KU LEUVEN

STADIUS
Center for Dynamical Systems, Signal Processing and Data Analytics

Citation/Reference	Hermans T., Smets L., Lemmens K., Dereymaeker A., Jansen K., Naulaers G., Zappasodi F., Van Huffel S., Comani S., De Vos M., ``A multi-task and multi-channel convolutional neural network for semi-supervised neonatal artefact detection'', Internal Report 23-04, ESAT-STADIUS, KU Leuven (Leuven, Belgium), 2023. Accepted for publication in Journal of Neural Engineering.
Archived version	Accepted manuscript before proofing and publication.
Published version	Klik hier als u tekst wilt invoeren.
Journal homepage	Journal of Neural Engineering
Author contact	your email tim.hermans@esat.kuleuven.be Klik hier als u tekst wilt invoeren.
Abstract	
IR	Klik hier als u tekst wilt invoeren.

(article begins on next page)

A multi-task and multi-channel convolutional neural network for semi-supervised neonatal artefact detection

Tim Hermans[1], Laura Smets[1,2], Katrien Lemmens[3,4], Anneleen Dereymaeker[3,4], Katrien Jansen[3,5], Gunnar Naulaers[3,4], Filippo Zappasodi[2,6,7], Sabine Van Huffel[1], Silvia Comani[2,6], Maarten De Vos[1,3]

[1]Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven, Leuven, Belgium.
[2]Department of Neuroscience, Imaging and Clinical Sciences, G. d'Annunzio University of Chieti–Pescara, Chieti, Italy.
[3]Department of Development and Regeneration, UZ Leuven, Leuven, Belgium.
[4]Neonatal Intensive Care Unit, UZ Leuven, Leuven, Belgium.
[5]Child Neurology, UZ Leuven, Leuven, Belgium.
[6]Behavioral Imaging and Neural Dynamics Center, G. d'Annunzio University of Chieti–Pescara, Chieti, Italy.
[7]Institute for Advanced Biomedical Technologies, G. d'Annunzio University of Chieti–Pescara, Chieti, Italy.

E-mail: tim.hermans@esat.kuleuven.be

Abstract

Objective: Automated artefact detection in the neonatal electroencephalogram (EEG) is crucial for reliable automated EEG analysis, but limited availability of expert artefact annotations challenges the development of deep learning models for artefact detection. This paper proposes a semi-supervised deep learning approach for artefact detection in neonatal EEG that requires few labelled data by training a multi-task convolutional neural network (CNN).
Approach: An unsupervised and a supervised objective were jointly optimised by combining an autoencoder and an artefact classifier in one multi-output model that processes multi-channel EEG inputs. The proposed semi-supervised multi-task training strategy was compared to a classical supervised strategy and other existing state-of-the-art models. The models were trained and tested separately on two different datasets, which contained partially annotated multi-channel neonatal EEG. Models were evaluated using the F1-statistic and the relevance of the method was investigated in the context of a functional brain age prediction model.
Main results: The proposed multi-task and multi-channel CNN methods outperformed state-of-the-art methods, reaching F1 scores of 86.2% and 95.7% on two separate datasets. The proposed semi-supervised multi-task training strategy was shown to be superior to a classical supervised training strategy when the amount of labels in the dataset was artificially reduced. Finally, we found that the error of a brain age prediction model correlated with the amount of automatically detected artefacts in the EEG segment.
Significance: Our results show that the proposed semi-supervised multi-task training strategy can train CNNs successfully even when the amount of labels in the dataset is limited. Therefore, this method is a promising semi-supervised technique for developing deep learning models with scarcely labelled data. Moreover, a correlation between the error of functional brain age estimates and the amount of detected artefacts in the corresponding EEG segments indicates the relevance of artefact detection for robust automated EEG analysis.

Keywords: neonatal EEG, artefact detection, deep learning, convolutional neural network, semi-supervised learning, multi-task learning

1. Introduction

Automated analysis of neonatal electroencephalography (EEG) is an active field of research, which has the potential to support clinical decision making during continuous brain monitoring in neonates admitted to the Neonatal Intensive Care Unit (NICU). In recent years, numerous algorithms have been developed to analyse neonatal EEG in an automated manner, including algorithms for seizure detection [1–4], sleep classification [5–7], stress quantification [8], and functional brain age estimation [9–11]. By performing such automated analyses at the bedside in the NICU, the vulnerable neonatal brain can be monitored more efficiently and objectively, which can guide clinical care and predict short- and long-term (neurodevelopmental) outcomes [12–14].

