size (pSize), and pooling stride length (pStride). The options for each parameter were selected from the following sets: nLayer ∈ {1, 2, 3}, nFilt ∈ {8, 16, 32}, kSize ∈ {3, 5, 7}, stride ∈ {1, 2}, pSize ∈ {4, 8}, and pStride ∈ {2, 4}. From these sets, there were 110 valid combinations of the variations of parameters.

Several stochastic gradient descent optimisers available in Keras (i.e. SGM, RMSProp, and Adam) were considered. Based on initial experimentation and analysis of learning curves, the Adam optimiser algorithm was used with a learning rate of 0.0009. Categorical cross entropy was used to compute the loss between the labels and predictions. Early stopping criteria was used to monitor the loss on the 20% test data and to stop the training if improvement was not achieved in over 400 iterations. Model checkpoint was subsequently used to then select the model with the best accuracy on the 20% test data. The learning curves were observed to ensure the above criteria were appropriate. Accuracy on the test data was used to rank the performance of the models and hyper-parameter sets. The five best performing models were subsequently assessed in a LOO routine to determine the best model for artefact detection.

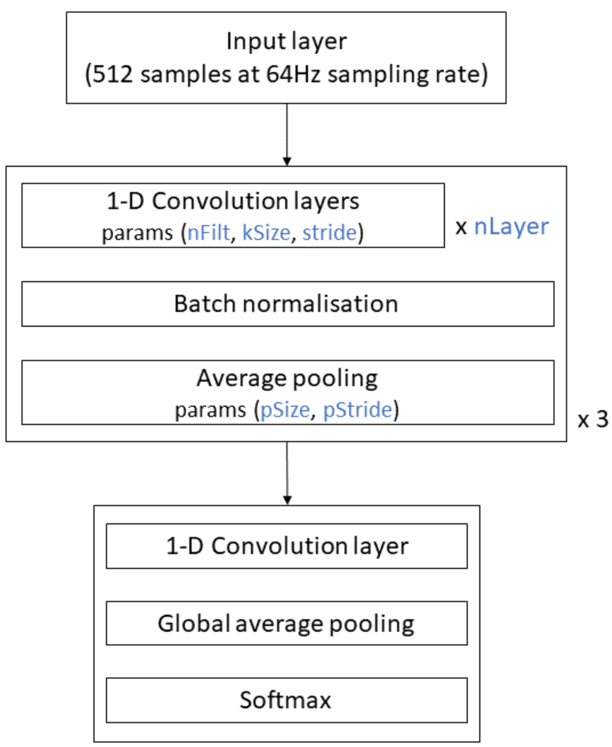

Figure 3. CNN architecture.

3.4. EEG grading algorithm

Based on the algorithm in (Raurale, et al., 2021), a new CNN model for EEG–HIE grading was developed by training the CNN on the dataset containing 653 hours of EEG, as outlined in Table 2. The pre-processing and CNN architecture are the same as before but with one change: the final classification layer (softmax) is replaced with a regression layer. This change was implemented to have an ordinal output, as opposed to categorical classes, which makes better use of the EEG information and provides continuous scale at output which could be used as an EEG score. Thus, the CNN is trained on 4 EEG grades represented as continuous values from the set {1, 2, 3, 4}. The grading algorithm is trained and tested using a 10-fold CV cross validation across the 181 neonates in the data set. The following post-processing routine determines the grade over the 1-hour epoch: 1) for each channel, the median value of the output from the CNN (the output of the regression layer) is calculated over all 5-minute segments per channel, 2) the median of this value is calculated over all channels, and then 3) rounded to the nearest integers. Estimates are bounded to be within [1,4].

3.5. Artefact detection as a post-processing routine for automated EEG grading

Following training and validation of the various artefact detection algorithms, the best performing model across the different methods was retrained on the full artefact detection dataset, creating one single model file. This model was then used in a post-processing routine for automated EEG grading using the newly trained model as per Section 3.4. The database of 653 1-hour epochs from 181 neonates, as described in Table 2, was used to test the automated grading algorithm with, and without, artefact detection in the post-processing routine. As described in Section 3.4, the algorithm makes a decision on 5 minute segments per channel. The best performing artefact detection model and pre-processing routine was used to predict the artefact probability in the 5-minute segment. If this artefact probability exceeded a certain threshold, then the 5-minute segment was removed from the median computation step in the HIE grading algorithm's post-processing. Threshold values ranging from 0.1 to 0.5 were assessed for removing the 5-minute segments. Threshold values of less than 0.1 were not considered as it would remove 5-minute segments with less than 10% artefact, which, in practice, was too low as it removed most of the 5-minute segments within the 1-hour epoch and therefore excluded the this epoch from further analysis. If the majority of segments (i.e. 50% of segments) in a single channel were classified as artefact, the entire channel was removed. Similarly, if the majority of segments (i.e. 50% of segments) across all channels at the same time point were classified as artefact, all channels at that time segment were removed. The testing accuracy within the 10-fold CV of the EEG-HIE grading algorithm was subsequently compared with and without artefact detection in the post-processing stage.

