

Title	Neonatal hypoxic-ischemic encephalopathy grading from multi-channel EEG time-series data using a fully convolutional neural network
Authors	Yu, Shuwen;Marnane, William P.;Boylan, Geraldine B.;Lightbody, Gordon
Publication date	2023-10-18
Original Citation	Yu, S., Marnane, W. P., Boylan, G. B. and Lightbody, G. (2023) 'Neonatal hypoxic-ischemic encephalopathy grading from multi-channel EEG time-series data using a fully convolutional neural network', 2023 19th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), Harbin, China, 29-31 July, pp. 1-9. doi: 10.1109/ICNC-FSKD59587.2023.10280986
Type of publication	Conference item
Link to publisher's version	10.1109/icnc-fskd59587.2023.10280986
Rights	© 2023, IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Download date	2026-10-11 16:23:23
Item downloaded from	https://hdl.handle.net/10468/15278

Neonatal Hypoxic-ischemic Encephalopathy Grading from Multi-channel EEG Time-series Data Using a Fully Convolutional Neural Network

Shuwen Yu^{*†}, William P Marnane^{*†}, Geraldine B. Boylan^{†‡}, and Gordon Lightbody^{*†}

^{*}Electrical and Electronic Engineering, University College Cork, Ireland.

[†]Infant Research Centre, Cork University Hospital, Ireland.

[‡]Pediatrics and Child Health, University College Cork, Ireland.

Abstract—A deep learning classifier is proposed for hypoxic-ischemic encephalopathy (HIE) grading in neonates. Rather than using any features, this architecture can be fed with raw EEG. Fully convolutional layers were adopted both in the feature extraction and classification blocks, which makes this architecture simpler, and deeper, but with fewer parameters. Here two large (335h and 338h respectively) multi-center neonatal continuous EEG datasets were used for training and test. The model was trained based on weak labels and channel independence. A majority vote method was used for the post-processing of the classifier results (across time and channels) to increase the robustness of the prediction. The proposed system achieved an accuracy of 86.09% (95% confidence interval: 82.41%~89.78%), an MCC of 0.7691, and an AUC of 86.23% on the large unseen test set. Two convolutional neural network architectures which utilized time-frequency distribution features were selected as the baseline as they had been developed or tested on the same datasets. A relative improvement of 23.65% in test accuracy was obtained as compared with the best baseline.

Index Terms—hypoxic-ischemic encephalopathy (HIE), EEG, fully convolutional neural network.

I. INTRODUCTION

Hypoxic-ischemic encephalopathy (HIE) is due to a lack of oxygen and blood supply to the neonatal brain around the time of birth. The severity of this injury can be assessed clinically, soon after birth, as mild, moderate or severe but clinical assessment can be subjective and also difficult to perform if the infant requires sedation. In the most severe cases, the baby may not survive the injury, and those that do survive may have permanent disabilities, such as epilepsy, cerebral palsy and learning difficulties [1]. The only treatment available for neonates with moderate and severe HIE is therapeutic hypothermia in which the neonate is cooled to 33.5 °C for 72 hours [2]. However, for this treatment to be effective, it ideally should be commenced within 6 hours after birth. Hence, an accurate assessment of the severity of encephalopathy is required soon after birth to instigate therapeutic hypothermia within this narrow window of time.

Electroencephalography (EEG) is an effective, portable, and non-invasive tool to monitor neonatal cortical activity. The severity of the HIE injury can be graded from the neonatal

EEG. However, the interpretation of EEG is not straightforward, requiring the expertise of a neurophysiologist, which may not be available in many Neonatal Intensive Care Units (NICU), especially on a 24/7 basis. An automated method is therefore attractive as a solution to provide some estimate of the severity of the HIE injury from the EEG, within the narrow 6-hour window of opportunity for the initiation of therapeutic hypothermia [3]. This could also be available as a 24/7 cotside service.

In the visual expert interpretation of HIE injury, the characteristics of the background EEG such as amplitude, frequency, discontinuity, and sleep wake cycling [4, 5] are analyzed and the resulting grade is determined. In [6, 7], some quantitative EEG features were investigated and selected from the amplitude, frequency, and continuity characteristics of the EEG to reproduce the process of visual HIE grading. Some machine learning methods have been developed to automatically grade the HIE, with the early techniques focusing on grade classification based on quantitative features which in some ways mimicked the expert visual interpretation. In particular Stevenson et al. [8] extracted the amplitude modulation and instantaneous frequency features from the time-frequency distribution extracted over short 64-second epochs. The distribution of these features over a 1-hour window provided the feature vector for a multiple linear discriminant classifier. This achieved a classification accuracy of 83% on a relatively small dataset consisting of 62 h of 8-channel EEG from 54 neonates. Ahmed et al. [9] utilized a feature set consisting of 55 basic descriptive features from the frequency, time, and information theory domains, which were extracted from short 8 seconds windows of EEG. In comparison to [8], this work utilized generic simple features and focused primarily on the use of a much more sophisticated dynamic classifier (the SVM supervector approach). This provided a grade every 80 seconds that was determined over a sequence of twenty 8-second windows (with 50% overlap). The final EEG grade for the neonate was then determined over an hour of multi-channel EEG. This approach yielded a significant improvement in grading performance over [8], with an accuracy of 87% achieved over the same small development dataset.

An alternative approach that does not depend on either finely crafted features [8] or a wide range of descriptive features [9]

The work was supported by Science Foundation Ireland (SFI) (19/FFP/6782). This study was also supported by a Strategic Translational Award and an Innovator Award from the Wellcome Trust (098983 & 209325).

is to shift from classical machine learning to deep learning, where the features can be automatically extracted from the data. In recent years, the convolutional neural network has dominated the field of computer vision. Ruarale et al. [10] extended the work of [8], where instead of manually extracting features from time-frequency analysis, they proposed that the time-frequency distribution feature map could be treated as an image that could be input to a convolutional neural network trained to classify HIE grades. Three parallel convolutional layers automatically extracted features in the direction of time, frequency, and time-frequency from a 2-dimensional time-frequency distribution feature map. The same architecture was then also used in [11] and trained with a large dataset, with the final softmax layer substituted by a regression layer (to reflect the ordinal nature of the grades). This approach achieved a classification accuracy of 82.8% on a large dataset in mismatched conditions (they were recorded at different times, centres, machines, sampling frequencies, and even protocols).

