A Deep Shared Multi-Scale Inception Network Enables Accurate Neonatal Quiet Sleep Detection with Limited EEG Channels

Amir H. Ansari [1], Kirubin Pillay [2], Anneleen Dereymaeker [3], Katrien Jansen [3,4], Sabine Van Huffel [1], Gunnar Naulaers [3], Maarten De Vos [1,3]

***Abstract*— In this paper, we introduce a new variation of the Convolutional Neural Network Inception block, called Sinc, for sleep stage classification in premature newborn babies using electroencephalogram (EEG). In practice, there are many medical centres where only a limited number of EEG channels are recorded. Existing automated algorithms mainly use multi-channel EEGs which perform poorly when fewer numbers of channels are available. The proposed Sinc utilizes multi-scale analysis to place emphasis on the temporal EEG information to be less dependent on the number of EEG channels. In Sinc, we increase the receptive fields through Inception while by additionally sharing the filters that have similar receptive fields, overfitting is controlled and the number of trainable parameters dramatically reduced. To train and test this model, 96 longitudinal EEG recordings from 26 premature infants are used. The Sinc-based model significantly outperforms state-of-the-art neonatal quiet sleep detection algorithms, with mean Kappa 0.77 ± 0.01 (with 8-channel EEG) and 0.75 ± 0.01 (with a single bipolar channel EEG). This is the first study using Inception-based networks for EEG analysis that utilizes filter sharing to improve efficiency and trainability. The suggested network can successfully detect quiet sleep stages with even a single EEG channel making it more practical especially in the hospital setting where cerebral function monitoring is predominantly used.**

***Index Terms*— Convolutional Neural Networks, CNN, Inception networks, multi-scale EEG analysis, sleep stage classification, neonatal EEG analysis.**

I. Introduction

Sleep Wake Cycling (SWC) is one of the main neuro-developmental markers for newborn infants (neonates) [1]. Preterm babies who are born at <37 weeks Postmenstrual Age (PMA, the age since the last menstrual cycle of the mother) are especially susceptible to long-term neurodevelopmental consequences if their early-stage sleep maturation is disturbed by environmental stresses. Similarly to adults and older babies, preterm neonates cycle between stages of Wakefulness, Rapid Eye Movement (REM, Active Sleep (AS)), and non-REM (Quiet Sleep (QS)). However, unlike in older ages, there are no sub-states in preterm AS and QS. AS is mostly associated with (semi)-continuous background EEG patterns across all frequency bands and synchronous occipital delta activities, while QS is discontinuous with burst (high amplitude) and inter-burst-interval (IBI, low amplitude) patterns [1]. AS EEG has similar morphologies to wakefulness EEG such that polygraphic signals (e.g. electrocardiogram, respiration) are required to differentiate these states if needed [2]. The current gold standard for detecting these sleep stages is visual labelling of polysomnographic recordings using predominantly continuous Electroencephalogram (EEG). However, this visual labelling is time-consuming and needs particular expertise that may not be available around the clock. Automated sleep staging can alleviate the workload of the clinicians and a bedside monitor equipped with such algorithms could reduce unnecessary sleep disturbances to the vulnerable baby (e.g. the timing of feeding) improving the quality of perinatal care [1].

A. H. A. , M. D. V. and S. V. H. are supported by: Bijzonder Onderzoeksfonds KU Leuven (BOF): The effect of perinatal stress on the later outcome in preterm babies : C24/15/036, Prevalentie van epilepsie en slaapstoornissen in de ziekte van Alzheimer: C24/18/097; Fonds voor Wetenschappelijk Onderzoek-Vlaanderen (FWO): PhD/Postdoc grants; Agentschap Innoveren en Ondernemen (VLAIO): 150466, OSA+, O&O HBC 2016 0184 eWatch; financial support of imec; EU: EU H2020 FETOPEN 'AMPHORA' #766456, EU H2020 MSCA-ITN-2018: 'INtegrating Magnetic Resonance SPectroscopy and Multimodal Imaging for Research and Education in MEDicine (INSPiRE-MED)', funded by the European Commission under Grant Agreement #813120, EU H2020 MSCA-ITN-2018: 'Integrating Functional Assessment measures for Neonatal Safeguard (INFANS)', funded by the European Commission under Grant Agreement #813483; EIT: 19263 – SeizeIT2: Discreet Personalized Epileptic Seizure Detection Device; Flemish Government: This research received funding from the Flemish Government (AI Research Program). S.V.H., A. H. A. and M. D. V. are affiliated to Leuven.AI - KU Leuven institute for AI, B-3000, Leuven, Belgium. AHA is furthermore supported by FWO postdoctoral fellowship. K. P. is funded by the Wellcome Trust UK.

[1] A. H. A., M. D. V., and S. V. H are with Department of Electrical Engineering (ESAT), STADIUS, KU Leuven, Belgium (e-mail: amirhossein.ansari@kuleuven.be).

[2] K. P. is with Department of Paediatrics, John Radcliffe Hospital, University of Oxford, Oxford, United Kingdom (e-mail: kirubin.pillay@paediatrics.ox.ac.uk).

[3] A. D., K. J., G. N., and M. D. V. are with Department of Development and Regeneration, University Hospitals Leuven, Neonatal Intensive Care Unit, KU Leuven, Leuven, Belgium (e-mail: gunnar.naulaers@uzleuven.be).

[4] K. J. is also with Department of Development and Regeneration, University Hospitals Leuven, Child Neurology, KU Leuven, Leuven, Belgium (e-mail: katrien.jansen@uzleuven.be).

Consequently, previous studies have developed various machine learning approaches to automated sleep staging (with a focus on QS) using multichannel EEG. Clustering of discriminative features extracted from adaptively segmenting the EEG was an early suggestion by Barlow et al. [3], improved by Krajca et al. [4], [5], before being developed into the fully-automated unsupervised 'CLASS' algorithm by Dereymaeker et al. [6]. As an alternative approach, Gerla et al. combined EEG features with a decision tree classifier and evolutionary optimization using a supervised approach [7]. Subsequently, more refined feature extraction and classifiers have been applied, such as the Support Vector Machine (SVM) with spectral, temporal and spatial features [8], the Least-Squares SVM (LS-SVM) with multi-fractal features [9], a Multilayer Perceptron (MLP) using entropy features [10], and other extensions that utilised multi-scale entropy [11], [12] . De Wel et al. also suggested a tensor decomposition approach for unsupervised classification [13], while Pillay et al. incorporated a large number of data-driven features in a combined Hidden Markov Model-Gaussian Mixture Model (HMM-GMM) which has been extended to additionally subclassify the 4 stages of sleep (AS I, AS II, QS I and QS II) that differentiate by term age (≥37 weeks PMA) [14].

Existing unsupervised methods have typically low accuracy, while supervised approaches still depend on large numbers of hand-crafted features which may not adequately capture the extent of the available information. Consequently, in the last decade, deep learning for biomedical signal analysis has emerged as a more accurate solution [15]–[20] extracting data-driven features directly from the raw EEG. More recently, similar approaches have found their way to neonatal sleep staging. Fraiwan et al. proposed a bi-directional Long-Short Term Memory (LSTM) for classifying term sleep [21]. Ghimatgar et al. further developed a hybrid approach using graph clustering, LSTM, and HMM for term sleep classification [22]. We have also suggested a network expanded to both preterm and term age groups, developing multiple deep Convolutional Neural Networks (CNNs) that outperformed the aforementioned feature-based approaches [23], [24]. However, new complications arise. The high dependency of these network architectures on spatial (multi-channel) EEG results in a large reduction in accuracy as the number of channels are reduced.