4. Results

4.1. ROCKET and pre-processing optimisation

Pre-processing parameters were assessed using the ROCKET classifier. Three parameters (filter cut-offs, sampling frequency, and epoch length) were altered resulting in 128 combinations that were evaluated, each in a 5-fold CV routine. Table 3 presents the 10 (from 128) best performing pre-processing parameter sets. Generally, the results show that greater temporal resolution, frequency bandwidth and epoch length yielded highest classification performance. The second-best performing parameter set was selected as the sampling frequency of 100 Hz offers better computational efficiency with only a very slight drop in performance (AUC of 0.897 versus 0.895). A bandpass filter of 0.5 to 40 Hz on an 8-second segment, with a sample frequency of 100 Hz, was used for all further analysis.

4.2. Artefact CNN and hyper-parameter optimisation

Table 4 presents the best 10 (out of 110) hyper-parameter sets based on the accuracy on a 1-fold test. The best performing model contains 10 convolutional layers in total, resulting in an accuracy of 95.16% and a loss of 0.1475. The second best performing model based on validation accuracy, which contains 7 convolutional layers, achieves a similar accuracy of 95.03% and a lower loss of 0.1475. The best 5 hyper-parameter sets from Table 4 were subsequently evaluated in a more robust test, using a LOO instead of 5-fold cross-validation, as presented in Table 5. The best performing model from the single split test in Table 4 is again the best performing model in LOO test, achieving 87% accuracy and 0.84 AUC. The CNN model with the most parameters is not the best performing model.

High-pass (Hz)	*Low-pass (Hz)*	*Sampling-rate (Hz)*	*Epoch length (seconds)*	*AUC (mean [max, min])*
0.5	40	200	8	0.8969 [0.8928, 0.9023]
0.5	**40**	**100**	**8**	**0.8952 [0.8897, 0.9019]**
0.5	40	200	6	0.8939 [0.8826, 0.9024]
0.5	40	100	6	0.8910 [0.8789, 0.9012]
0.5	30	200	8	0.8860 [0.8833, 0.8907]
0.5	30	100	8	0.8859 [0.8729, 0.8935]
0.5	40	200	4	0.8855 [0.8799, 0.8902]
0.5	30	200	6	0.8836 [0.8768, 0.8874]
0.5	30	100	6	0.8814 [0.8726, 0.8911]
0.5	40	100	4	0.8794 [0.8758, 0.8822]

Table 3. Top 10 set of pre-processing parameters selected using ROCKET in a 5-fold cross validation. The full set of parameters included is high-pass {0.5 to 2 Hz}, low-pass {10 to 40 Hz}, sampling-rate {25 to 200 Hz}, and epoch duration {2 to 8 seconds}.

Model no.	*nLayer*	*nFilt*	*sKern*	*stride*	*sPool*	*sPStride*	*Train loss*	*Train accuracy*	*Val loss*	*Val accuracy*
1	3	32	3	1	8	4	0.048	0.979	0.160	0.952
2	2	32	7	1	8	2	0.048	0.981	0.148	0.950
3	3	32	5	1	4	4	0.033	0.985	0.239	0.950
4	3	16	3	1	8	4	0.085	0.966	0.145	0.950
5	2	16	3	1	8	4	0.103	0.960	0.142	0.950
6	2	32	7	1	4	4	0.029	0.987	0.157	0.950
7	3	32	3	1	8	2	0.059	0.978	0.159	0.950
8	2	16	7	1	4	4	0.047	0.979	0.155	0.949
9	2	32	5	1	8	4	0.040	0.982	0.157	0.949
10	3	16	3	1	8	2	0.099	0.964	0.145	0.949

Table 4. CNN hyper-parameter optimisation 1-fold results.

Model no.	*nLayer*	*nFilt*	*sKern*	*stride*	*sPool*	*sPStride*	*No. of params*	*AUC*	*Accuracy*
1	3	32	3	1	8	4	25,538	0.844	0.872
2	2	32	7	1	8	2	36,834	0.779	0.759
3	3	32	5	1	4	4	41,986	0.834	0.855
4	3	16	3	1	8	4	6,626	0.823	0.822
5	2	16	3	1	8	4	4,274	0.682	0.632

Table 5. CNN hyper-parameter optimisation LOO results.

4.3. Comparison of artefact detection algorithms

Table 6 present a comparison between the NEURAL algorithm on the full dataset (as no training was required), the best ROCKET model and the best CNN model, both for testing LOO. The CNN achieved the best performance with 0.844 AUC. The NEURAL algorithm achieved high specificity (0.97) at the expense of low sensitivity (0.4), suggesting that its engineered thresholds for determining artefact could be improved. As expected, the method performed well on artefacts such as movement and poor electrode contact but performed poorly on more complex artefacts, such as ECG and respiration, which were not considered in the design of the method. Considering that the ROCKET requires zero design and parameter tuning, its overall performance of 0.815 AUC is relatively high.

4.4. Artefact detection as a post-processing routine for automated HIE grading

The best model from Section 4.3 was used to provide artefact probabilities for each file in the HIE graded dataset in Table 2. Examples of 8-second epochs from the HIE graded dataset and their probabilistic outputs from the artefact detection CNN are shown in Figure 4.

The HIE grading algorithm detailed in Section 3.4 was trained on the large 653-hour dataset detailed in Table 2. The confusion matrix for the classifier's HIE predictions is shown in Table 7. The classifier achieved an overall accuracy of 82.8% in a 10-fold LOO routine, as shown in Table 8.