Rather than using any features of the EEG as input, some convolutional neural network architectures have been successfully applied to detect neonatal seizures directly from the raw EEG signal, with minimal preprocessing. A fully convolutional neural network (FCN) architecture was presented in [12], and successfully solved the problem of neonatal seizure detection from the raw EEG signal. An FCN architecture with residual connections was developed in [13] that yielded further improvements in the seizure detection performance. HIE grading requires a much wider receptive field than seizure detection, in order to capture the sort of changes in the EEG signal that may occur over say 80 seconds as opposed to 8 seconds for neonatal seizure detection. In [14], a convolutional neural network (CNN) architecture was used for sleep state detection with raw EEG, which achieved the highest classification accuracy both in preterm and term neonates. An initial pilot study based on a small dataset, [15] demonstrated however that good performance could be obtained using a CNN architecture for HIE grading with raw EEG, despite the large number of parameters in the fully connected layer and the relatively long window of raw EEG required.

Unlike seizure detection, where a seizure event in the EEG can be marked with a one-second resolution, each long segment (typically 1-hour) of EEG in the HIE grading problem, has only one HIE grade annotation. It is tempting therefore to address this problem as a sequence classifier using an architecture such as the Transformer which is suited to long sequence information [16]. Recently some transformer models, such as vision transformer and swin transformer, have achieved very good performance in many tasks, and have outperformed convolutional neural networks. Recently however there have been some notable improvements in convolutional networks, such as ConvNets [17], which have achieved a better performance than the swin transformer in vision tasks and which maintained the simplicity and efficiency of the convolutional network.

In this paper, we used a fully convolutional neural network with raw EEG to classify the HIE grades. Without the fully

connected layers in the classifier block, this network has much fewer parameters and also maintains the temporal position information. This network is trained on a large data set of continuous unedited multi-channel EEG from a multi-center study. The performance of the HIE grading network is then demonstrated on another large unseen and unedited multi-center dataset, under mismatched conditions.

II. METHOD

A. Dataset

A multi-centre, randomised, controlled study was performed in parallel in 4 European countries (Ireland, the UK, the Netherlands, and Sweden), from 8 neonatal intensive care units (NICU). Neonates who met the recruitment criteria were admitted, with gestational age between 36 and 44 weeks. In total, 472 neonates were recruited, but for the purpose of our analysis, only 284 infants with HIE were included.

Of the 284 HIE neonates, 35 infants were subsequently excluded, and 68 were held out for future validation. For the remaining 181 neonates, EEG epochs each lasting exactly 1 hour were extracted from the EEG records at postnatal ages of 6, 12, 24, 36 and 48 h (if available). The ANSeR1 dataset provided 338 hours of multi-channel EEG, from 91 neonates across the 6 centres. The ANSeR2 dataset, provided 315 hours of multi-channel EEG, from 90 infants across 8 centres. Compared with other EEG datasets – SIMKAP (48 subjects) for mental workload classification [18], the SEED emotion recognition task dataset (15 subjects, 45 h) [19], and the DEAP dataset (32 subject, 21.3 h) [20], our dataset is sizable.

All neonates were recorded with cEEG (Nihon Kohden Neurofax, EEG-1200, Japan or NicoletOne ICU Monitor & Xtek, Natus, USA), with a sampling frequency of either 256 Hz or 250 Hz. The electrodes were positioned according to the international 10:20 system with disposable electrodes at F3, F4, C3, C4, T3, T4, Cz, O1/P3, and O2/P4.

Each 1-hour segment of multi-channel EEG was graded by an expert neurophysiologist for the severity of the HIE injury using 5 grades: normal, mild, moderate, major, and inactive. In this work the normal and mild grades were combined as grade 1; this reduces the number of grades to 4. Fig. 1 presents clear 30-second EEG examples of different HIE grades. It can be appreciated from Fig. 2, that the ANSeR1 and ANSeR2 datasets have similar distributions of the HIE grades. The classes are however substantially unbalanced, (which reflects the distribution in a typical NICU), with Grade 1 accounting for over 50% of the annotated 1-hour long epochs.

B. Fully convolutional neural network

1) *Pre-processing*: A bipolar montage was used, with 8 EEG channels: F4-C4, C4-O2/C4-P4, F3-C3, C3-O1/C3-P3, T4-C4, C4-Cz, Cz-C3, and C3-T3. Since the key information in neonatal EEG is located below 13 Hz, a bandpass filter (0.5~12.8 Hz) was employed – the 0.5 Hz cut-off frequency was adopted to avoid very low frequencies such as DC drift. The EEG was then resampled from 256 or 250 Hz to 32