In general, the number of EEG electrodes for neonatal brain monitoring mostly does not exceed 19 - 21 electrodes which are usually placed according to the international 10-20 system. However, for premature neonates with smaller head circumferences, it is mostly limited to 8 - 9 electrodes placed using a restricted version of the 10-20 system [2], [25]. However, in practice, multi-channel EEG monitoring is not commonplace across Neonatal Intensive Care Units (NICUs) due to the clinical training and time required to apply and acquire each recording. It is more typical for centres to use Cerebral Functional Monitoring (CFM) as an alternative, which is a derivative of EEG and uses one or two bipolar EEG channels. In neonatal sleep stage classification, respecting the fact that many EEG characteristics synonymous with sleep staging (e.g. EEG 'bursts') have strong spatiotemporal dependencies, reducing the number of EEG channels leads to a severe drop in performance.

In this study, we present a new end-to-end deep learning architecture for automated QS detection. We introduce a multi-scale deep CNN, that utilises a novel 'Sinc' block to better extract temporal features across multiple timescales. The Sinc block applies a filter sharing approach in Inception which reduces massively the number of required parameters when different scales are required. We show that including this Sinc block allows the model to outperform existing approaches for neonatal QS detection especially when fewer numbers of EEG channels are available.

II. Materials and Methods

A. Database

The database used in this study were recordings from 26 neonates admitted in the NICU of the University Hospitals Leuven, Belgium, between 2012 and 2014. All babies were born prematurely with Gestational Age (GA) <32 weeks and were recruited after approval by the Medical Ethics Committee of the hospitals and parental consent. 2-4 EEGs were recorded during each baby's stay in the NICU resulting in 96 longitudinal EEG recordings with Postmenstrual Age (PMA) range 27-42 weeks. All 26 neonates had a normal neurodevelopmental outcome at 9- and 24-months follow-up (Bayley Scores of Infant motor Development-II) [26]. Furthermore, patients with severe cerebral lesions or use of sedative or anti-seizure medication were excluded during recording.

The original EEG recordings included F1, F2, C3, C4, T3, T4, O1, O2 and referenced Cz channels (totalling 8 channels) according to the restricted standard 10-20 electrode placement system [2]. Recordings were acquired at a sampling frequency of 250 Hz using the BrainRT EEG recording system (OSG BVBA Rumst, Belgium) and were annotated as states of QS and non-QS (NQS – combining AS and Wakefulness) by an expert clinical neurophysiologist (AD). All recordings were anonymized by the clinicians before analysis and further details about the demographics are given in [1], [27]. This database was also previously used in [6], [9], [13], [23], [24]. Additional technical and medical details about this database are provided in [1].

B. Data preparation and pre-processing

In order to develop and validate the proposed model, the database was randomly split into a training and testing dataset by *patient* such that each group included the recordings from 13 neonates (50-50 split). This fixed split was previously used in [6], [23], [24] for developing the previous QS detection methods and thus provides a fair comparison when evaluating model performance. Each EEG recording was under-sampled to 64 Hz with an antialiasing (zero-phase low-pass FIR) filter. Data was then segmented into epochs of length 30s with 50% overlap.

In order to evaluate these considered algorithms using fewer number of EEG channels, we simulated the channel reduction in a structured manner to resemble the common outputs of CFM monitors, using all 8 available channels as 8 channel arrangement, C3, C4, T3, and T4 as a reduced 4 channel arrangement, C3 and C4 for 2 channels, and finally a bipolar C3

– C4 montage for the single channel case. From a clinical point of view, these central EEG channels can reflect many of the sleep-related patterns that mature from preterm to term age [1].

C. Multi-scale deep learning networks

An important feature in the proposed model architecture is multi-scaling which results in different Receptive Fields (RFs) across the network. The term 'receptive field' emerged in early 20th century research to describe the region of the body surface where a stimulation results in a specific neuronal response [28]. Subsequently, the deep learning field adopted this concept for artificial neural networks to define the region or cluster of data which stimulates a specific neuron (node) in the network.

Typically, the RF size for a CNN layer is directly defined by the layer's kernel (or filter) length used in the convolution operation. For instance, the RF of a CNN layer with kernel size 3 and stride (step size) of 1 would be 3. This essentially means that each output sample of the layer is the result of the combination of 3 samples in the input to the layer. In a time-series signal (such as EEG), calculation of the RF with respect to the initial input to the network can be very helpful for designing and interpreting the network dynamics. As an example, if the RF of the output of an arbitrary CNN layer (with respect to some input EEG) were 50, it means that this layer effectively extracts features over 0.5 seconds of the original EEG (assuming a sampling frequency of 100 Hz.). Thus, we cannot expect that an EEG burst pattern (present in QS) lasting for 1s to be fully characterised by this particular layer.

In conventional CNNs, the RF of each layer with respect to the network input can be mathematically formulated in a simple manner. A layer with kernel size k increases the RF by $(k-1)\prod D_i$, where D_i is the downsampling factor of the previous layers (e.g. a D_i is 2 after a maxpool with stride of 2). One drawback of such a conventional, sequential CNN is that the input RF to the subsequent layer is a single value. Previous EEG studies have shown that the extraction of features across multiple scales provides a better representation of the EEG information [29], [30]. One classic example are frequency band decompositions (using Fourier Transforms, Wavelet decomposition, Empirical mode decompositions etc.) that require longer timescales to represent the lower frequencies but are proven strong candidates for successful sleep classification [23].

To similarly extract features across multiple frequencies using a CNN layer, a multi-scaling approach that provides multiple input RFs should be provided. This idea was first proposed in the GoogLeNet architecture and dubbed the 'Inception' block [31]. An inception block typically consists of processing the same input using 3 parallel convolutional 'streams' with kernel size (k) 1, 3, 5, respectively (and an additional pooling layer). This produces 3 different RF outputs for a subsequent layer. Subsequently, it has been shown that the inception block can perform better if the larger convolutions are split (or factorized) to two consecutive convolutions that achieve the same output RF [32]. We will see that this factorization approach is a key element of our proposed architecture.

While these inception blocks are well validated in computer vision tasks [32]–[35], initial testing of such model derivatives to preterm EEG analysis were inefficient due to the large increase in the number of parameters as a result of the parallelised architecture. In EEG time series, when the sampling frequency is tens or hundreds of Hz, some scales may simply represent noise and cause overfitting. The proposed architecture modifies the inception block in order to share some of the filters as a means to control this overfitting risk. The resulting shared inception block is called 'Sinc'.

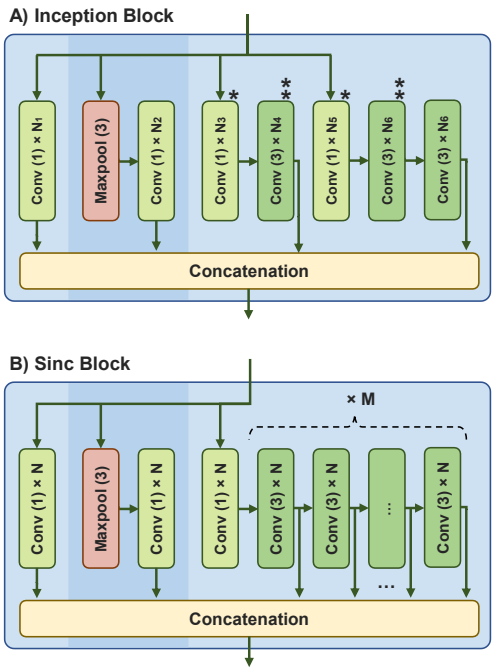

Fig. 1. the block-diagram of the original factorized Inception block (A) and the proposed Sinc block (B). The layers that are marked with asterisks denote the layers in Inception with similar receptive fields that should have shared weights to be equal to a Sinc layer with M = 2.