Artefact type	*NEURAL*				*ROCKET*				*CNN*			
	AUC	**MCC**	**Sens**	**Spec**	**AUC**	**MCC**	**Sens**	**Spec**	**AUC**	**MCC**	**Sens**	**Spec**
All	**0.685**	**0.478**	**0.400**	**0.969**	**0.815**	**0.614**	**0.726**	**0.904**	**0.844**	**0.649**	**0.794**	**0.894**
Poor electrode	0.799	0.554	0.628		0.853	0.506	0.803		0.901	0.550	0.908	
Movement	0.865	0.648	0.761		0.925	0.565	0.946		0.892	0.243	0.890	
Muscle	0.539	0.045	0.108		0.747	0.162	0.590		0.722	0.292	0.550	
Sweat	0.572	0.143	0.175		0.808	0.361	0.713		0.913	0.536	0.932	
Dummy sucking	0.505	0.002	0.042		0.744	0.061	0.583		0.822	0.316	0.750	
ECG	0.495	-0.008	0.020		0.706	0.215	0.508		0.799	0.339	0.705	
Respiration	0.546	0.111	0.123		0.756	0.344	0.608		0.947	0.107	1.000	
Pulsatile	0.703	0.126	0.438		0.482	-0.010	0.061		0.824	0.173	0.754	

Table 6. Performance metrics of NEURAL, ROCKET, and CNN artefact detection in LOO routine.

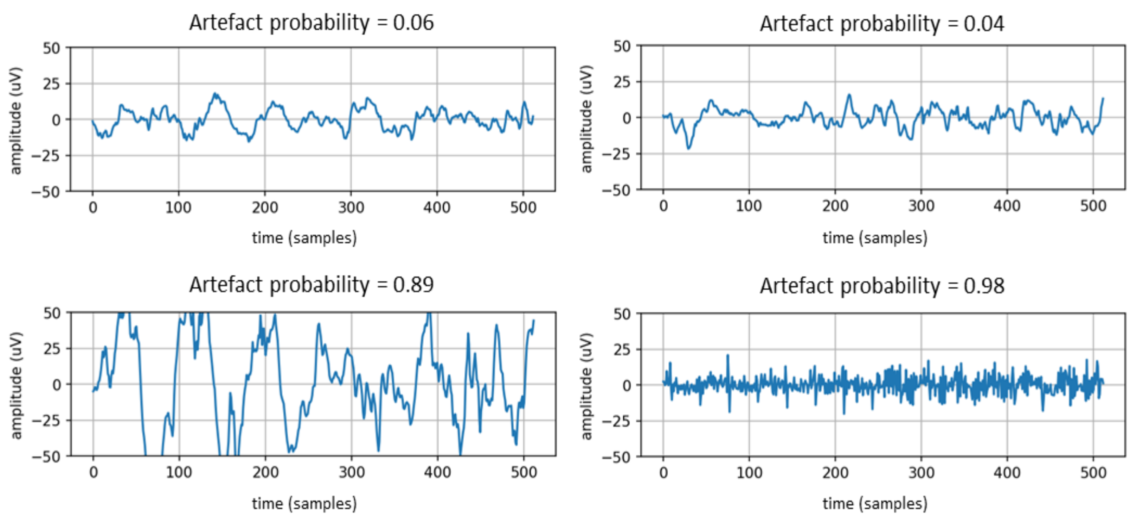

Figure 4. Examples of 8-second EEG epochs with probabilistic outputs of the artefact detection CNN.

The methods described in Section 3.5 were used to omit 5-minute segments from the median computation step in the HIE grading algorithm's post-processing. A threshold of 0.25 was used to classify segments as artefacts and omit segments. On average, in each 1-hour window, 19.65% of 5-minute segments were detected as artefact and omitted. There were 12 files out of the 653 1-hour epochs in which the artefact detector omitted all 5-minute segments, and thus, a grade could not be computed. Using the remaining 641 files, the HIE grading accuracy with the incorporated artefact detection post-processing routine was 83.6%, a 0.8% increase in accuracy, as shown in Table 8. In addition, the average artefact probability across the full 1-hour, 8-channel files for both correct and incorrect EEG grades were computed. Using 1,000 iterations of a bootstrap procedure sampled on a baby rather than epoch level, the average percentage of artefact for the correct classifications, 11.76% (95% CI: 9.37 to 13.72%), did not significantly differ to the percentage of artefact for the incorrect classifications, 12.28% (95% CI: 11.13 to 13.43%), p=0.568. There is not a statistically significant difference between the amount of artefact in correct versus incorrect grade predictions.

a) *Predictions* (rows: *Labels*)

Grade	*1*	*2*	*3*	*4*
1	318	35	0	0
2	41	122	3	0
3	0	11	50	5
4	0	1	16	51

b) *Predictions* (rows: *Labels*)

Grade	*1*	*2*	*3*	*4*
1	309	33	1	0
2	38	124	3	0
3	0	9	51	5
4	0	0	16	52

Table 7. Confusion matrix of the EEG grading classifier a) without and b) with artefact detection.