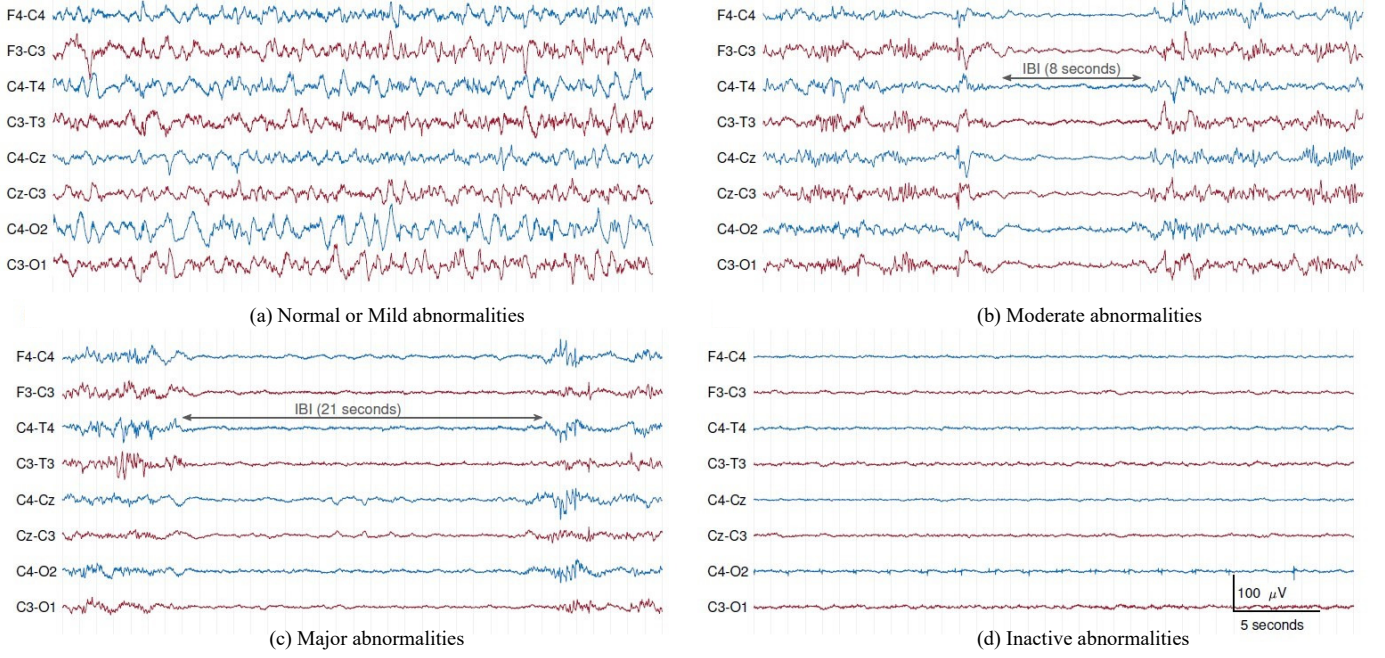


Fig. 1. EEG examples (30 seconds long) for different HIE grades. (a) normal or mild abnormalities (grades 1); (b) moderate abnormalities (grade 2); (c) major abnormalities (grade 3); (d) inactive abnormalities (grade 4); grades 2 and 3 with inter-burst intervals (IBI) annotation. (From [21])

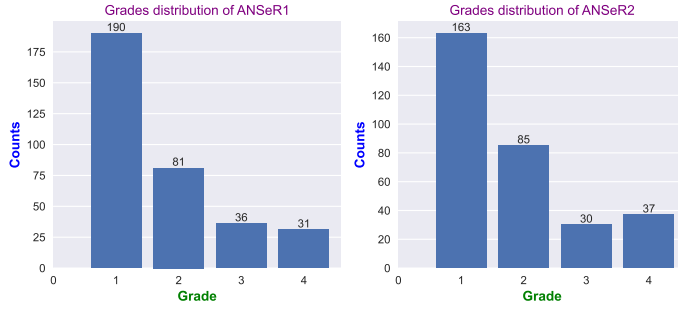


Fig. 2. The distribution of the HIE grades for a), the ANSeR1 HIE dataset & b), the ANSeR2 HIE dataset.

Hz. No other artefact removal procedures were used. It is common in EEG based classification tasks to reduce the effect of artefacts on performance, using techniques such as independent component analysis to separate the artefacts from the EEG [22]. However, in deep learning, with enough training data, the key features can be learned even from noisy EEG data when artefacts are present. In this work we are keen to demonstrate that excellent performance can be achieved even when there is minimal preprocessing of the raw EEG.

As HIE is a global injury, all the channels should display EEG characteristics in a similar way. Thus, channel-independent EEG epochs could be used for training a single channel-independent classifier. This allows the number of training samples to increase by 8 folds. In keeping with the approach proposed by [8] and [9] each 1-hour graded EEG segment was split into 60-second epochs using a 60s moving window with 50% of overlap; this was deemed long enough

TABLE I
MODEL PARAMETER OF DIFFERENT ARCHITECTURES

	L	S1	S2	K
FCN10	3	2	3	3
FCN13	4	1	4	3
FCN16	5	1	3	2

to capture the expected variation in the inter-burst interval (IBI) across grades. Each window was labelled with the grade assigned to that particular 1-hour epoch. Technically each 60-second window has what is called a “weak label”, as the actual grade may possibly fluctuate over the one hour.

Because the bandpass filter has removed the DC drift, the mean of each channel in the bipolar montage was zero. The standard derivation of each channel across all the training data ranged from $1 \times 10^{-6} V$ to $1 \times 10^{-4} V$. In a very basic normalization, all EEG channels were simply scaled up by a factor of 1×10^4 – as input samples which are very small might cause difficulty in the training of the initial convolutional layers which come before the first batch normalization layer.

2) *FCN architecture*: A 1-dimensional fully convolutional neural network (FCN) was used for the HIE grading. The FCN10/13/16 model architectures can be seen in Fig. 3, these were composed of 10/13/16 convolutional layers, respectively. The input of the model was a 60-second raw EEG segment sampled at 32 Hz. These 1920 samples can be thought of as an image for a single channel.