D. The proposed Sinc block

Fig. 1 shows the block-diagram of the factorized Inception (A) and the proposed Sinc (B). In the proposed Sinc, inspired from the original inception architecture, three streams are used: 1) a convolution $(k = 1)$ with N filters to project the input feature maps (projection convolution), namely Conv1, 2) a maxpool (of size 3) followed by a second projection Conv1, 3) a third projection Conv1 followed by a series of M convolutional layers $(k = 3)$ with N filters, namely Conv3. Projection convolutions decrease the computational cost as it performs dimensionality reduction, while simultaneously increasing the depth and nonlinearity of the network [31]. The key difference between Sinc and the original inception model is the efficient generation of additional RF outputs within the 3rd parallel stream by using the factorisation principle. In the 3rd stream of Inception, for each RF, there is an isolated path between the input and the concatenation layer. In Fig. 1 A) there are two paths (= Conv1-Conv3 for RF=3 and Conv1-Conv3-Conv3 for RF=5) as proposed in the original paper. One can show that the size of the block, and therefore the number of parameters, is quadratically expanded by increasing the RF of

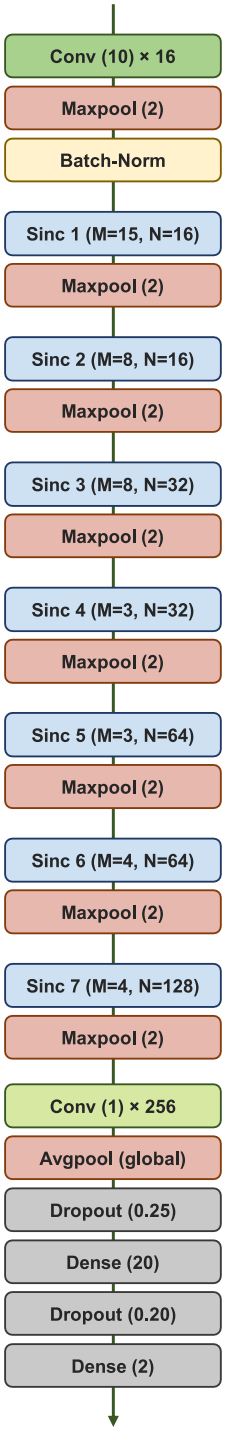

Fig. 2. the proposed Sinc network.

the Inception, as needed in this EEG analysis. Now, one can make the first Conv3s of all paths (RF = 3, 5, 7, etc.) identical (or share their parameters) and similarly for the second Conv3s and so on (see the asterisks in Fig. 1 A). This will convert the Inception block to a Sinc block. In this way, the first Conv3 produces an output RF of 3. This is then combined with a second Conv3 to produce the next RF of size 5, and these two layers link with a third layer to generate the RF output of size 7, and so on. In general, M convolutions are used consecutively

TABLE I
THE LAYERS OF THE PROPOSED NETWORK

Layer/ Block	Parameters	Output shape	Max RF*
Input	-	1920 * 8	-
Conv	k: 10, s: 1, N: 16	1920 * 16	10 (0.2 s)
Maxpool	k: 2, s: 2	960 * 16	11 (0.2 s)
B-Norm		960 * 16	11 (0.2 s)
Sinc 1	M: 15, N: 16	960 * 272	71 (1.1 s)
Maxpool	k: 2, s: 2	480 * 272	73 (1.1 s)
Sinc 2	M: 8, N: 16	480 * 160	137 (2.1 s)
Maxpool	k: 2, s: 2	240 * 160	141 (2.2 s)
Sinc 3	M: 8, N: 32	240 * 320	269 (4.2 s)
Maxpool	k: 2, s: 2	120 * 320	277 (4.3 s)
Sinc 4	M: 3, N: 32	120 * 160	373 (5.8 s)
Maxpool	k: 2, s: 2	60 * 160	389 (6.1 s)
Sinc 5	M: 3, N: 64	60 * 320	581 (9.1 s)
Maxpool	k: 2, s: 2	30 * 320	613 (9.6 s)
Sinc 6	M: 4, N: 64	30 * 384	1125 (17.6 s)
Maxpool	k: 2, s: 2	15 * 384	1189 (18.6 s)
Sinc 7	M: 4, N: 128	15 * 768	1920 (30.0 s)
Maxpool	k: 2, s: 2	7 * 768	1920 (30.0 s)
Conv	k: 1, s: 1, N: 256	7 * 256	1920 (30.0 s)
Avgpool	k: 7, s: 1, global	256	1920 (30.0 s)
Dropout	α = 25%	256	-
Dense	-	20	-
Dropout	α = 20%	20	-
Dense	softmax	2	-

Max RF: the maximum receptive field (scale) with respect to the input layer in samples and (in seconds).
M: number of shared filters in each Sinc block
N: number of filters in each convolution
k: kernel size
s: stride

and the outputs of these and the first two parallel streams are concatenated together for subsequent processing. Unlike Inception, the number of parameters in a Sinc linearly increases by increasing the RF size.

In Sinc, for all convolutions (Conv1/Conv3), N filters are trained such that each Sinc block has only two hyperparameters: M = the number of consecutive convolutions (with maximum RF size = $M + 1$), and N = number of filters. Similar to the original inception block, the stride of the maxpool and the convolutional layers are 1. As with typical CNN layers, an activation layer is also used after the convolutions to add non-linearity to the network. In this Sinc block, an exponential linear unit (ELU) is used. Unlike the more typical Rectified Linear Unit (ReLU), ELU permits negative values that can reduce the bias shift and tends to converge faster in training [36].

E. The complete network architecture

Fig. 2 shows the block-diagram of the complete Sinc-based network and Table I shows the output size and receptive fields of each layers, with respect to the input layer. The input of the network is a normalized multichannel 30-second EEG segment. By the first convolution stage, the EEG has been temporally filtered and the channels spatially integrated. We subsequently refer to this as 'early integration'. Afterwards, a maxpool with stride 2 and a batch-normalization layer downsamples and normalizes the activations respectively. This is then processed by a first Sinc block with $M = 15$ convolutions (producing 16 different scales/RFs covering 0.2s to 1.1s of the EEG). A further maxpool downsamples the data by a factor 2, with additional

Sinc blocks and maxpool layers repeated (with different choices of M and N) until each output sample contains information from the whole 30-second EEG input. Finally, a projection convolution and a global average-pooling decreases the dimensionality to control the parameter numbers, before two dense layers (with regularizing dropout layers) perform the classification. Across the network, ELU is used as the activation layer with the exception of the last layer where a softmax generates the class probabilities.

F. Post-processing

After running the network on all the segments for each recording, a moving average filter (of length 3 minutes) smoothens these outputs. As shown in previous models, this decreases the risk of spurious false positives due to short, high power EEG artefacts [23], [24]. As a side-analysis, we also tested an LSTM layer after the CNNs as is a typical approach in time series analyses [20]. However, it did not meaningfully improve the results over the simpler moving average filter.

G. Benchmark algorithms and the state-of-the-art

1). Multiscale Entropy Tensor Decomposition (METD): an unsupervised QS detection algorithm proposed in *[13]*. In this method, first the multichannel EEG is tensorized via multiscale entropy and is then factorized by canonical polyadic decomposition (CPD) in a sum of multiple rank-1 tensors. Next, the best CPD component corresponding to the sleep stages is automatically detected using an autocorrelation analysis. Finally, the QS intervals are detected by applying k-means clustering on the smoothed temporal signature of the selected component. Since the used core tensor in this algorithm needs spatial information, it does not support channel reduction.