		Grade 1	*Grade 2*	*Grade 3*	*Grade 4*	*Overall accuracy (95% C.I.)*	*Overall MCC*
Without art. detection	Recall	0.901	0.735	0.758	0.750	0.828 (0.805 to 0.852)	0.722
	Precision	0.886	0.722	0.725	0.911		
With art. detection	Recall	0.875	0.747	0.773	0.765	0.836 (0.814 to 0.862)	0.737
	Precision	0.890	0.747	0.718	0.912		

Table 8. EEG grading performance with and without artefact detection.

The EEG grade was only changed in 9 cases, and an additional 12 cases were determined to be entirely artefact. The EEG grading algorithm was developed to compute the median grade across all channels and 5-minute segments to obtain the final predicted grade for a 1-hour window of EEG. Figure 5 shows the performance of the grading algorithm when the median is taken across shorter window lengths from 5 minutes to 60 minutes. The results show that the grading algorithm performs relatively well even at shorter durations. An accuracy of 80% is achieved in all window lengths greater than 10 minutes. The inclusion of the artefact detection algorithm in its postprocessing marginally improves performance across all window durations greater than 25 minutes.

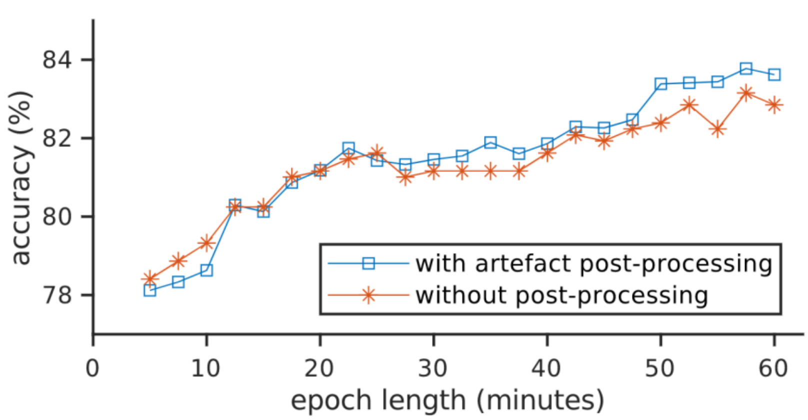

Figure 5. EEG grading algorithm accuracy with and without artefact detection using various window lengths.

5. Discussion

This paper presents the development of neonatal EEG artefact detection algorithms using DSP, ridge classifier and deep learning methods. The CNN was the best performing model, achieving an AUC of 0.844 for artefact detection in neonatal EEG. The detector is sensitive to many types of artefacts including multiple types of biological artefacts, which are often the most problematic for human interpretation. An EEG-HIE grading algorithm was previously developed on a small dataset containing 54 hours of EEG, achieving 69.5% accuracy on an unseen test set (Raurale, et al., 2021). In this paper, a new CNN model for HIE grading was developed using a large, multi-centre dataset containing 653 hours of EEG. Although cross-validation testing results will often be inflated compared to testing results on a completely unseen validation dataset, the increase in performance was large: an increase in accuracy from 69.5% (95% CI: 65.3 to 73.6) to 83.6% (95% CI: 81.4 to 86.2). The use of the artefact detection algorithm showed that the misclassifications of this new EEG-HIE grading algorithm are not significantly associated with artefact, indicating that algorithm performance is not sensitive to artefacts. Consistent with this finding, combining the artefact detection algorithm in a post-processing routine for grading resulted in only a marginal increase in overall classifier performance by 0.8%, an increase too small to be considered of practical significance.

5.1. Artefact detection classifier development

The NEURAL method provides a high specificity of 0.969 but a low sensitivity of 0.40. Unsurprisingly, it achieves highest performance on movement artefacts, which are typically very high amplitude artefacts and simply removed by applying an amplitude threshold. However, this method is not able to identify physiological artefacts, such as ECG and respiration, as it is not designed to include these, and hence the relatively lower performance compared to the other two methods.

The ROCKET algorithm offers a rapid prototyping methodology which does not require deep learning expertise to design the network architecture or to train and deploy a model. The kernel parameters and their weights and biases are randomly generated. The only hyper-parameter that requires tuning is the regularisation parameter of the ridge regression. Without design, manual tuning or optimisation, the ROCKET model achieved 0.815 AUC in a LOO routine, taking on average 8.9 minutes on a CPU to complete a single iteration of LOO. In comparison, the CNN took 30.7 minutes on a GPU to complete a single LOO iteration.

One of the main tasks in developing a machine learning algorithm, is first developing a suitable pre-processing routine. ROCKET was used in this work to optimise the parameters of the time series pre-processing routine, including setting the sampling rate, filtering, and epoch length. The EEG data was recorded at sampling frequencies between 200 and 256 Hz. The majority of pertinent information in EEG is sub-50 Hz. It has been shown that dominant frequencies of neonatal EEG seizures are in the 0.5-6 Hz range (Kitayama, et al., 2003), for example. Artefacts, such as respiration and pulsatile, are slow-moving waveforms similar to seizures. However, muscle artefact is predominantly over 15 Hz (Muthukumaraswamy, 2013). Therefore, a validation routine was used to select the optimal pre-processing routine. The results in Table 3 suggest that the ROCKET algorithm performs best with as much temporal and spectral information as possible. The largest sampling rate (200 Hz), widest filter bandwidth (0.5-40 Hz) and longest epoch length (8 seconds) achieved the best performance in 5-fold CV tests with 0.897 AUC. As the upper limit of the filter was 40Hz, reducing the sampling rate to 100Hz was a logical decision, as there is negligible trade-off in performance (0.895 AUC) and it reduces oversampling making it less computationally and memory intensive during training.