Table I presented the configuration of different model parameters. L is the number of layers of the feature extraction

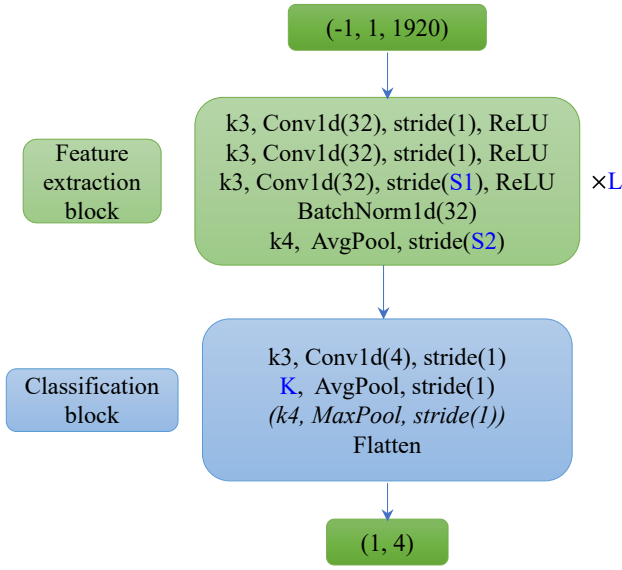


Fig. 3. Architecture of FCN10/13/16. FCN10 was composed of 3 (L) feature extraction blocks with 32 channels of convolutional layers and a classification block with 4 channels of a convolutional layer. FCN13/16 has 4/5 (L) feature extraction blocks with 32 channels of convolutional layers (S1 stride of 1) and a classification block without MaxPool layer.

block. S1 and S2 are the strides of the third convolutional layer and the Average Pooling layer in the feature extraction blocks, respectively. K is the kernel size of the average pooling layer in the classification block. Only the FCN10 has a Max Pooling layer (this helps the feature map size reduce to one within the limited number of layers). The deeper model has a smaller stride and kernel size, which maintains a gradual reduction of the feature map size from the network input to its output.

L layers of identical feature extraction blocks were used to learn the features from the raw EEG segments. Each of these feature extraction blocks consists of three 1-D convolutional layers with 32 channels. The BatchNorm layer normalizes the learned features and maintains them within a suitable distribution and hence helps the training process. The pooling layer not only reduces the sensitivity of the convolutional layer to positional information and increases the robustness of the model, but it also plays a downsampling role – this is useful in controlling the receptive field. In the classifier block, the convolutional layer projects from 32 channels down to 4 channels and summarizes the characteristics of the 4 HIE grades. Compared with the usual fully connected classifier layers, these convolutional layers ensure that the model has much fewer parameters, is simpler in structure and less likely to overfit.

3) *Post-processing*: For the validation and test set, a post-processing procedure was implemented to yield the final one-hour epoch prediction result. This procedure is based on a majority vote, which can be seen in Fig. 4. Each 60-second segment in each channel is classified as grade 1~4, independently using a single common FCN classifier. The grades are then collated, for the 119 (60 seconds) segments for each of the 8 channels. The resultant HIE grade for each one-

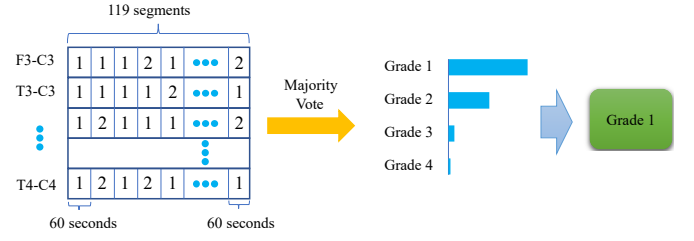


Fig. 4. Majority vote procedure – HIE grade estimated over 1 hour and multiple channels.

hour epoch is then determined by the majority vote (across channels and time).

4) *Metrics*: In our work, the prediction accuracy and the area under the receiver-operating curve (AUC-ROC) are used for the primary metrics of the classification performance. In HIE grading tasks, almost all of the published work has reported classification accuracy. However, due to the unbalanced distribution of each class, the classification accuracy cannot sufficiently represent the model performance. AUC measures the ability of a model to correctly classify positive and negative classes, even if one of the classes is more dominative than the other and represents the overall quality of the model performance for different thresholds. The Matthews Correlation Coefficient (MCC) is another common metric in classification tasks, which can also represent the classification performance regardless of the number of samples in each class. Compared with accuracy, MCC takes into account the 4 values of the confusion matrix, thus even if the positive class and negative class are swapped, its value does not change. It can be calculated by equation 1:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (1)$$

Where TP and TN are the true positive rate and the true negative rate in the confusion matrix; FP and FN are the false positive rate and the false negative rate, respectively.

The confidence interval (CI) provides a bound of the uncertain estimation. In our work, a 95% confidence interval was combined with the test accuracy, which not only provided a better measure of model performance but also facilitated a fair comparison between models.

III. RESULTS

In our experiments, the ANSeR2 dataset was used for training and the completely different, unseen, ANSeR1 dataset was used for the test. Although in many EEG-based classification tasks, the individual model's performance is better than that of a general model [23]. Furthermore, cross-dataset evaluation is more challenging as many subjects may have different environments, devices, and protocols [24]. An early stopping mechanism was adopted, in which a relatively large validation set was selected (for internal cross-validation) from the ANSeR2 training set (20% of epochs) in order to better evaluate the performance of different models. The significantly

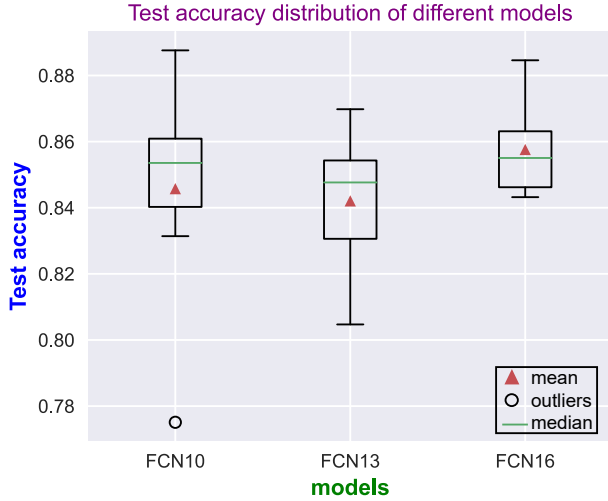


Fig. 5. Distributions of the test accuracy results (on ANSeR1 dataset) for each of the network models. For each model, 8 networks were trained, each with a different random initialization.

large unseen test dataset in ANSeR1 presents a compelling assessment of the generalization ability of the trained classifier. The training set includes 251 h of EEG. The validation set and test set have 63 h and 338 h of EEG, respectively. The cross-entropy function was used as the loss function. AdamW was selected as the optimizer, which allows for the decoupling of the weight decay from the setting of the learning rate in the training steps [25].