One factor that complicates the development and application of such automated analyses is noisy artefacts in the EEG. Especially during continuous brain monitoring as done in the NICU, where EEG can be monitored for several hours, it is inevitable that there are segments in which artefacts contaminate and dominate the EEG signal, obscuring the actual brain signal. In neonatal EEG, movements are an example of a major source of recurring transient artefacts, but artefacts can also originate from other external or internal sources, such as eye movements, cardiac interference, faulty electrodes and electrode impedance [15]. During the development of algorithms for automated EEG analyses, experts typically visually inspect the EEG to remove artefactual EEG segments and extract only the clean EEG segments for further processing. Therefore, most analysis algorithms require that the input signals are clean and free from artefacts. However, when applying such an analysis algorithm to a novel, entire EEG recording, the accuracy and reliability of this algorithm may be reduced since it is unlikely that the entire recording would be artefact-free [16–18]. Manually preselecting clean EEG data for automated analysis is not a preferred solution to this problem, as this breaks the automated pipeline. Instead, automated identification of artefacts in neonatal EEG, and thereby the identification of clean, artefact-free EEG is a crucial step towards a robust and reliable application of fully automated EEG analyses.

Two main approaches are generally used to deal with artefacts in EEG: artefact detection and artefact correction [19]. Artefact detection algorithms identify data segments that contain artefacts. With such a method, segments with artefacts can be excluded from further analysis. Alternatively, artefact correction algorithms first identify artefacts and then additionally filter out these artefacts from the signal to reconstruct the artefact-free signal. Therefore, artefact correction methods are useful in situations where every part of an EEG recording is to be analysed, whereas artefact detection methods are useful when it can be afforded to discard some parts of the recording. An additional difference is that artefact correction algorithms typically focus on only one specific type of artefact with well-known dynamics and characteristics, whereas artefact detection can more easily cover several different types of artefacts. Given that continuously recorded neonatal EEG is generally long and can contain several types of artefacts, identification of clean EEG epochs using artefact detection methods is preferred over artefact correction methods.

Various methods have been proposed in the literature for detecting artefacts in neonatal EEG. These methods typically use a supervised machine learning approach, where artefacts in the data are labelled by experts to train the machine learning algorithm. For example, one common approach is to compute a specific set of features for an EEG segment and then classify that segment as artefact or clean using a feature-based classification model such as a support vector machine (SVM) [20,21]. In more recent developments, the feature extraction step has been incorporated in the machine learning process with deep learning models [22]. By incorporating this feature extraction in the training process, deep learning models can outperform feature-based models. Furthermore, deep learning models are typically faster at test time compared to feature-based models because of efficient parallel vector multiplications.

One common challenge in machine learning in medical applications is the limited availability of labelled data to train machine learning models. In neonatal EEG, artefact annotations are also limited

because they require time and expertise, whereas many unlabelled neonatal EEG data are available. In supervised machine learning models, this unlabelled data is left unused. In contrast, semi-supervised machine learning modes can learn from both labelled and unlabelled data. One of the few examples of semi-supervised methods for detecting artefacts in neonatal EEG is the Gaussian mixture model (GMM) proposed by Kauppila et al. [23]. This GMM is a feature-based semi-supervised method that learns from unlabelled data by finding clusters in the feature space.

To our knowledge, deep learning has not yet been applied for semi-supervised detection of artefacts in neonatal EEG, despite the versatility and popularity of unsupervised and semi-supervised deep learning frameworks, such as (variational) autoencoders, generative adversarial networks and pseudo-labelling methods [24]. Nonetheless, semi-supervised deep learning has been used for other EEG applications. For example, Wen et al. proposed a convolutional neural autoencoder for unsupervised feature extraction in adult EEG [25]. They showed that the features learnt by the unsupervised autoencoder could be used to train a supervised classification algorithm for detecting seizures in adults. A limitation of their method is that the feature learning and the classification tasks are optimised separately using unsupervised and supervised training, respectively. Therefore, the available labels for the supervised task are not exploited for feature learning. Another semi-supervised approach that could work for artefact detection is the deep semi-supervised anomaly detection framework proposed by Ruff et al. Here, labelled and unlabelled data are combined to train a deep anomaly detection algorithm [26]. Although this last method has not yet been applied to EEG data, it could be a suitable framework for semi-supervised artefact detection, assuming that artefacts are anomalies.

The aim of this paper is to develop a novel semi-supervised deep learning method able to detect artefacts in neonatal EEG. To this end, we propose a multi-task convolutional neural network that jointly optimises an unsupervised and supervised objective by combining an autoencoder (unsupervised) and a classifier (supervised) into a single multi-output network. The remainder of this paper describes this novel proposed semi-supervised method in more detail and compares its performance to several variations of the method and existing state-of-the-art methods. Finally, the relevance of the novel automated artefact detection is illustrated for the application of functional brain age estimation.

2. Materials and methods

2.1 Datasets

In this study, two separate neonatal EEG datasets (D1 and D2) were used to test the proposed method. Dataset D1 was obtained from our own research group, while D2 is an external publicly available dataset that we used to further validate our method. We did not merge these two datasets, but instead we trained artefact detection models for both datasets separately to investigate whether the proposed method works on datasets other than our own.