2). Cluster-based Adaptive Sleep Staging (CLASS): an unsupervised QS detection algorithm developed in [6]. In this method, after removing high-power, short-duration muscle artefacts using the artefact subspace reconstruction technique, the EEG is split into smaller quasi-stationary segments. Then, nine time and frequency-domain features are extracted for each segment. The segments are then grouped into 12 clusters using k-means. Smoothing and thresholding the resulting 'cluster vs time' signals define the QS cycles. This method does not support channel reduction.

3). Feature-based QS Detection (FBD): a supervised classical machine learning approach using hand-crafted features and SVM proposed in [37]. In this method, 9 time and frequency-domain features are extracted from each channel of every 30-second EEG epoch. Then, all features are input to an SVM with Radial Basis Function (RBF) kernel to detect QS epochs. This method does not support channel reduction.

4) and 5). A 2-dimensional (2D) CNN network proposed in [23] and its improved version proposed in [24] for neonatal sleep staging: these networks exploit both spatial and temporal information by using 2D kernels. The EEG channels are integrated in the intermediate layers across multiple steps. These are the current published state-of-the-art [24].

6). A 1-dimensional (1D) CNN network proposed in [16] first developed for neonatal seizure detection: The main feature of this architecture is that the EEG channels are integrated at the last layer of the CNN. This design makes the network able to exploit temporal information for the majority of the processing, requiring limited retraining in cases with fewer numbers of EEG channels.

7). A modified version of [24] with earlier channel integration: In the first convolutional layer, all EEG channels are integrated and the remaining CNNs are 1D. The kernel size of the 1D filters and number of filters are the same as the original network described in 2.) [24].

8). A newly designed Conv-LSTM network: a Conv-LSTM unit is an extension of the LSTM, where both input and recurrent transformations are convolutional [38]. This network has 4 layers of Conv-LSTM and 2 further convolutional layers. The initial filtering layers and final dense layers are the same as the Sinc network. More details are provided in Appendix A.

9) and 10). 'Dilated' and 'Dense': the proposed Sinc network where the Sinc blocks are replaced by 'Dilated CNN' and 'Dense blocks'. These two blocks are two of the well-known multi-scale versions of CNN for computer vision and image segmentation tasks, which were respectively proposed in [39] and [40]. The former uses dilation of (2, 4, 6, and 8) and the latter utilizes 4 densely connected convolutional layers in each block. The number of filters increases corresponding to the layer depth so that the total number of trainable parameters are comparable with the proposed Sinc network.

11). 'Inc': the proposed Sinc network where the convolution layers within the Sinc blocks are not shared and ReLU is used as the nonlinear function (as proposed in the original Inception network). This comparison highlights the effects of the proposed modifications.

H. Training methodology

The Sinc network, as well as the benchmarked algorithms, was implemented in Python using Tensorflow 2.1 (Keras) and trained with an NVIDIA GPU (RTX 2080Ti) using the same training and testing data splits. An early stopping was used which monitored the validation loss and checked 10 further epochs after the minimum loss. For model-selection and early-stopping, a fixed validation set was additionally separated from the training dataset. This was generated by removing one random recording from each PMA group: (27-29, 29-31, …, 41-43 weeks). The loss function and the optimization approach were the categorical cross-entropy and Adam [41], respectively. An L_2 norm regularization (weight decay) was used across all convolutional layers in the main stem and the Sinc blocks (with regularization factor = 0.001). Finally, the moving average post-processing step in the Sinc model was also applied to the estimates from the alternative networks, for consistency.

All models were trained with 5 different random weight initialisations, and the best performing version (assessed on the validation dataset) was taken forward for final assessment on

TABLE II
COMPARISON OF THE BENCHMARKED NETWORKS FOR NEONATAL QUIET SLEEP DETECTION

Method	chan. integ.[1]	Cohen's Kappa (95% confidence interval) 8-ch	4-ch	2-ch	1-ch
METD [13]	-	0.50 (std[2]: 0.38)	NA	NA	NA
CLASS [5]	-	0.66 (std: 0.24)	NA	NA	NA
FBD [21]	-	0.70 (std: 0.21)	NA	NA	NA
Conv-2D [23]	Mid	0.71 (0.704 - 0.718)	NA	NA	NA
Conv-2D [24]	Mid	0.75 (0.742 - 0.756)	0.70 (0.691 - 0.705)	0.65 (0.645 - 0.660)	0.51 (0.505 - 0.523)
Conv-1D [16]	Late	0.58 (0.574 - 0.591)	0.70 (0.695 - 0.709)	0.59 (0.582 - 0.597)	0.26 (0.251 - 0.264)
Conv-1D [3]	Early	0.72 (0.715 - 0.728)	0.69 (0.681 - 0.695)	0.64 (0.630 - 0.645)	0.61 (0.607 - 0.622)
Conv-LSTM	Early	0.67 (0.661 - 0.676)	0.73 (0.719 - 0.734)	0.60 (0.590 - 0.606)	0.70 (0.691 - 0.706)
Dilated	Early	0.71 (0.699 - 0.714)	0.70 (0.692 - 0.707)	0.54 (0.536 - 0.554)	0.54 (0.531 - 0.548)
Dense	Early	0.68 (0.675 - 0.689)	0.71 (0.703 - 0.717)	0.69 (0.687 - 0.702)	0.69 (0.680 - 0.695)
Inc	Early	0.76 (0.756 – 0.769)	0.74 (0.737 – 0.750)	0.70 (0.696 – 0.711)	0.69 (0.678 – 0.693)
Sinc	**Early**	**0.77 (0.761 - 0.774)**	**0.76 (0.751 - 0.765)**	**0.74 (0.730 - 0.744)**	**0.75 (0.746 - 0.760)**

[1] channel integration indicating the network depth where the EEG channels are merged together
[2] for these reference algorithms the 95% confidence intervals were not reported. Instead, the standard deviations (std) are listed.
[3] a modified version of [24] with an early EEG integration

the test set. This was also repeated using a series of reduced channel arrangements. We simulated the channel reduction in a structured manner to resemble the common outputs of CFM monitors, using C3, C4, T3, and T4 as a reduced 4 channel arrangement, C3 and C4 for 2 channels, and finally a bipolar C3 – C4 montage for the single channel case.

I. Statistical analysis, ablation study, and visualization

The main metric that we used in this paper to evaluate the performance is Cohen's Kappa. Kappa is a metric that was originally proposed for interrater agreement measurement. Nowadays, it is also widely being used in machine learning as a normalized version of typical accuracy. A key feature of Kappa is that it is less sensitive to unbalanced classes, such as in neonatal sleep staging (where NQS predominates QS), and is thus a good alternative for accuracy. It is calculated as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (1)$$

where p_o denotes the observed agreement (equals accuracy) and p_e is the expected agreement (agreement by chance). Kappa values range from -1 (absolute disagreement) to +1 (absolute agreement) and 0 represent chance agreement. The standard error of the kappa is also analytically calculable [42].

In order to understand the role of each Sinc block, an ablation study inspired from [43] is performed on the network. To this end, after training, the outputs of the Sinc blocks are used as input of a simple classifier composed of a Flatten, Dropout (25%) and Dense layers to predict the sleep stages. This new classifier block is trained using the same training data while the rest of the network is frozen. The kappa value for both validation and test sets are reported for each block. This analysis shows how the features extracted by each of the Sinc blocks are discriminative for the QS detection task.

To visualize the network parameters and outputs during training, two methods are applied. To visualize the (many) features generated by each of the Sinc blocks, a 2D UMAP [44] is used. The key feature of the UMAP is to preserve both local and most of the global data structures in a reduced representation (a form of dimensionality reduction). From this approach, we can monitor how the Sinc blocks discriminate between QS and NQS and the contributions of each layer to the overall classification. We also generate a UMAP to show the structure of the data with respect to the PMA distribution of the recordings.