In the training process, 8 different networks were trained with different random initialisations. For each network initialization, the best model was selected in training based on the highest moving average validation AUC. The model performance for each of the 8 trained networks was then reported on the test set. Fig. 5 shows the test accuracy distribution on the ANSeR1 dataset from these 8 trained networks. Three different architectures were compared in this manner. The FCN16 architecture achieved the best performance with a higher mean and smaller spread. The FCN10 network performed better than the FCN13 (with a larger mean, median, maximum, and minimum), but with a significant outlier.

To reduce the bias of the reported model performance, the predictions provided by the 8 trained FCN16 networks were ensembled using a simple majority vote across networks. Table II presents the confusion matrix obtained using the ensemble of 8 FCN16 model prediction results from the 338 one-hour HIE grades (the ANSeR1 test set). The sensitivity indicates how likely the positive class (the other three classes were treated as the negative class) was correctly classified. The positive predictive rate (PPV) measures how confident an example was predicted as positive. The sensitivity of grade 1 (0.953) and the PPV of grade 4 (0.963) achieved the highest values, whereas the PPV of grade 3 (0.762) and sensitivity of grade 2 (0.642) have the lowest values, respectively. No epoch was wrongly predicted by more than one grade from the expert labelled grades. An 86.09% test accuracy from the ensembled

TABLE II
CONFUSION MATRIX OF 8 ENSEMBLED FCN16 MODELS PREDICTION RESULTS OVER THE ANSeR1 TEST SET BASED ON THE MAJORITY VOTE

		Prediction				
True Label		1	2	3	4	sensitivity
	1	181	9	0	0	0.953
	2	24	52	5	0	0.642
	3	0	3	32	1	0.889
	4	0	0	5	26	0.839
PPV		0.883	0.813	0.762	0.963	

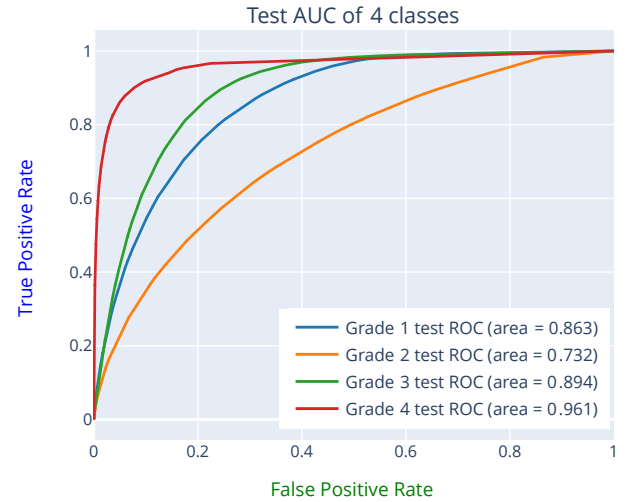


Fig. 6. ROC curve on the test set from one of the FCN16 networks predictions. Each class of AUC was obtained by treating them as a binary classification (the other 3 classes were the negative class).

prediction was achieved with a 95% CI of 82.41% 89.78% and 0.7691 of MCC. One of the FCN16 model prediction ROC curves on the test set can be seen in Fig. 6, which yielded an average test AUC of 86.23%.

IV. DISCUSSION

A. Comparison with CNN baseline

Many of the previous works were trained and tested on different datasets, thus it is difficult to directly compare their performance. Two of them were selected here as the baseline. In [10], a complex convolutional neural network extracted features from quadratic time-frequency distributions (TFD-CNN) was used. Although it was trained on a small dataset (54 neonates), it was also tested on the ANSeR1 dataset. The same architecture was also employed in [11], but the softmax

TABLE III
COMPARISON BETWEEN DIFFERENT METHODS' CONFIGURATIONS

	Dev. set	Test set	Dev. Method	Input
TFD-CNN[10]	54 neonates (Cork)	ANSeR1	Leave one out	TFD Features
TFD-CNN-Reg[11]	ANSeR1&2	ANSeR1&2	10-fold CV	TFD Features
Proposed	ANSeR2	ANSeR1	Internal CV	Raw EEG

TABLE IV
TEST ACCURACY, CONFIDENCE INTERVAL AND MCC COMPARISON BETWEEN DIFFERENT METHODS (THE ENSEMBLED FCN16 PREDICTION TOOK THE MAJORITY VOTE OF 8 TRAINED NETWORKS)

	TFD-CNN	TFD-CNN-Reg	FCN16(ensemble)
Accuracy	69.5%	82.8%	86.09%
95% CI	65.3%~73.6%	80.5%~85.2%	82.41%~89.78%
MCC	-	0.722	0.7691

layer was replaced as a regression layer (TFD-CNN-Reg). The whole ANSeR1&2 dataset was used for development, and a 10-fold cross-validation method was adopted. Therefore, a more direct comparison can be implemented based on this work.