Unlike ROCKET, CNN development involves the tuning and optimisation of an almost infinite number of hyper-parameters and design choices. Testing each network architecture and tuning each parameter in an automated routine is therefore not feasible. There are several well-known convolutional networks including AlexNet, VGGNet, and ResNet (similar to that used previously for artefact detection in (Webb, Kauppila, Roberts, Vanhatalo, & Stevenson, 2021)). In this work, following a literature review and preliminary experimentation with commonly used models, a fully-convolutional architecture was selected as it is quicker to train and requires fewer parameters compared to networks with fully connected layers. Several hyper-parameters, including the number of convolutional layers, number of filters, kernel size, stride length, pooling size, and pooling stride length were then optimised in an automated routine. Over 110 sets of hyper-parameter combinations were assessed in the same single fold test, as presented in Table 4. As presented in Table 5, the best five models were tested in a LOO routine. Generally, the results show that deeper networks tend to achieve higher performance, as the models with 10 convolutional layers achieve highest performance. However, there is risk of overtraining, as the two models with the most parameters were not the best performing models in the LOO routine. The CNN performed particularly well on the biological artefacts, achieving the highest AUC on respiration artefact. The average AUC across the biological artefact classes is 0.856, while the average AUC on non-biological artefact classes in 0.850. According to (Mathieson, et al., 2016), respiration artefacts are the biggest contributor to false detections in neonatal seizure detection, and as such, that is a considerable strength of the proposed artefact detection CNN in this work.

The residual CNN presented in (Webb, Kauppila, Roberts, Vanhatalo, & Stevenson, 2021) was trained on a dataset containing 112 hours of data with 5 types of artefact and achieved 0.872 AUC. Our CNN was trained on a considerably smaller dataset of 19 hours of EEG with 8 types of artefact, and achieved 0.844 AUC. The artefact detector in this work was not designed to provide multi-class classification, unlike in (Webb, Kauppila, Roberts, Vanhatalo, & Stevenson, 2021). Our binary classifier was designed to distinguish between artefact and non-artefact, as our goal was to generate a signal quality measure to be used in parallel to diagnostic decision support algorithms. Adding residual connections to a deep neural network typically allows for deeper networks, which could potentially yield improved feature extraction and classification performance if supported also by an increase in the size of the training set. Further improvements to the CNN in this work may be achieved by investigating residual connections, in addition to exploring other promising architectures, such as ConvNeXt (Liu, et al., 2022).

The best model in Table 5 was subsequently taken forward as the model deployed for use in the HIE application. It achieved 0.844 AUC and 87.2% accuracy in the LOO routine. It is considerably better than the performance of NEURAL of 0.685 AUC. The ROCKET algorithm, however, performed comparatively well, despite requiring minimal knowledge of the classification task or deep learning theory in general. If nothing else, the ROCKET algorithm serves as a valuable method for establishing baseline performance with minimal training time or effort. Development of fine-tuned models can thus be completed with this baseline performance in mind.

5.2. Artefact detection as a post-processing routine for automated HIE grading

The average percentage of detected artefact in misclassifications is not significantly different to that in correct classifications. This indicates that there is minimal association between EEG grade classification quantity of artefact, and thus, that the EEG grading algorithm is somewhat robust to artefacts.

Using the artefact detection algorithm as a post-processing routine to the EEG grading algorithm resulted in a slight increase in performance at the 1-hour level. The accuracy of the grading algorithm without and with artefact detection was 82.8% and 83.6%, respectively; the MCC also increased from 0.722 to 0.737, respectively. Nonetheless, this increase in performance is minimal and likely not of practical significance considering the range of variability of CNN performance due to random initialisations of parameters during development, in addition to the performance linked specifically to our dataset. Retrospectively reviewing the epochs rejected by the artefact detection algorithm indicates that the artefact detection algorithm is correctly identifying and removing artefactual epochs, as the example in Figure 4 shows. On average, 19.65% of the 5-minute segments within the 1-hour windows were detected as artefact and removed from the final EEG grading step. This resulted in a change of the 1-hour HIE grade predictions in only 3.7% of cases.

During the grading of the HIE database, EEG epochs from each infant were selected at approximately 6, 12, 24, 36 and 48 hours after birth. Within reason, the neurophysiologist selected an hour epoch that was close to the correct time point but of relatively good quality and suitable for grading, in their view. As such, the HIE dataset may be somewhat of higher quality (less artefacts) than continuous long-term EEG. Based on the artefact percentage results in Section 4, between 11.76% to 12.28% of the HIE epochs were artefacts. Future research will include an analysis of the HIE grading algorithm with artefact detection on continuous, long-term EEG which has not been carefully pruned by a human expert. The inclusion of artefact detection in the automated grading algorithm may be of greater significance for grading EEG from long-term recordings.