The second visualization method is based on the activation of the convolutional layers in the Sinc blocks. To this end, for each convolution, the output of the ELU is stored for all QS and NQS segments in the test dataset. Then, the average values for the QS and NQS segments are separately calculated and compared together to determine at which sleep stage the neuron is most activated. The neurons that always produce negative values are called dead neurons and can be pruned in further steps.

III. RESULTS

Table II compares the Cohen's Kappa values with analytical confidence intervals (CIs) of the benchmarked networks and with the reduced channel arrangements. For the first three algorithms, as the confidence interval was not reported in the corresponding papers, the standard deviation is listed. In this table, it is shown that the proposed Sinc network significantly outperforms the alternative algorithms. When all 8 EEG channels are available, the proposed Sinc has a better performance. When decreasing the number of available channels, the Sinc network has an increasing superiority and its performance with a single EEG channel is comparable or higher than the alternative networks (even when these utilise the full EEG). In addition to the listed 95% confidence intervals, a bootstrap hypothesis test also indicates the statistical superiority (p<0.05) of the Sinc model over all cases (except 8-ch "Inc", though the proposed model still maintains a higher performance). For more details on the bootstrap hypothesis testing, see [45], [46].

These results also suggest that integrating the EEG channels in the first layers resulted in better performance. This is especially evident when the two Conv-1D models (with late and early integrations) are compared. Another observation is that the Dense network, which also has a multi-scale characteristic, has almost constant performance when fewer numbers of channels

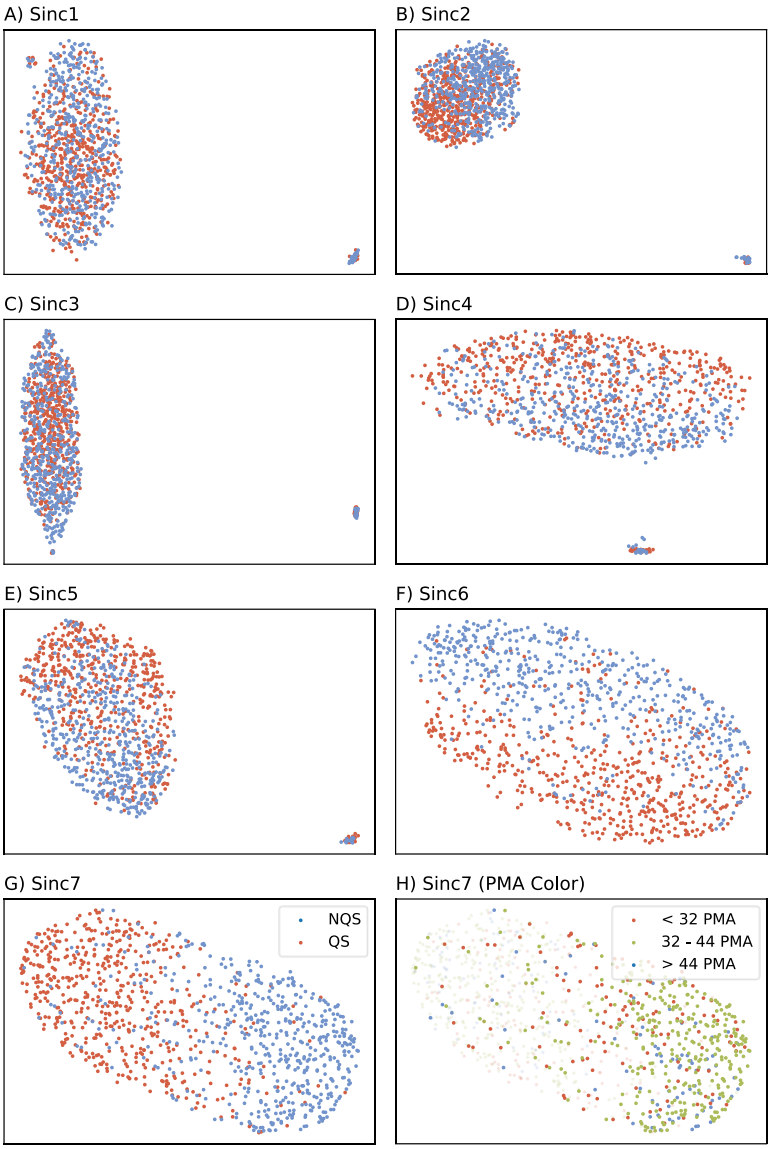

Fig. 3. UMAP visualization of the outputs of different Sinc blocks from 1 to 7. In A to G, blue and red dots correspond to the NQS and QS EEG segments, respectively. In H, red, olive, and blue colors correspond to different post menstrual ages of the non-quiet sleep segments (PMA groups are indicated in the legend).

are used (similarly to Sinc) though Sinc has an overall higher performance. To see more performance metrics, similar tables for the Area Under the Curve (AUC) and Accuracy (ACC) are also provided in Appendix-B.

TABLE III
COHEN'S KAPPA VALUES OF THE ABLATION STUDY

	Validation		Test	
	8-ch	1-ch	8-ch	1-ch
Sinc1 input	0.39	0.26	0.36	0.44
Sinc1 output	0.54	0.40	0.51	0.50
Sinc2 output	0.66	0.54	0.63	0.68
Sinc3 output	0.66	0.58	0.65	0.67
Sinc4 output	0.70	0.62	0.69	0.70
Sinc5 output	0.70	0.66	0.73	0.72
Sinc6 output	0.73	0.66	0.76	0.75
Sinc7 output	0.73	0.67	0.77	0.75

Table III shows the results of the ablation study of the Sinc blocks. The study was done for both validation and test datasets and for the 8-channel and 1-channel configurations. The first observation is that each of the first 6 Sinc blocks play an important role so that excluding any of them can drop the performance. Furthermore, the other finding is that the layers after Sinc6 have no meaningful added value so that the Sinc7 and following Conv1, pool, and the first Dense layers can be eliminated with no cost. These findings are also supported by the UMAP visualization. Fig. 3 (A-G) provides the UMAP of the outputs of each Sinc blocks for the test dataset. The blue and red dots correspond to the NQS and QS segments respectively. In these maps, we observe that the Sinc blocks can be categorized into three groups: (Sinc 1,2,3) extracts local features with receptive fields of 0.2 to 4.2s (such as EEG bursts) and no clear class discrimination, (Sinc 4,5) combines the local features with bigger receptive fields of 4.3 to 9.1s, and finally (Sinc 6,7) enables the class separation using predominantly