Table III shows the comparison between the two baselines and the proposed method configuration. Table IV presents the performance of the different methods. The proposed FCN16 architecture achieved the best performance (average over 8 networks; each with random initialization). Not surprisingly, the TFD-CNN method achieved the lowest performance, due in some ways to a smaller development dataset (62 h). However, even with a substantially larger training set, this architecture with a slight modification of the final layer (TFD-CNN-Reg) only obtained a test accuracy of 82.8%. Moreover, a 10-fold cross-validation approach was adopted on 653 h of EEG. Compared with the proposed method (251 h for training, 63 h for validation and 338h for test), the TFD-CNN-Reg has much more data (about 588 h for training, 65 h for test). In terms of the model structure, the FCN presented here is much simpler. Although TFD-CNN only has 5 layers, it has more parameters (44,574) compared with FCN10 (25,540) and FCN16 (44,292). Rather than extracting the complex time-frequency distribution (TFD) features, the FCN16 directly used the raw EEG as the input. A higher MCC and 95% CI were still achieved. A 23.65% (86.09–82.8)/(100–86.09) relative performance improvement in test accuracy was obtained compared with the TFD-CNN-Reg method.

B. Receptive field

The receptive field is how many samples one element in the feature map can see from the input. A wider receptive field means that the model makes its decision on a longer window of the time series. The receptive field can be increased either by using a deeper network or by employing longer strides in the pooling layers.

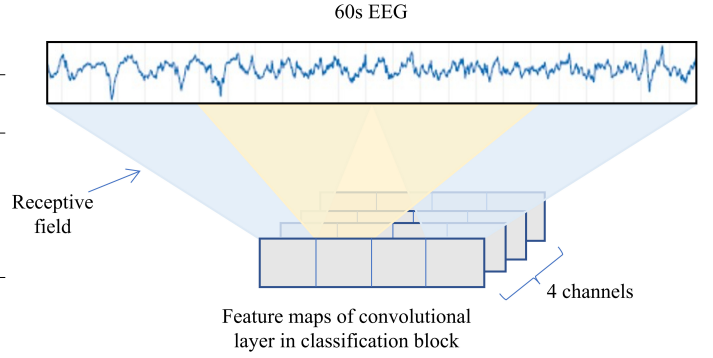


Fig. 7. Illustration of the receptive field of the convolutional layer in the classification block.

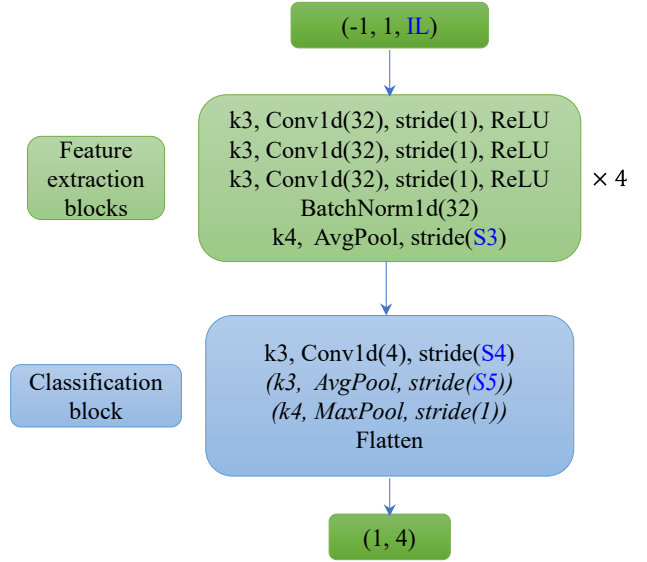


Fig. 8. FCN13_30/60/90s (16Hz) model architecture. Three models have the depth (13 layers) but different input length (IL: 30s, 60s and 90s of EEG). FCN13_90s model does not have the Average Pooling and Max Pooling layers in the classification block, and the FCN13_60s model only includes the convolutional layer and a Average Pooling layer.

Fig. 7 illustrates the receptive field of the convolutional layer in the classifier. The decisions made after the final pooling layer simply combine the outputs of the previous feature maps, yielding an overall decision which is based on approximately 60 seconds of EEG. However, these constituent feature maps may each have a receptive field which is significantly smaller than 60 seconds. We will concentrate on these as they are trainable – their receptive fields can be modified by depth, stride etc.

Table V presents the receptive field of the three network architectures and their validation (Val.), test accuracy (Acc.) and AUC (each averaged over the 8 trained models, with 8 random initializations). Surprisingly, the increase in the receptive field did not yield a large performance improvement. Although FCN16 has many more parameters and a wider receptive field, it is only slightly better in performance than the others. This means that a receptive field of approximately 30 s is large enough to extract details of the EEG to make decisions.

TABLE V
TEST ACCURACY, CONFIDENCE INTERVAL AND MCC COMPARISON
BETWEEN DIFFERENT METHODS (THE ENSEMBLED FCN16 PREDICTION
TOOK THE MAJORITY VOTE OF 8 TRAINED NETWORKS)

	Receptive field	No. of Params	Test Acc.	Test AUC	Val. Acc.	Val. AUC
FCN10	29.66s	25,540	0.8458	0.8693	0.8005	0.8674
FCN13	39.94s	34,916	0.8421	0.8643	0.7917	0.8689
FCN16	49.25s	44,292	0.8572	0.8620	0.8199	0.8639

TABLE VI
FCN13_30/60/90s MODEL CONFIGURATION

	IL	S3	S4	S5
FCN13_30s	480	2	2	2
FCN13_30s	960	3	2	1
FCN13_30s	1440	4	1	-

TABLE VII
FCN13_30/60/90s FCN13_30/60/90s MODEL RECEPTIVE FIELD AND
PERFORMANCE COMPARISON (AVERAGED ON 8 TIMES RANDOM
INITIALIZATION).

	Receptive field	MCC	Test Acc.	Test AUC	Val. Acc.	Val. AUC
FCN13_30s	10.5s	0.7261	0.8336	0.8784	0.6747	0.8609
FCN13_60s	32.69s	0.7692	0.8539	0.8694	0.7193	0.8693
FCN13_90s	79.88s	0.7773	0.8661	0.8793	0.7529	0.8893

It can also be noted that the test and validation performance across all models are quite similar, which could suggest that the proposed FCN model performance is very stable.