2.1.1 D1

Dataset 1 (D1) consists of 329 multi-channel EEG recordings of 133 preterm neonates with gestational age (GA) ranging from 23.86 weeks to 33.86 weeks were used in the current study. Of each neonate, one up to four EEG recordings were obtained at different times during their stay in the NICU of the University Hospitals Leuven (Belgium). This resulted in 329 recordings with post menstrual ages (PMA) at time of recording ranging from 24.00 to 46.57 weeks with a median (Q1-Q3) PMA of 34.14 (32.00 - 38.00) weeks. The EEG data were collected using the Brain RT EEG recording system (OSG BVBA, Kontich, Belgium) with a sampling frequency of 250 Hz and eight electrodes (Fp1, Fp2, C3, C4, T3, T4, O1, O2) and Cz as reference electrode, following the guidelines of the international 10-20 system. The research was conducted in accordance with the principles embodied in the Declaration of Helsinki and in accordance with local

statutory requirements. The data were completely anonymised and no personal data were used. In total, the amount of single-channel EEG data was 25,098 hours (329 recordings, 8 channels, average duration 9.5 h).

Out of the 329 recordings, 21 EEGs from 15 different patients were (partially) annotated, identifying periods with artefacts and periods with clean EEG (see Figure 1(a) for an example). Annotations were made separately for each channel. On average, 75% of the annotated EEG was labelled as clean and 25% as artefact. Out of the 21 labelled EEG recordings, 11 were annotated by a clinical expert and the remaining 10 recordings by two non-expert authors.

For the development and validation of the algorithm, the dataset was split in a training, validation, and test set. No cross-validation was done, but instead we divided the 21 labelled recordings among these three sets as follows. The eleven recordings that were annotated by the expert were included in the test set, because those annotations are the most precise and therefore are most reliable for final evaluation. From the remaining labelled recordings, six were assigned to the training set and the other four to the validation set, ensuring that there was a fair split in labeller and a similar distribution of PMA. All remaining unlabelled recordings were assigned to either the training or validation set, with a training/validation ratio of 60/40. When making the split, we ensured that all recordings of the same patient were assigned to the same set, to prevent dependence among the training, validation and test sets. As a result, the total amount of single-channel EEG data assigned to the training, validation, and test sets (and the amount that was labelled) was 14994 h (42 h), 9300 h (33 h) and 804 h (64 h), respectively.

2.1.2 D2

Dataset 2 (D2) contains 79 multi-channel EEG recordings from 79 neonates with gestational ages ranging between 35 and 44 weeks and has been used before for neonatal artefact detection by Webb et al. [22,27]. Artefacts and clean segments were annotated by experts using an 18-channel bipolar montage (Fp2-F8, F8-T4, T4-T6, T6-O2, Fp1-F7, F7-T3, T3-T5, T5-O1, Fp2-F4, F4-C4, C4-P4, P4-O2, Fp1-F3, F3-C3, C3-P3, P3-O1, Fz-Cz, Cz-Pz). In contrast to D1, where only a small subset of recordings was labelled, in this second dataset, each recording was partially annotated. The types of artefacts that were identified are: device interference, muscle, movement, electrode, and biological rhythm artefacts. All types of artefacts were merged into one artefact class, analogous to D1. More details about this dataset are described in Webb et al. [22]. The dataset was obtained from https://github.com/LockyWebb/NeonatalEEGArtefactDetection.

Similarly to D1, we split the recordings in dataset D2 into a training, validation and test set. This was done by randomly assigning 60% of the recordings to the training set, 20% of the recordings to the validation set and the remaining 20% to the test set. As a result, the total amount of single-channel EEG data assigned to the training, validation, and test sets (and the percentage that was labelled) was 1162 h (171 h), 461 h (46 h) and 391 h (42 h), respectively.