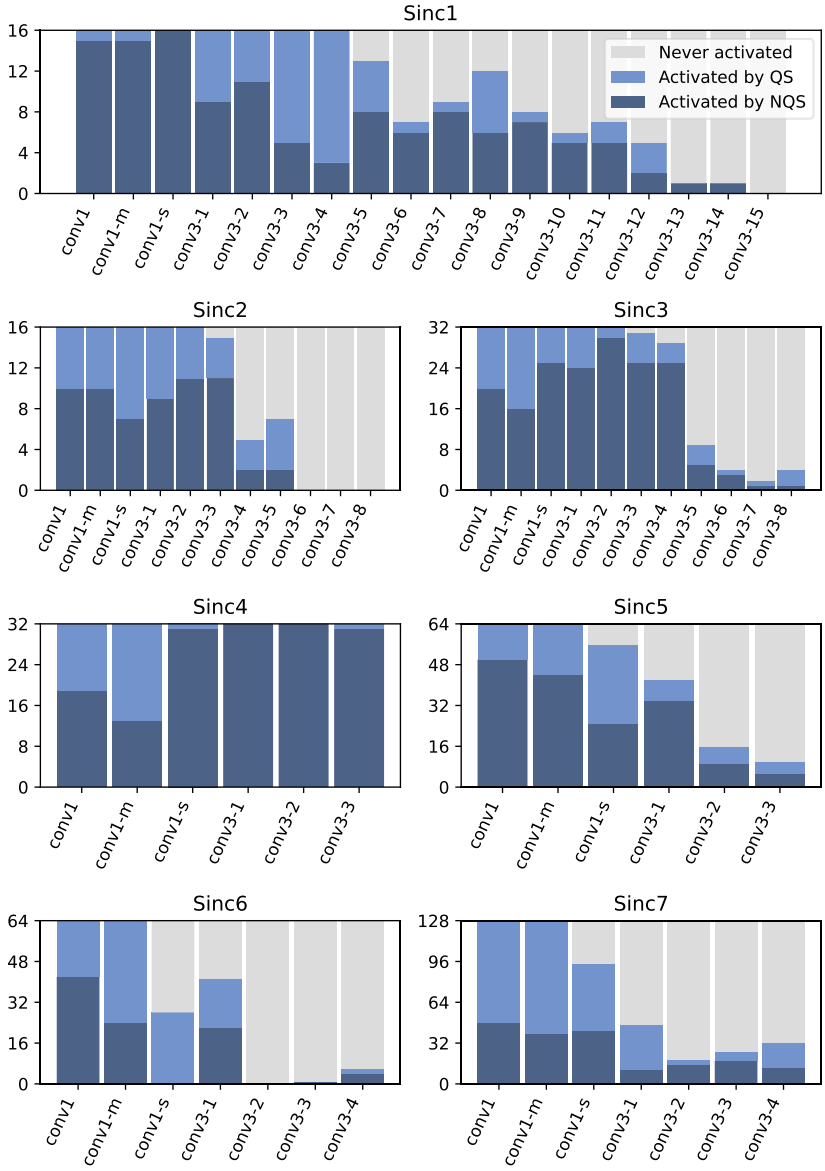

Fig. 4. Activation of the neurons in the Sinc blocks. In each graph, conv1, conv1-m, and conv1-s respectively correspond to the projection convolution, the projection convolution before the maxpool, and the projection convolution before the shared convolutions (see also Fig. 1). The navy blue, light blue and grey bars correspond to the number of neurons that were activated by non-quiet sleep, quiet sleep, and none, respectively.

more global features. Similarly to the results of the ablation study, it seems that Sinc7 does not increase the discrimination. In the last graph (H), the colors indicate three different PMA groups (<32w, 32w-44w, >44w). This reveals that the extremely premature babies (<32 PMA weeks) have the most difficult EEG segments to discriminate.

Fig. 4 presents the number of feature-maps activated by QS and NQS for each neuron (node) in each Sinc block. Here, 'conv1', 'conv1-m', and 'conv1-s' correspond to the isolated projection convolution, the projection convolution before the maxpool, and the projection convolution before the shared filters, respectively (see also Fig. 1). The y-axis denotes the number of feature-maps (N). The dead neurons always produced negative values and were counted as being 'never activated'.

IV. Discussion

In this study, a novel deep learning architecture for automated QS detection is proposed. The proposed network can successfully outperform the state-of-the-art single-scale approaches trained and tested on the same database. This proposed multi-scale CNN can be seen as an extension of the well-known Inception networks. Inception networks have been shown as a successful classifier in computer vision problems. Although the multi-scale characteristic of Inception seems useful for EEG analysis, it had not been previously used in this domain to the best of our knowledge. The main reason, as mentioned, may be the fact that the original structure of the Inception network is not well suited for time-series with very noisy patterns, such as EEG. In this paper, we increased the number of convolutional layers in the Sinc block to increase the

number of receptive fields, while sharing the parameters of the convolutional filters to reduce the number of trainable parameters and help to improve model generalization (reduce overfitting risk). As a result, in Table II, it was shown that the proposed Sinc significantly outperforms the other algorithms, including the Sinc network without filter sharing (i.e. Inc). Furthermore, comparing across a different number of channels showed that with more spatial information (8 or 4 channels), the Sinc model has only marginally better performance because the spatial characteristics of sleep-related patterns, such as EEG bursts and 'inter-bursts' (the low-amplitude suppression of the EEG between successive bursts), are also extractable by the alternative single-scale networks. However, when decreasing the number of EEG channels to only 1 or 2, temporal information must be solely used to detect such patterns. In this case, the superiority of the proposed multi-scale Sinc is clear.

Another interesting finding from Table II is that the performances of the considered methods with 2 channels are not necessarily better than with 1 channel (and sometimes even worse). This seems somewhat counterintuitive. In the 2-channel configuration, C3-Cz and C4-Cz are used (as Cz is the reference electrode) while, in the 1-channel case, a bipolar C3-C4 is used. It seems that Cz does not provide extra information for this task and, therefore, the simpler model (1-channel) with less parameters performs better.

In the UMAPs of Fig. 3 and Table III, we observe how the two sleep stages (QS and NQS) are separated block-by-block and to what extent the derived features depend on PMA. It is also shown that the most difficult segments for this classification are from the extremely premature infants (<32 weeks PMA). There are two possible reasons: 1) the biological fact that the neural sleep organization system is still developing in these babies *[1]* and, therefore, the corresponding EEG sleep patterns do not manifest clearly in this age group, and 2) the methodological fact that the number of training samples for this age group is smaller than the middle-age group. Nevertheless, the group with the oldest age (>44 PMA weeks) have almost the same data size as the <32 weeks group but exhibit a much better discrimination.

In the original inception block, the number of filters for each of the 1×1, 3×3, and 5×5 kernels are a separate hyperparameter and should be chosen before training. Although the network is not very sensitive to these numbers, a small or large value can lead to underfitting or overfitting, respectively. An ideal solution could be to optimize these hyperparameters using a big validation dataset. However, in the Sinc model with this limited database, we use an equal number of filters (defined by N) in all convolutions including the projections. This reduces the number of hyperparameters in the Sinc block and made the domain less diverse for designing and tuning the model. This said, a possible limitation is that it can cause redundant computations. Fig. 4 reveals that there are some dead filters in this network particularly when the receptive field increases. This did not cause overall adverse performance in the model but, in the presence of a bigger validation set, a further pruning could exclude such dead neurons to increase the computational efficiency and potentially aid generalization.

Another finding from Fig. 4 is that the number of neurons activated by the NQS inputs are greater than those activated by QS in the first Sinc blocks. However, in the latter blocks, this is the opposite. This could be due to the fact that QS is mainly constructed by cycles of burst and inter-burst interval patterns. In order to extract this cyclic pattern, a wider scale is needed to capture at least two bursts. This longer timescale is only provided in latter Sinc blocks, such that excluding the final Sinc (Sinc7) block decreases Kappa in both the validation and test datasets.

A limitation of this study is that since the number of recordings are limited and deep learning generally needs large amounts of training data, a minimum size for the validation set is chosen. This prevents further hyperparameter optimization. The hyperparameters used in this proposed Sinc network (M, N, RFs, number of blocks, and dropout ratios), are selected largely based on intuition. Nevertheless, we attempted to keep all benchmarked networks under similar conditions, with identical pre/post processing, overlapping, training, early stopping etc. routines to ensure a fair comparison. In the future, incorporating more data (when available) to facilitate extra hyperparameter optimization may increase the performance even further. In addition, extended techniques such as using an ensemble model or incorporating data augmentation could have additional improvements on the results.