Models with different depths but similar receptive field demonstrated a similar performance. Another experiment was implemented to explore how the receptive field over a wide range effects the performance. Fig. 8 illustrates the model

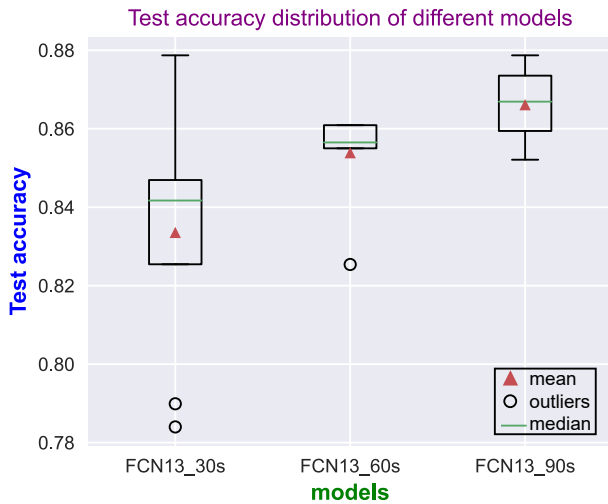


Fig. 9. Test accuracy distribution of FCN13_30/60/90s (16Hz) model on ANSeR1 dataset. Each model was trained with 8 times random initialization.

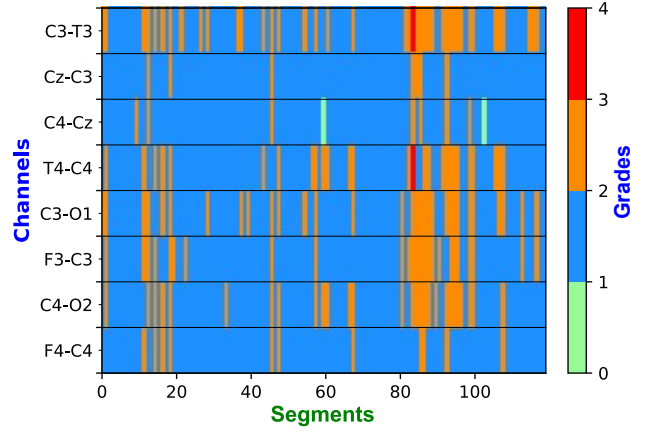


Fig. 10. Example of the color map over the one-hour epoch showing the grading decisions for 119 segments per channel (60 s for each segment with 50% of overlap). This one hour of EEG was expertly graded as Grade 2. After postprocessing, the decision of the classifier is also Grade 2.

structure of FCN13_30/60/90s. They have the same depth but different input lengths. EEG segments sampled at 16 Hz of length 30s, 60s and 90s were used as input in order to speed up the training. Due to the same depth, three models have the same number of parameters. Table VII shows their receptive field and the average test, validation performance over 8 times random initialization. It can be seen that the wider the receptive field, the better the model performance will be. However, compared with the difference between FCN13_30s and FCN13_60s model performance, the difference between 60s and 90s model is smaller, which can also be seen in the test accuracy distribution on Fig. 9. Although FCN13_90s has higher mean and median values, FCN13_60s has the smallest spread. It confirms again that an approximately 30s receptive field is longer enough to capture the key information.

C. Weak labels

The one-hour epoch of EEG only has one label (global label), which was assigned to all of its constituent channels and 60s segments; this can effectively be thought of as weak labelling of the dataset. Each segment was treated as an independent sample during the training and testing process. As a matter of fact, not all weak labels are correct. Neuro-physiologists annotated the one-hour EEG based on the whole one-hour multi-channel EEG trace. During the one hour, for example, the HIE might change from grade 1 to 2 and then back to 1. However, if most of the time is grade 1, the global annotation would be grade 1. This is the reason why majority voting was adopted in the post-processing process. Fig. 10 presents the colour map of a prediction example over one hour of 8-channel EEG which has been graded as Grade 2 (Moderate abnormalities) by the expert. Although many of the segments were marked as grade 3, the majority vote across all channels and time clearly provides the correct estimate of the HIE grade as 2 (Moderate). As a matter of fact, weak labels are sort of data augmentation, which not only inject some noise but also increased the number of samples. In [26], three kinds of

data augmentation methods were used and evaluated on three different EEG-based classification tasks, which consistently produced a performance improvement.

V. CONCLUSION

This paper presented a fully convolutional neural network architecture for HIE grading. The model was developed and assessed on two large HIE datasets containing unedited continuous 1-hour segments of multichannel EEG – there was a substantial mismatch between training and testing conditions. The training process was based on 60s segments with weak labels. The channel independence and majority voting made the prediction more robust. The minimally pre-processed EEG time-series data can be directly inputted, without the need for feature engineering. The raw EEG was only downsampled, filtered and scaled, and no other artefact removal and or normalization procedures were used. Compared with state-of-the-art, this architecture is simpler and deeper, but with fewer parameters. Compared with the TFD-CNN-Reg method [11], it was trained on a smaller amount of dataset but achieved comparable performance on a much larger, completely different test set.

ACKNOWLEDGMENT

We would like to thank the ANSeR consortium: INFANT Research Centre and Department of Paediatrics and Child Health, University College Cork, Cork, Ireland; Institute for Women's Health, University College London, London, UK; Utrecht Brain Center, University Medical Center Utrecht, Utrecht University, The Netherlands; Department of Neonatal Medicine and Division of Paediatrics, Department CLIN-TEC, Karolinska University Hospital, Karolinska Institutet, Stockholm, Sweden; Rotunda Hospital, Dublin, Ireland; Royal London Hospital and Queen Mary University of London, London, UK; Department of Clinical Neurophysiology, Great Ormond Street Hospital for Children NHS Trust, London, UK; Homerton University Hospital NHS Foundation Trust, London, UK; Clinical Neurophysiology, University Medical Center Utrecht, Utrecht, The Netherlands.