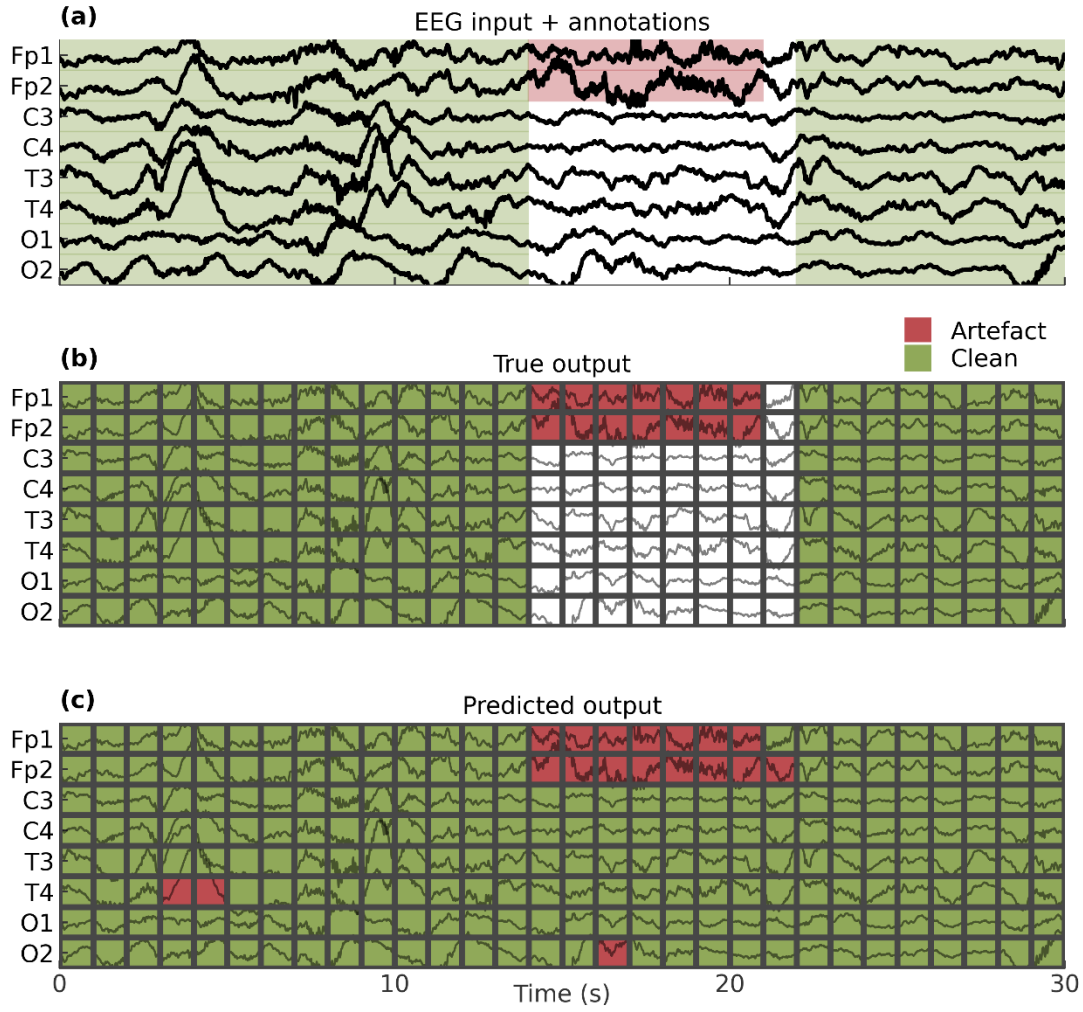

Figure 1 (a): Illustration of a typical input from D1 to the neural network; (b): the true (target) output; (c): the predicted output. The data and the annotations in (a) are divided into a grid of 30 seconds and 8 channels, where each second of each channel is assigned a label based on the annotations in (b). The network receives the full 30 seconds and all 8 channels simultaneously, but also makes predictions per second per channel, as visualised in (c). This is an example from the test dataset.

2.2 Pre-processing

The EEG was filtered using a notch filter (50 Hz for D1 and 60 Hz for D2) and a 0.27-30 Hz first-order Butterworth band-pass filter. The latter filter is identical to the one implemented in the software that was used by the expert who labelled the EEG for artefacts in D1. After filtering, the signals were downsampled to a frequency of 128 Hz using linear interpolation to reduce the size of the data, while still retaining the wave shape of the signal. A power of 2 was conveniently chosen for the sampling frequency as the neural network contains downsampling steps that reduce the time dimension by a factor 2, as explained in the next section. Subsequently, the EEG recordings were normalised by subtracting the mean and dividing by the standard deviation (SD) of the input signals in the training set. As explained in more detail in the remainder of this section, we have trained both single-channel models that are completely channel-independent and multi-channel models designed for one specific channel set-up. Considering that also the normalisation should be channel-independent and channel-dependent, the computation of the normalisation parameters (mean and SD) differed between the single- and multi-channel models. For the single-channel models, the normalisation parameters consisted of one channel-independent mean and SD computed after pooling all

channels. For the multi-channel models, the normalisation parameters consisted of the eight means and standard deviations computed for each channel separately.

The artefact and clean annotations were used to provide a label for each second and each channel of the EEG data. I.e., a 1-second grid was applied to the EEG data and each second of every EEG channel was assigned to one of three classes: artefact, clean, or unlabelled (see Figure 1(b)). A length of 1 second was chosen since this was considered a decent resolution to capture most typical artefacts in neonatal EEG. For each 1-second segment of single-channel EEG, if more than 50% (i.e., 0.5s) was annotated as clean or artefact, it was labelled as such. Additionally, 1-second segments that contained a short (<0.5s) artefact that started and ended within that same segment were also labelled as artefact. In all other cases, the segment was assigned to the unlabelled class. Figure 1 illustrates this segmentation and labelling.

2.3 Semi-supervised multi-task model

The proposed semi-supervised multi-task model consists of a convolutional neural network (CNN) autoencoder that learns to compress and decompress a 30-second EEG segment and a CNN classifier that predicts the locations of artefacts in the EEG segment. Instead of having two separate networks for the autoencoder and classifier, we merged them into one multi-output network, as schematically depicted in Figure 22. The network can be split into three parts: an encoder (blue), a decoder (red) and a classification part (green). The autoencoder consists of the encoder and decoder, and the classifier consists of the encoder and the classification layers. As is clear from the arrows in Figure 2, for each input of EEG, two outputs are generated: a reconstruction and a classification. Whereas the classification output is the output of interest during testing and application of the model, the reconstruction output makes it possible to incorporate unlabelled EEG data in the training process of the model. Therefore, please note that the reconstruction output of the autoencoder is not used to detect artefacts. Instead, the role of the autoencoder is merely to steer the training of the encoder. This is achieved by having the autoencoder and classifier share the layers of the encoder, and by training the autoencoder and classifier simultaneously, as explained in more detail in the section *Training.*