The human expert deemed all 653 1-hour epochs to be of sufficient quality for HIE grading. It is therefore a potential inaccuracy of the artefact detection algorithm and the post-processing routine that in 12 cases to be classed as artefact and therefore unable to provide a grade.

The median voting step which takes the median across all 5-min segments and across all available channels to provide the grade for the 1-hour epoch may reduce the potential of the overall 1-hour grade being impacted by artefacts. It is unlikely that, in this dataset, there are many artefacts which span both the majority of time points in the 1-hour, as well as the majority of channels in the EEG. As the dataset only contains grades at the 1-hour level, it was not possible to assess accuracy improvements at the 5-minute level. However, as shown in Figure 5, the grading is relatively robust to shorter window lengths, with an accuracy of over 80% when using window lengths of greater than 10 minutes. A shorter window length may provide a relatively accurate substitute for the 1-hour epoch, which may be advantageous when the duration of EEG monitoring is time constrained, such as with EEG monitoring not long after birth.

With regard to clinical usefulness of the EEG grading algorithm, the artefact detection CNN and routine may have a significant role to play. The artefact detection CNN provides additional decision support by informing the medical team of the quality of the EEG data that both they and the HIE grading algorithm are interpreting. Clinical deployment of the HIE grading algorithm will require a continuous output of estimated HIE grades at the 5-minute segment level, as opposed to providing a predicted grade for each hour alone. The artefact detection model may therefore be useful in shorter windows, providing a confidence score alongside the grading algorithm at the 5-minute segment level.

Prior art in neonatal seizure detection indicates that artefacts are the largest contributor to incorrect classifications of seizures (Mathieson, et al., 2016). The algorithm used in that work was an SVM-based algorithm. The results in this paper indicate that misclassifications with the CNN-based automated EEG grading algorithm are not related to higher artefact percentage. This is perhaps because HIE is a global injury and by using the median across channels and by averaging over a long period of time (1-hour windows), there is a degree of protection from artefacts using the HIE grading algorithm. Seizures, by contrast, can be localised to specific channels and are detected in much shorter periods of time (i.e. 8-second windows). As such, it is not surprising that artefacts are a significant cause of misclassification of seizures, but not as relevant for long-term HIE grading. It is also possible that the robustness of the HIE grading to artefacts is due to the design of HIE grading algorithm, as the TFD input to the CNN may offer robustness to artefacts.

Future work will look at testing the artefact detection post-processing routine for HIE grading on continuous, long-duration EEG, as opposed to pruned 1-hour epochs selected here. The artefact detection algorithm will also be tested in other neonatal EEG applications, such as seizure detection. It is also worth investigating if the signal quality scores for each segment could be used as an input to the grading CNN. Prior art has shown that the inclusion of data quality metrics as an input to a multimodal CNN alongside the time series data can improve learning (Petrozziello, Redman, Papageorghiou, Jordanov, & Georgieva, 2019). Improvement in automated grading performance may be achievable by using the artefact detection algorithm in this manner.

6. Conclusion

We found that a fully convolutional neural network was a suitable and accurate architecture for the task of artefact detection in neonatal EEG. We also found ROCKET to be a quick and easy machine-learning implementation, which in many applications may provide suitable performance without needing in-depth knowledge of deep learning models.

The EEG-HIE grading algorithm was trained on a significantly larger dataset in this work, achieving 82.8% in a 10-fold CV. The misclassifications of the algorithm were not influenced by the presence of artefacts and combining the artefact detection algorithm in a post-processing routine provided only a marginal increase in classification performance to 83.6%. The artefact detection algorithm’s value in clinical translation and implementation is important to ensure the medical team have confidence in estimating the EEG grades associated with HIE. We found that at the 5-minute segment level, the artefact detection algorithm correctly omits artefactual segments from the HIE grading decision, and as such, provides clinicians with a greater degree of confidence in the HIE grading algorithm’s outputs.

The artefact detection outputs may be used as a continuous score of EEG quality alongside the predicted EEG grades to provide a more informative decision support system.

7. Funding

This study was supported by a Strategic Translational Award and an Innovator Award from the Wellcome Trust (098983 & 209325) and a Research Infrastructure grant from Science Foundation Ireland (15/RI/3239).

8. Acknowledgements

The authors acknowledge the hospitals and personnel who coordinated, supervised and collected data as part of the ANSeR Consortium:

INFANT Research Centre and Department of Paediatrics and Child Health, University College Cork, Cork, Ireland (*Andreea M Pavel, MD; Eugene Dempsey, MD; Vicki Livingstone, PhD; Elena Pavlidis, MD; Liudmila Kharoshankaya, MD; Sean R Mathieson, PhD; Liam Marnane PhD; Gordon Lightbody PhD; Jackie O'Leary MSc; Mairead Murray MSc; Jean Conway MSc; Denis Dwyer BSc; Andrey Temko, PhD; Taragh Kiely, MSc; Anthony C Ryan, MD; Geraldine B Boylan PhD*). Institute for Women's Health, University College London, London, UK (*Janet M Rennie, MD; Subhabrata Mitra, PhD*). Utrecht Brain Center, University Medical Center Utrecht, Utrecht University, The Netherlands (*Linda S de Vries, PhD; Lauren C Weeke, PhD; Mona C Toet, PhD; Johanneke C Harteman PhD*). Department of Neonatal Medicine and Division of Paediatrics, Department CLINTEC, Karolinska University Hospital, Karolinska Institutet, Stockholm, Sweden (*Mikael Finder, PhD; Mats Blennow PhD; Ingela Edqvis, MD*). Rotunda Hospital, Dublin, Ireland (*Adrienne Foran MD; Raga Mallika Pinnamaneni, MD; Jessica Colby-Milley, MSc*). Royal London Hospital and Queen Mary University of London, London, UK (*Divyen K Shah, MD; Nicola Openshaw-Lawrence*). Department of Clinical Neurophysiology, Great Ormond Street Hospital for Children NHS Trust, London, UK (*Ronit M Pressler, PhD*). Homerton University Hospital NHS Foundation Trust, London, UK (*Olga Kapellou, MD*). Clinical Neurophysiology, University Medical Center Utrecht, Utrecht, The Netherlands (*Alexander C van Huffelen, PhD*).