V. Conclusion

The purpose of this study is to present a new algorithm that can satisfactorily detect neonatal QS from preterm EEG recordings. To this end, we suggest a novel multi-scale convolutional neural network based on the newly-introduced Sinc block. In this proposed network, a variety of features are extracted across different EEG temporal scales at each layer. In order to decrease the number of parameters, the filters of those convolutions are shared. We show that this network significantly outperforms the state-of-the-art algorithms across all reduced channel arrangements. Further research will explore other capabilities of such a multi-scale deep network for supervised and unsupervised EEG problems including sleep staging in older age groups and seizure detection. Crucially, the high accuracy of the Sinc network with only a single-channel EEG (often the only existing EEG configuration in many centres) makes it primed for translation into a real-time clinical assessment tool.

Appendix A – Conv-LSTM layers

In Table IV, more details about the architecture of the considered Conv-LSTM are provided.

Appendix B – AUC and Accuracy

In Table V and Table VI, the performance of the benchmarked networks is respectively presented with the Area Under the Curve (AUC) and Accuracy metric.

TABLE IV
THE LAYERS OF THE CONV-LSTM NETWORK

Layer/ Block	Output shape	Parameters
Input	1920 * 8	-
Conv1D	1920 * 32	k: 8, s: 1, N: 32
Maxpool [BN]	960 * 32	k: 2, s: 2
Reshape	30 * 32 * 32	
ConvLSTM [BN]	30 * 32 * 32	k: 3, s: 1, N: 32
ConvLSTM [BN]	30 * 32 * 32	k: 3, s: 1, N: 32
Maxpool	30 * 16 * 32	k: 2, s: 2
ConvLSTM [BN]	30 * 16 * 32	k: 3, s: 1, N: 32
ConvLSTM [BN]	30 * 16 * 32	k: 3, s: 1, N: 32
Conv2D [BN]	28 * 14 * 64	k: 3, s: 1, N: 64, vld
Maxpool	14 * 7 * 32	k: 2, s: 2
Conv2D [BN]	12 * 5 * 64	k: 3, s: 1, N: 64, vld
Maxpool	6 * 2 * 64	k: 2, s: 2
Flatten	768	
Dropout	768	$\alpha = 25\%$
Dense	20	-
Dropout	20	$\alpha = 20\%$
Dense	2	softmax

BN: followed by a batch-normalizer
vld: valid size without zero-padding
N: number of filters
k: kernel size
s: stride

TABLE V
THE AUC VALUES FOR DIFFERENT NETWORKS

	Area under the curve (AUC)			
Method	8-ch	4-ch	2-ch	1-ch
METD [13]	0.85	NA	NA	NA
CLASS [6]	0.92	NA	NA	NA
FBD [23]	0.94	NA	NA	NA
Conv-2D [23]	0.93	NA	NA	NA
Conv-2D [24]	0.95	0.94	0.91	0.86
Conv-1D [16]	0.90	0.93	0.89	0.74
Conv-1D [*]	0.94	0.94	0.92	0.92
Conv-LSTM	0.91	0.93	0.88	0.93
Dilated	0.92	0.94	0.89	0.88
Dense	0.95	0.95	0.94	0.94
Inc	0.96	0.95	0.95	0.94
Sinc	0.95	0.96	0.95	0.96

[*] a modified version of [24] with an early EEG integration

TABLE VI
THE ACCURACY VALUES FOR DIFFERENT NETWORKS

	Area under the curve (AUC)			
Method	8-ch	4-ch	2-ch	1-ch
METD [13]	0.79	NA	NA	NA
CLASS [6]	0.87	NA	NA	NA
FBD [23]	0.89	NA	NA	NA
Conv-2D [23]	0.88	NA	NA	NA
Conv-2D [24]	0.90	0.88	0.87	0.84
Conv-1D [16]	0.85	0.89	0.84	0.61
Conv-1D [*]	0.89	0.87	0.85	0.84
Conv-LSTM	0.88	0.90	0.84	0.89
Dilated	0.89	0.88	0.84	0.84
Dense	0.87	0.89	0.89	0.88
Inc	0.91	0.90	0.90	0.89
Sinc	0.91	0.91	0.90	0.91

[*] a modified version of [24] with an early EEG integration

REFERENCES:

[1] A. Dereymaeker *et al.*, "Review of sleep-EEG in preterm and term neonates," *Early Hum. Dev.*, vol. 113, pp. 87–103, 2017, doi: 10.1016/j.earlhumdev.2017.07.003.

[2] P. J. Cherian, R. M. Swarte, and G. H. Visser, "Technical standards for recording and interpretation of neonatal electroencephalogram in clinical practice," *Ann. Indian Acad. Neurol.*, vol. 12, no. 1, pp. 58–70, 2009, doi: 10.4103/0972-2327.48869.

[3] J. S. Barlow, "Computer characterization of tracé alternant and REM sleep patterns in the neonatal EEG by adaptive segmentation—an exploratory study," *Electroencephalogr. Clin. Neurophysiol.*, vol. 60, no. 2, pp. 163–173, Feb. 1985, doi: 10.1016/0013-4694(85)90024-0.

[4] V. Krajca, S. Petranek, K. Paul, M. Matousek, J. Mohylova, and L. Lhotska, "Automatic Detection of Sleep Stages in Neonatal EEG Using the Structural Time Profiles," in *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, Jan. 2005, pp. 6014–6016, doi: 10.1109/IEMBS.2005.1615862.

[5] V. Krajca, S. Petránek, J. Mohylová, K. Paul, V. Gerla, and L. Lhotská, "Neonatal EEG Sleep Stages Modelling by Temporal Profiles," in *Computer Aided Systems Theory – EUROCAST 2007*, 2007, pp. 195–201.

[6] A. Dereymaeker *et al.*, "An Automated Quiet Sleep Detection Approach in Preterm Infants as a Gateway to Assess Brain Maturation," *Int. J. Neural Syst.*, vol. 27, no. 6, p. 1750023, Sep. 2017, doi: 10.1142/S012906571750023X.

[7] V. Gerla, M. Bursa, L. Lhotska, K. Paul, and V. Krajca, "Newborn sleep stage classification using hybrid evolutionary approach," *Int J Bioelectromagn*, vol. 9, no. 1, pp. 25–26, 2007.

[8] N. Koolen *et al.*, "Automated classification of neonatal sleep states using EEG," *Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol.*, vol. 128, no. 6, pp. 1100–1108, 2017, doi: 10.1016/j.clinph.2017.02.025.

[9] M. Lavanga *et al.*, "Automatic quiet sleep detection based on multifractality in preterm neonates: Effects of maturation," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2017, pp. 2010–2013, doi: 10.1109/EMBC.2017.8037246.

[10] L. Fraiwan, K. Lweesy, N. Khasawneh, M. Fraiwan, H. Wenz, and H. Dickhaus, "Time Frequency Analysis for Automated Sleep Stage Identification in Fullterm and Preterm Neonates," *J. Med. Syst.*, vol. 35, no. 4, pp. 693–702, Aug. 2011, doi: 10.1007/s10916-009-9406-2.

[11] O. De Wel *et al.*, "Complexity Analysis of Neonatal EEG Using Multiscale Entropy: Applications in Brain Maturation and Sleep Stage Classification," *Entropy*, vol. 19, no. 10, Art. no. 10, Oct. 2017, doi: 10.3390/e19100516.

[12] L. Fraiwan and K. Lweesy, "Newborn sleep stage identification using multiscale entropy," in *2nd Middle East Conference on Biomedical Engineering*, Feb. 2014, pp. 361–364, doi: 10.1109/MECBME.2014.6783278.