The architecture of the autoencoder (encoder + decoder) is inspired by the work of Wen et al. [25], which compresses the EEG into a latent representation using repeated convolution and pooling layers and decompresses the latent feature vector using deconvolutional layers. Even though the basic structure is similar to the one of the method proposed by Wen et al., our model contains several novel adaptations. The first adaptation is that instead of computing one latent vector characterising the entire 30-second input segment, our model computes one latent feature vector (of length 8) for each second of the input data to retain some temporal resolution. This is achieved by implementing 7 downsampling (max pooling) steps in the encoder, without any flattening layer at the end. Each downsampling step halves the resolution, therefore downsampling the 128 Hz signal 7 times by a factor 2 yields the desired latent resolution of 1 Hz. Therefore, a 30-second EEG input segment sampled at 128 Hz with C channels has shape $(30 * 128, C)$, and will be decoded into a latent representation with shape $(30, C, 8)$, i.e., each second and each channel of the EEG data is encoded by eight latent features. A symmetrical architecture is used for the decoder, applying (strided) transposed convolutional layers to upsample the compressed latent representation back to the original input size, aiming to reconstruct the input as close as possible. A second adaptation is that we changed the 1D convolutional layers in the encoder and decoder to 2D convolutional layers to make the network compatible with multi-channel inputs. Even though we use 2D kernels, the kernel size in the channel dimension was kept equal to 1, such that the autoencoder network processes (compress and decompress) each channel independently (i.e., the layers in the encoder and decoder are no spatial filters, but instead represent temporal filters that process multi-channel data in parallel). A third adaptation is the inclusion of one convolution layer with a wider kernel in order to incorporate temporal context (indicated by [1] in Figure 2).

As mentioned before, the architecture of the classifier starts with the encoder that is shared with the autoencoder. On top of this encoder, convolutional layers are added to classify each second of each channel as artefact or clean. Thus, the output of the classifier is a tensor with class probabilities with shape $(30, C, 2)$, corresponding to the number of input seconds, the number of channels and the two classes (artefact, clean), respectively. Two versions of the network were used: a single-channel network (CNN SS-MT) with $C = 1$, and a multi-channel network (CNN SS-MT MCh) with $C = 8$ for D1 and $C = 18$ for D2. In the single-channel case ($C = 1$), the 1x1 convolutional layers in the classifier can be regarded as fully-connected layers that act on the latent feature vector of each second of data independently. In the multi-channel case ($C > 1$), spatial information is exploited by means of a valid $(1, C)$ convolution (annotated by [2] in Figure 2). Here, the classification layers can be interpreted as fully-connected layers acting on the pooled feature vectors of all channels. A reshape layer is needed at the end to re-order the neurons to still obtain one prediction per second per channel. Note that the model does not return a single label, but instead returns a mask that predicts for each second and each channel whether it is artefact or clean, much like image segmentation models that predict a class for each pixel.

Although the model was trained on input segments with a fixed length of 30 seconds, in practice, the input time dimension is variable, which is possible due to the convolutional nature of the model. The same holds for the channel dimension in the single-channel model. This does not affect the training process but makes it possible to run the model on data of any length, and with any number of channels when making predictions. Furthermore, batch normalisation and spatial dropout layers were used after convolutional layers for regularisation. The models were built with Python using Tensorflow and Keras and the code has been made publicly available at https://gitlab.com/timhermans/artefact_detection_public [28,29].