9. References

Abbasi, H., & Unsworth, C. P. (2020). Applications of advanced signal processing and machine learning in the neonatal hypoxic-ischemic electroencephalography. *Neural Regeneration Research*, *15*(2), 222-231.

Abdur Rahman, M., Abul Hossain, M., Raihan Kabir, M., Hossain Sani, M., Al-Mamun, A., & Abdul Awal, M. (2019). Optimization of Sleep Stage Classification using Single-Channel EEG Signals. *International Conference on Electrical Information and Communication Technology (EICT).* Khulna, Bangladesh.

Ahmed, R., Temko, A., Marnane, W., Lightbody, G., & Boylan, G. (2016). Grading hypoxic–ischemic encephalopathy severity in neonatal EEG using GMM supervectors and the support vector machine. *Clinical Neurophysiology*, *127*(1), 297-309.

Alvarez, V., & Rossetti, A. O. (2015). Clinical use of EEG in the ICU: technical setting. *Journal of Clinical Neurophysiology*, *32*(6), 481-485.

Dai, J., Li, Y., He, K., & Sun, J. (2016). R-FCN: Object Detection via Region-based Fully Convolutional Networks. *Advances in Neural Information Processing Systems, 29*, 379-397.

Davies, A., Wassink, G., Bennet, L., Gunn, A. J., & Davidson, J. O. (2019). Can we further optimize therapeutic hypothermia for hypoxic-ischemic encephalopathy? *Neural regeneration research*, *14*(10), 1678-1683.

De Vos, M., Deburchgraeve, W., Cherian, P. J., Matic, V., Swarte, R. M., Govaert, P., . . . Van Huffel, S. (2011). Automated artifact removal as preprocessing refines neonatal seizure detection. *Clinical Neurophysiology*, *122*(12), 2345-2354.

Dempster, A., Petitjean, F., & Webb, G. I. (2020). ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, *34*, 1454-1495.

Garvey, A. A., Pavel, A. M., O'Toole, J. M., Walsh, B. H., Korotchikova, I., Livingstone, V., . . . Boylan, G. B. (2021). Multichannel EEG abnormalities during the first 6 hours in infants with mild hypoxic–ischaemic encephalopathy. *Pediatric Research*, *90*, 117-124.

Gluckman, P. D., Wyatt, J. S., Azzopardi, D., Ballard, R., Edwards, A. D., Ferriero, D. M., . . . Gunn, A. J. (2005). Selective head cooling with mild systemic hypothermia after neonatal encephalopathy: multicentre randomised trial. *The Lancet*, 663-670.

J, G., Cheng, X., & Wu, D. (2020). Grading Method for Hypoxic-Ischemic Encephalopathy Based on Neonatal EEG. *Computer Modeling in Engineering & Sciences, 122*(2), 721-741.

Khatwani, M., Hosseini, M., Paneliya, H., Mohsenin, T., Hairston, D. W., & Waytowich, N. (2018). Energy Efficient Convolutional Neural Networks for EEG Artifact Detection. *IEEE Biomedical Circuits and Systems Conference (BioCAS).* Cleveland.

Khlif, M. S., Mesbah, M., Boashash, B., & Colditz, P. (2010). Influence of EEG artifacts on detecting neonatal seizure. *10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010).* Kuala Lumpur, Malaysia.

Kitayama, M., Otsubo, H., Parvez, S., Lodha, A., Ying, E., Parvez, B., . . . Snead, O. C. (2003). Wavelet analysis for neonatal electroencephalographic seizures. *Pediatric Neurology, 29*(4), 326-333.

Kurinczuk, J. J., White-Koning, M., & Badawi, N. (2010). Epidemiology of neonatal encephalopathy and hypoxic–ischaemic encephalopathy. *Early Human Development, 86*(6), 329-338.

Lee, A. C., Kozuki, N., Blencowe, H., Vos, T., Bahalim, A., Darmstadt, G. L., . . . Lawn, J. E. (2013). Intrapartum-related neonatal encephalopathy incidence and impairment at regional and global levels for 2010 with trends from 1990. *Pediatric Research, 74*(1), 50-72.

Liu, Z., Mao, H., Wu, C., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 11976-11986).

Lundy, C., & O'Toole, J. M. (2021). Random Convolution Kernels with Multi-Scale Decomposition for Preterm EEG Inter-burst Detection. *European Signal Processing Conference (EUSIPCO).* Dublin, Ireland.