[13] O. De Wel, M. Lavanga, A. Caicedo, K. Jansen, G. Naulaers, and S. Van Huffel, "Decomposition of a Multiscale Entropy Tensor for Sleep Stage Identification in Preterm Infants," *Entropy*, vol. 21, no. 10, Art. no. 10, Oct. 2019, doi: 10.3390/e21100936.

[14] K. Pillay, A. Dereymaeker, K. Jansen, G. Naulaers, S. Van Huffel, and M. De Vos, "Automated EEG sleep staging in the term-age baby using a generative modelling approach," *J. Neural Eng.*, vol. 15, no. 3, p. 036004, Jun. 2018, doi: 10.1088/1741-2552/aaab73.

[15] H. Cecotti and A. Graser, "Convolutional neural networks for P300 detection with application to brain-computer interfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 433–445, 2011.

[16] A. H. Ansari, P. Cherian, A. Caicedo Dorado, G. Naulaers, M. De Vos, and S. Van Huffel, "Neonatal Seizure Detection Using Deep Convolutional Neural Networks," *Int. J. Neural Syst.*, no. accepted, 2018, doi: 10.1142/S0129065718500119.

[17] L. Deckers, N. Das, A. H. Ansari, A. Bertrand, and T. Francart, "EEG-based detection of the attended speaker and the locus of auditory attention with convolutional neural networks," *bioRxiv*, p. 475673, Dec. 2018, doi: 10.1101/475673.

[18] U. B. Baloglu, M. Talo, O. Yildirim, R. S. Tan, and U. R. Acharya, "Classification of myocardial infarction with multi-lead ECG signals and deep CNN," *Pattern Recognit. Lett.*, vol. 122, pp. 23–30, May 2019, doi: 10.1016/j.patrec.2019.02.016.

[19] R. Donida Labati, E. Muñoz, V. Piuri, R. Sassi, and F. Scotti, "Deep-ECG: Convolutional Neural Networks for ECG biometric recognition,"

Pattern Recognit. Lett., vol. 126, pp. 78–85, Sep. 2019, doi: 10.1016/j.patrec.2018.03.028.

[20] H. Phan, F. Andreotti, N. Cooray, O. Y. Chén, and M. De Vos, "SeqSleepNet: End-to-End Hierarchical Recurrent Neural Network for Sequence-to-Sequence Automatic Sleep Staging," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 3, pp. 400–410, Mar. 2019, doi: 10.1109/TNSRE.2019.2896659.

[21] L. Fraiwan and M. Alkhodari, "Neonatal sleep stage identification using long short-term memory learning system," *Med. Biol. Eng. Comput.*, vol. 58, no. 6, pp. 1383–1391, Jun. 2020, doi: 10.1007/s11517-020-02169-x.

[22] H. Ghimatgar *et al.*, "Neonatal EEG sleep stage classification based on deep learning and HMM," *J. Neural Eng.*, vol. 17, no. 3, p. 036031, 2020.

[23] A. H. Ansari *et al.*, "Quiet sleep detection in preterm infants using deep convolutional neural networks," *J. Neural Eng.*, vol. 15, no. 6, p. 066006, 2018, doi: 10.1088/1741-2552/aadc1f.

[24] A. H. Ansari *et al.*, "A convolutional neural network outperforming state-of-the-art sleep staging algorithms for both preterm and term infants," *J. Neural Eng.*, vol. 17, no. 1, p. 016028, Jan. 2020, doi: 10.1088/1741-2552/ab5469.

[25] A. M. Husain, "Review of neonatal EEG," *Am. J. Electroneurodiagnostic Technol.*, vol. 45, no. 1, pp. 12–35, 2005.

[26] B. H. Connolly, L. Dalton, J. B. Smith, N. G. Lamberth, B. McCay, and W. Murphy, "Concurrent Validity of the Bayley Scales of Infant Development II (BSID-II) Motor Scale and the Peabody Developmental Motor Scale II (PDMS-2) in 12-Month-Old Infants," *Pediatr. Phys. Ther.*, vol. 18, no. 3, pp. 190–196, Fall 2006, doi: 10.1097/01.pep.0000226746.57895.57.

[27] K. Pillay, A. Dereymaeker, K. Jansen, G. Naulaers, and M. De Vos, "Applying a data-driven approach to quantify EEG maturational deviations in preterms with normal and abnormal neurodevelopmental outcomes," *Sci. Rep.*, vol. 10, no. 1, Art. no. 1, Apr. 2020, doi: 10.1038/s41598-020-64211-0.

[28] J.-M. Alonso and Y. Chen, "Receptive field," *Scholarpedia*, vol. 4, no. 1, p. 5393, Jan. 2009, doi: 10.4249/scholarpedia.5393.

[29] A. H. Ansari *et al.*, "Improved multi-stage neonatal seizure detection using a heuristic classifier and a data-driven post-processor," *Clin. Neurophysiol.*, vol. 127, no. 9, pp. 3014–3024, Sep. 2016, doi: 10.1016/j.clinph.2016.06.018.

[30] C.-E. Kuo and S.-F. Liang, "Automatic stage scoring of single-channel sleep EEG based on multiscale permutation entropy," in *2011 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, Nov. 2011, pp. 448–451, doi: 10.1109/BioCAS.2011.6107824.

[31] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," 2016, pp. 2818–2826, Accessed: May 25, 2020. [Online]. Available: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html.

[33] Xiaoling Xia, Cui Xu, and Bing Nan, "Inception-v3 for flower classification," in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, Jun. 2017, pp. 783–787, doi: 10.1109/ICIVC.2017.7984661.

[34] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," 2017.

[35] F. Baldassarre, D. G. Morín, and L. Rodés-Guirao, "Deep Koalarization: Image Colorization using CNNs and Inception-ResNet-v2," *ArXiv171203400 Cs*, Dec. 2017, Accessed: Jun. 09, 2020. [Online]. Available: http://arxiv.org/abs/1712.03400.

[36] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *ArXiv Prepr. ArXiv151107289*, 2015.

[37] A. Piryatinska, G. Terdik, W. A. Woyczynski, K. A. Loparo, M. S. Scher, and A. Zlotnik, "Automated detection of neonate EEG sleep stages," *Comput. Methods Programs Biomed.*, vol. 95, no. 1, pp. 31–46, Jul. 2009, doi: 10.1016/j.cmpb.2009.01.006.

[38] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. Wong, and W. Woo, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," *ArXiv150604214 Cs*, Jun. 2015, Accessed: Jun. 09, 2020. [Online]. Available: http://arxiv.org/abs/1506.04214.

[39] F. Yu and V. Koltun, "Multi-Scale Context Aggregation by Dilated Convolutions," *ArXiv151107122 Cs*, Apr. 2016, Accessed: Apr. 30, 2021. [Online]. Available: http://arxiv.org/abs/1511.07122.

[40] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," *ArXiv160806993 Cs*, Jan. 2018, Accessed: Apr. 30, 2021. [Online]. Available: http://arxiv.org/abs/1608.06993.

[41] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *ArXiv14126980 Cs*, Jan. 2017, Accessed: May 27, 2020. [Online]. Available: http://arxiv.org/abs/1412.6980.

[42] J. L. Fleiss, B. Levin, and M. C. Paik, "Statistical methods for rates and proportions," in *Statistical Methods for Rates and Proportions*, John Wiley & Sons, Inc., 2003, pp. 598–626.

[43] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[44] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *ArXiv180203426 Cs Stat*, Dec. 2018, Accessed: Jun. 05, 2020. [Online]. Available: http://arxiv.org/abs/1802.03426.

[45] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*, Boca Raton, FL: Chapman&Hall. CRC press, 1993.

[46] M. D. Smucker, J. Allan, and B. Carterette, "A comparison of statistical significance tests for information retrieval evaluation," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, 2007, pp. 623–632.