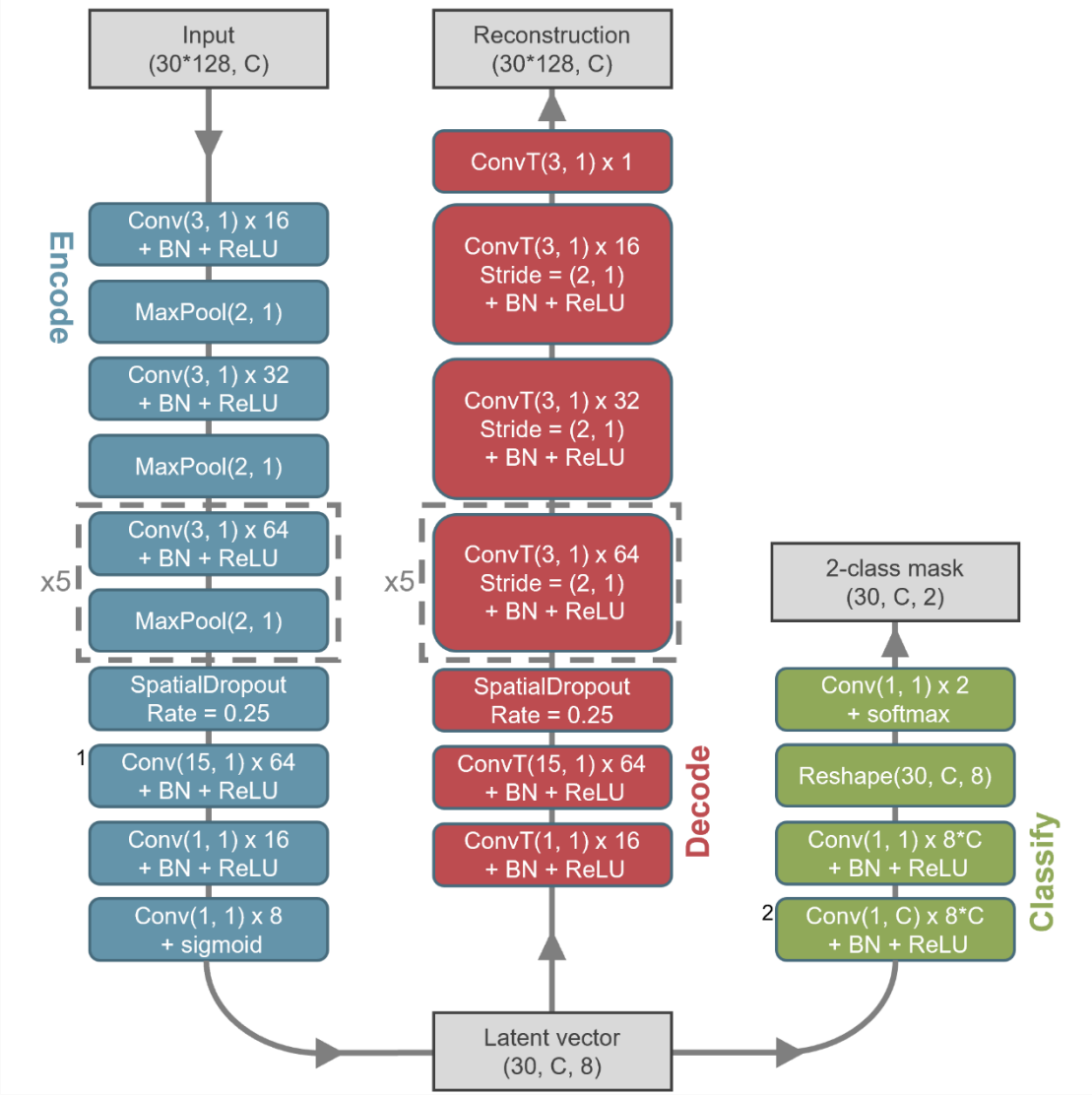

Figure 2 Architecture of the proposed multi-task model (for data sampled at 128 Hz). C is the number of channels. Conv(a, b) are 2D convolution layers with kernel size (a, b) and with "same" padding, except at [2] where padding is "valid". Here, a and b refer to the time and channel dimensions, respectively. ConvT are 2D transpose convolution layers. The reconstruction is used for the unsupervised autoencoder loss and the class probabilities for the supervised classification loss. [1]This layer captures the temporal context. [2]This is a valid convolution that captures the spatial context. BN: batch normalisation; ReLU: rectified linear unit activation function.

2.4 Training

As mentioned before, the proposed multi-task model learns to extract features using an unsupervised objective and to classify using a supervised objective. Whereas Wen et al. used a two-step approach, where they first trained the CNN autoencoder for unsupervised feature extraction and then used the latent features computed by the trained encoder to train a feature-based machine learning classifier, our model requires only one single end-to-end training stage by combining these two objectives in one loss function.

The loss function of the multi-task model (L_{MT}) consists of two parts, one for each task, i.e., autoencoder (unsupervised) and classification (supervised):

$$L_{MT} = L_{AE} + \eta L_C \qquad \text{Eq. 1}$$

Here, L_{AE} is the autoencoder loss, L_C is the classification loss and η is a hyperparameter scaling the classification loss. For the autoencoder, loss the mean squared error was used between the original signal ($\boldsymbol{x}$) and reconstructed signals ($\hat{\boldsymbol{x}}$):

$$L_{AE}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \frac{1}{NC} \sum_{i=1}^{N} \sum_{j=1}^{C} \left(x_{i,j} - \hat{x}_{i,j}\right)^2 \quad \text{Eq. 2}$$

Here, $x_{i,j}$ refers to the jth time sample of the ith channel in the input segment, and C and $N(= 30 * 128)$ are the number of channels and time samples in the input segment, respectively. For the classification loss the categorical cross-entropy was used:

$$L_C(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{2CT} \sum_{i=1}^{T} \sum_{j=1}^{C} \sum_{k=1}^{2} y_{i,j,k} \log(\hat{y}_{i,j,k}) \quad \text{Eq. 3}$$

Here, $\boldsymbol{y}_{i,j}$ is a 2-element vector with a one-hot-like encoding of the true labels for the jth second in the ith channel, $\hat{\boldsymbol{y}}_{i,j}$ is a 2-element vector with probabilities for the corresponding classes as predicted by the classifier, k is the class index, and $T(= 30)$ is the number of seconds in the input segment. Importantly, the encoding of the true labels in $\boldsymbol{y}_{i,j}$ were done as follows: [1, 0] for artefact, [0, 1] for clean and [0, 0] for unlabelled. The latter makes it possible to exclude unlabelled parts from the classification loss (as all terms in Eq. 3 are zero for unlabelled data).