Mannan, M. M., Kamran, M. A., & Jeong, M. Y. (2018). Identification and Removal of Physiological Artifacts From Electroencephalogram Signals: A Review. *IEEE Access, 6*, 30630-30652.

Mathieson, S., Rennie, J., Livingstone, V., Temko, A., Low, E., Pressler, R. M., & Boylan, G. B. (2016). In-depth performance analysis of an EEG based neonatal seizure detection algorithm. *Clinical Neurophsiology, 127*(5).

Matic, V., Cherian, P. J., Koolen, N., Naulaers, G., Swarte, R. M., Govaert, P., . . . De Vos, M. (2014). Holistic approach for automated background EEG assessment in asphyxiated full-term infants. *Journal of Neural Engineering, 11*(6).

Moghadam, S. M., Pinchefsky, E., Tse, I., Marchi, V., Kohonen, J., Kauppila, M., . . . Vanhatalo, S. (2021). Building an Open Source Classifier for the Neonatal EEG Background: A Systematic Feature-Based Approach From Expert Scoring to Clinical Visualization. *Frontiers in human neuroscience, 15*.

Murray, D. M., Boylan, G. B., Ryan, C. A., & Connolly, S. (2009). Early EEG Findings in Hypoxic-Ischemic Encephalopathy Predict Outcomes at 2 Years. *Pediatrics, 124*(3), 459-467.

Muthukumaraswamy, S. D. (2013). High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations. *Frontiers in Human Neuroscience, 7*.

Nandy, A., Alahe, M. A., Nasim Uddin, S. M., Alam, S., Nahid, A., & Awal, M. A. (2019). Feature Extraction and Classification of EEG Signals for Seizure Detection. *International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST).* Dhaka, Bangladesh.

Nolan, H., Whelan, R., & Reilly, R. B. (2010). FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection. *Journal of Neuroscience Methods, 192*(1), 152-162.

O'Shea, A., Lightbody, G., Boylan, G., & Temko, A. (2020). Neonatal seizure detection from raw multi-channel EEG using a fully convolutional architecture. *Neural Networks, 123*, 12-25.

O'Toole, J. M., & Boylan, G. B. (2017). *NEURAL: quantitative features for newborn EEG using Matlab*. (ArXiv) Retrieved Nov 5th, 2021, from https://arxiv.org/abs/1704.05694v1

Pavel, A. M., Rennie, J. M., de Vries, L. S., Blennow, M., Foran, A., Shah, D. K., . . . Boylan, G. B. (2021). Neonatal Seizure Management: Is the Timing of Treatment Critical? *The Journal of Pediatrics*.

Petrozziello, A., Redman, C. W., Papageorghiou, A. T., Jordanov, I., & Georgieva, A. (2019). Multimodal Convolutional Neural Networks to Detect Fetal Compromise During Labor and Delivery. *IEEE Access, 7*, 112026-112036.

Rankine, L., Stevenson, N., Mesbah, M., & Boashash, B. (2007). A Nonstationary Model of Newborn EEG. *IEEE Transactions on Biomedical Engineering*, 19-28.

Raurale, S. A., Boylan, G. B., Lightbody, G., & O'Toole, J. M. (2020). Grading the severity of hypoxic-ischemic encephalopathy in newborn EEG using a convolutional neural network. *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC).* Montreal.

Raurale, S. A., Boylan, G. B., Mathieson, S. R., Marnane, W. P., Lightbody, G., & O'Toole, J. M. (2021). Grading hypoxic-ischemic encephalopathy in neonatal EEG with convolutional neural networks and quadratic time–frequency distributions. *Journal of Neural Engineering, 18*(4).

Sai, C. Y., Mokhtar, N., Arof, H., Cumming, P., & Iwahashi, M. (2018). Automated Classification and Removal of EEG Artifacts With SVM and Wavelet-ICA. *IEEE Journal of Biomedical and Health Informatics, 22*(3), 664-670.

Shuyue, J. (n.d.). *EEG-DL*. (Github) Retrieved December 9, 2021, from https://github.com/SuperBruceJia/EEG-DL

Song, A., Choi, J., Han, Y., & Kim, Y. (2018). Change Detection in Hyperspectral Images Using Recurrent 3D Fully Convolutional Networks. *Remote Sensing, 10*(11).

Stevenson, N. J., O'Toole, J. M., Korotchikova, I., & Boylan, G. B. (2014). Artefact detection in neonatal EEG. *36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.* Chicago.

Tamburro, G., Croce, P., Zappasodi, F., & Silvia, C. (2021). Automated Detection and Removal of Cardiac and Pulse Interferences from Neonatal EEG Signals. *Sensors, 21*(19).

Webb, L., Kauppila, M., Roberts, J. A., Vanhatalo, S., & Stevenson, N. J. (2021). Automated detection of artefacts in neonatal EEG with residual neural networks. *Computer Methods and Programs in Biomedicine, 208*.

Yasoda, K., Ponmagal, R. S., Bhuvaneshwari, K. S., & Venkatachalam, K. (2020). Automatic detection and classification of EEG artifacts using fuzzy kernel SVM and wavelet ICA (WICA). *Methodologies and Application, 24*, 16011-16019.