REFERENCES

- [1] R. C. Vannucci, "Hypoxic-ischemic encephalopathy," *American journal of perinatology*, vol. 17, no. 03, pp. 113–120, 2000.
- [2] D. V. Azzopardi, B. Strohm, A. D. Edwards, L. Dyet, H. L. Halliday, E. Juszczak, O. Kapellou, M. Levene, N. Marlow, E. Porter *et al.*, "Moderate hypothermia to treat perinatal asphyxial encephalopathy," *New England Journal of Medicine*, vol. 361, no. 14, pp. 1349–1358, 2009.
- [3] S. Roth, J. Baudin, E. Cady, K. Johal, J. Townsend, J. S. Wyatt, E. R. Reynolds, and A. L. Stewart, "Relation of deranged neonatal cerebral oxidative metabolism with neurodevelopmental outcome and head circumference at 4 years," *Developmental Medicine & Child Neurology*, vol. 39, no. 11, pp. 718–725, 1997.
- [4] D. M. Murray, G. B. Boylan, C. A. Ryan, and S. Connolly, "Early eeg findings in hypoxic-ischemic encephalopathy predict outcomes at 2 years," *Pediatrics*, vol. 124, no. 3, pp. e459–e467, 2009.
- [5] R. A. Hrachovy and E. M. Mizrahi, *Atlas of neonatal electroencephalography*. Springer Publishing Company, 2015.
- [6] I. Korotchikova, N. Stevenson, B. Walsh, D. Murray, and G. Boylan, "Quantitative eeg analysis in neonatal hypoxic ischaemic encephalopathy," *Clinical Neurophysiology*, vol. 122, no. 8, pp. 1671–1678, 2011.
- [7] L. Lacan, N. Betrouni, M.-D. Lamblin, L. Chaton, A. Delval, J.-L. Bourriez, L. Storme, P. Derambure, and S. N. Tich, "Quantitative approach to early neonatal eeg visual analysis in hypoxic-ischemic encephalopathy severity: Bridging the gap between eyes and machine," *Neurophysiologie Clinique*, vol. 51, no. 2, pp. 121–131, 2021.
- [8] N. Stevenson, I. Korotchikova, A. Temko, G. Lightbody, W. Marnane, and G. Boylan, "An automated system for grading eeg abnormality in term neonates with hypoxic-ischaemic encephalopathy," *Annals of biomedical engineering*, vol. 41, pp. 775–785, 2013.
- [9] R. Ahmed, A. Temko, W. Marnane, G. Lightbody, and G. Boylan, "Grading hypoxic-ischemic encephalopathy severity in neonatal eeg using gmm supervectors and the support vector machine," *Clinical Neurophysiology*, vol. 127, no. 1, pp. 297–309, 2016.
- [10] 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.
- [11] M. E. O'Sullivan, G. Lightbody, S. R. Mathieson, W. P. Marnane, G. B. Boylan, and J. M. O'Toole, "Development of an eeg artefact detection algorithm and its application in grading neonatal hypoxic-ischemic encephalopathy," *Expert Systems with Applications*, vol. 213, p. 118917, 2023.
- [12] A. O'Shea, G. Lightbody, G. Boylan, and A. Temko, "Neonatal seizure detection from raw multi-channel eeg using a fully convolutional architecture," *Neural Networks*, vol. 123, pp. 12–25, 2020.
- [13] A. Daly, A. O'Shea, G. Lightbody, and A. Temko, "Towards deeper neural networks for neonatal seizure detection," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 920–923.
- [14] A. H. Ansari, O. De Wel, K. Pillay, A. Dereymaeker, K. Jansen, S. Van Huffel, G. Naulaers, and M. De Vos, "A convolutional neural network outperforming state-of-the-art sleep staging algorithms for both preterm and term infants," *Journal of Neural Engineering*, vol. 17, no. 1, p. 016028, 2020.

- [15] S. A. Raurale, G. B. Boylan, G. Lightbody, and J. M. O'Toole, "Grading the severity of hypoxic-ischemic encephalopathy in newborn eeg using a convolutional neural network," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 6103–6106.
- [16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [17] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986.
- [18] W. Lim, O. Sourina, and L. P. Wang, "Stew: Simultaneous task eeg workload data set," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 11, pp. 2106–2114, 2018.
- [19] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks," *IEEE Transactions on autonomous mental development*, vol. 7, no. 3, pp. 162–175, 2015.
- [20] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.
- [21] J. M. O'Toole, S. R. Mathieson, S. A. Raurale, F. Magarelli, W. P. Marnane, G. Lightbody, and G. B. Boylan, "Neonatal eeg graded for severity of background abnormalities in hypoxic-ischaemic encephalopathy," *arXiv preprint arXiv:2206.04420*, 2022.
- [22] A. X. Stewart, A. Nuthmann, and G. Sanguinetti, "Single-trial classification of eeg in a visual object task using ica and machine learning," *Journal of neuroscience methods*, vol. 228, pp. 1–14, 2014.
- [23] M. Stikic, R. R. Johnson, V. Tan, and C. Berka, "Eeg-based classification of positive and negative affective states," *Brain-Computer Interfaces*, vol. 1, no. 2, pp. 99–112, 2014.
- [24] Z. Lan, O. Sourina, L. Wang, R. Scherer, and G. R. Müller-Putz, "Domain adaptation techniques for eeg-based emotion recognition: a comparative study on two public datasets," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 11, no. 1, pp. 85–94, 2018.
- [25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
- [26] R. Li, L. Wang, P. Suganthan, and O. Sourina, "Sample-based data augmentation based on electroencephalogram intrinsic characteristics," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 10, pp. 4996–5003, 2022.