For training the CNN models, the Adam optimiser was used with a learning rate of 0.001. The batch size was set to 1500 for the single-channel model, where one sample consists of one 30-second segment of 1-channel EEG data, and to 187 for the multi-channel model, where one sample consists of one 30-second segment of 8-channel EEG data. Batches were created by iterating over the shuffled samples of the pooled labelled and unlabelled data. Learning was stopped if the validation loss did not decrease for 7 epochs (early stopping). Given that the classification task is easier than the reconstruction task, only the autoencoder loss is activated for the initial 4 epochs (by setting η in the loss function of Eq. 1 to zero for the first 4 epochs). After the fourth epoch, the classification loss was added to the total loss, by specifying a non-zero η. The procedure for finding a good value for η was as follows. First, the autoencoder (L_{AE}) and classifier loss (L_C) were computed on a randomly initialized model and an initial guess for η was determined at the value that it brings both losses to a similar order of magnitude, i.e., $\eta_0 = L_{AE}/L_C$. Second, we did a random hyperparameter search, by trying values around η_0 and selecting the value that resulted in the best classification results on the validation data. Following this procedure, we set $\eta = 4000$ for D1 and $\eta = 50$ for D2.

2.5 Reference methods

Besides the proposed semi-supervised multi-task deep learning method, we applied some reference methods that we optimised for the described dataset, as we did for the proposed method. As a baseline, we applied two feature-based methods from previously published studies: a supervised support vector machine (SVM) and a semi-supervised Gaussian mixture model (GMM) [20] [23]. To maintain comparability, the same pre-processing was used as described above. Although the GMM paper contains some different features, we used the same feature set as in the SVM paper for better comparison of the training methods. A list of the 23 features and a more detailed description of these methods can be found in the Appendix (see supplementary materials).

As an alternative deep learning method for semi-supervised artefact detection, we applied the semi-supervised anomaly detection (SS-AD) method as proposed by Ruff et al. [26]. This method consists of two training steps. In the first training step, an autoencoder is trained on all available data (labelled and unlabelled). Then, all latent feature vectors of the clean data were averaged to define the centre of a

hypersphere ($\boldsymbol{c}$) in the latent dimension. In the second training step, the encoder from the trained autoencoder was taken and trained with the deep SS-AD loss as proposed in their paper [26]. Briefly, the objective of this loss function is to map clean and unlabelled samples as closely as possible to the defined hypersphere centre $\boldsymbol{c}$, while mapping labelled artefacts far away from the hypersphere centre. As in the classification loss, this loss is computed per second and per channel. We use the same architecture for the autoencoder as in Figure 2 with two modifications to prevent the hypersphere collapse as explained by Ruff et al., i.e., all bias terms (in the convolution layers) were removed and the sigmoid activation at the end of the encoder was replaced by a linear activation.

As a final reference method, we trained our proposed CNN classifier in a fully supervised fashion. Here, the model consists only of the encoder and the classification layers, and the model is trained on the labelled data only using the classifier loss function in Eq. 3. Like for the semi-supervised multi-task classifier, this supervised classification model was trained in a single-channel (CNN S) and multi-channel fashion (CNN S mCh).

2.6 Evaluation metrics

The models were evaluated based on the confusion matrix computed on the labelled portion of the data. Each second of single-channel EEG that was labelled was counted as one observation in the confusion matrix. This is the same for the single-channel and the multi-channel models, as the model makes predictions per second and per channel in both cases. The artefact class was considered the positive class, obtaining counts for true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). The models were compared based on the F1 score, which is the harmonic mean of sensitivity and precision and is identical to the Dice-coefficient from a segmentation point of view, which quantifies the percentage of overlap between the true and predicted artefact labels. Furthermore, we computed the accuracy (percentage of correct predictions), miss rate (percentage of artefacts missed) and false discovery rate (percentage of predicted artefacts that are actually clean):

$$F1 = \frac{2TP}{2TP + FP + FN} \times 100\% \quad \text{Eq. 4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad \text{Eq. 5}$$

$$Miss\ rate = \frac{FN}{TP + FN} \times 100\% \quad \text{Eq. 6}$$

2.7 Experiments

We first compare our proposed model to the reference methods. For each method, 10 models were trained with different random initializations and, for each method, the model with the highest F1 score on the validation data was selected as the final model. For these selected models, the evaluation metrics were computed on the test dataset and reported. This was done separately for the two datasets D1 and D2.

After training had finished, we computed the runtime of each model when making predictions on a new recording. To compute this runtime, we applied all models to 1 h of EEG data on the same PC and repeated this 10 times and took the mean of these 10 runs.

To investigate the effect of the amount of labelled data on the training of the different methods, we reduced the total amount of labelled seconds in the training and validation set by 50%, 75% or 87.5%. This was achieved by excluding a portion of annotations at the end of each recording in the training and validation set. All models were then trained from scratch on the new dataset with fewer labels. Importantly, the labels in the test set were not reduced to ensure a fair comparison since reducing the test data would not

affect the model itself, but instead would provide us with noisier estimates of its performance. Again 10 models were trained per method with different random initialisations and the best version was selected based on the maximum F1 score on the validation data. As before, this analysis was done for both datasets separately.

Finally, the relationship between the amount of detected artefacts in an EEG segment and the error made by a functional brain age (FBA) estimation model was investigated to show the relevance of the proposed automated artefact detection method. To this end, a deep shared multi-scale inception network model was applied to the EEG recordings to estimate the FBA [30,31]. This FBA model provides one estimate of FBA for every 30 seconds of multi-channel EEG data. Besides estimating the FBA for each segment, we applied the novel single- and multi-channel CNN SS-MT models to the EEG and computed the percentage of detected artefacts for each 30-second segment (artefact contamination). To test if there was a relationship between the level of artefact contamination and the FBA error, we assigned each 30-second EEG segment to one of 10 levels of artefact contamination: 0%-10%, 10%-20%, etc. Then, for each level of artefact contamination, the median FBA of the segments with that level of artefact contamination was computed. Finally, the error of these median FBAs was computed as the absolute difference between the median FBA and the PMA of the neonate at the time of the recording. This analysis was performed only for the 329 recordings in dataset D1, as the data from D2 was incompatible with the input requirements of the FBA model (different recording protocol).

3. Results

The test scores of the different approaches are presented in Table 1. It shows the accuracy, miss rate, false discovery rate and F1 scores for each of the two datasets, computed on the recordings in the corresponding test data. Confusion matrices for the multi-task models are provided in the appendix (see supplementary materials).

We first describe the results obtained with D1. Comparing the supervised single-channel models, the results show that the proposed deep learning model (CNN S) performs similarly to the feature-based baseline (SVM) with F1 scores of 74.9% and 76.2%, respectively. Looking at semi-supervised methods, the feature-based GMM performs worse than its fully supervised counterpart SVM, with an F1 score of 67.6%. Likewise, the semi-supervised anomaly detection algorithm (CNN SS-AD) performs significantly worse than its supervised counterpart (CNN S), with an F1 score dropping to 62.5%. In contrast, the proposed semi-supervised multi-task learning model (CNN SS-MT) has a similar performance as the supervised CNN and the supervised SVM (F1 is 74.7%). When going to the multi-channel implementation of our proposed network, we see a significant increase in performance towards F1 scores of 83.8% for the supervised model (CNN S mCh) and an even higher F1 score of 86.2% for the semi-supervised multi-task model (CNN SS-MT mCh).

The results obtained with D2 show that all CNN models (supervised and semi-supervised) perform similarly with F1 scores ranging between 94.0%-95.7%, with the semi-supervised models having the highest F1 scores. All CNN models outperform the feature-based GMM and SVM, which have F1 scores of 88.9% and 91.7%, respectively. Similarly to D1, the supervised feature-based method (SVM) outperforms the semi-supervised feature-based method (GMM). In contrast to D1, for D2 there is no significant improvement in performance when comparing the multi-channel to the single-channel CNN models.

Table 1 Performance metrics for the trained models per dataset (D1 and D2). Metrics are computed on the corresponding test set. SVM: Support Vector Machine (supervised, feature-based), GMM: Gaussian Mixture Model (semi-supervised, feature-based), CNN: convolutional neural network, S: supervised model, SS-AD: semi-supervised anomaly detection, SS-MT: semi-supervised multi-task model, mCh: multi-channel.

Model	**Accuracy (%)**		**Miss rate (%)**		**False discovery rate (%)**		**F1 (%)**	
	D1	**D2**	**D1**	**D2**	**D1**	**D2**	**D1**	**D2**
CNN SS-MT mCh	**96.6**	**97.3**	11.7	**3.4**	**15.8**	5.1	**86.2**	**95.7**
CNN S mCh	95.8	96.3	**9.8**	7.3	21.7	4.6	83.8	94.0
CNN SS-MT	93.5	97.2	19.9	6.7	30.0	**2.1**	74.7	95.5
CNN SS-AD	88.4	97.2	19.8	5.7	48.8	3.3	62.5	95.5
CNN S	94.0	96.7	19.6	8.4	27.5	**2.1**	76.2	94.7
GMM	91.6	93.3	26.6	13.9	37.3	8.1	67.6	88.9
SVM	93.4	94.7	17.8	5.3	31.1	11.1	74.9	91.7

Figure 3 shows the output of the single- and multi-channel approaches on one exemplary EEG epoch from the test set that is partly labelled. In Figure 3(a), the EEG data is shown and the labelled artefacts and clean parts are coloured in red and green, respectively. Figures 3(b) and 3(c) show the same EEG data, where the colours indicate the predictions by the single-channel and multi-channel model, respectively. Especially for the multi-channel model, there is a good agreement between the model predictions and the true labels.

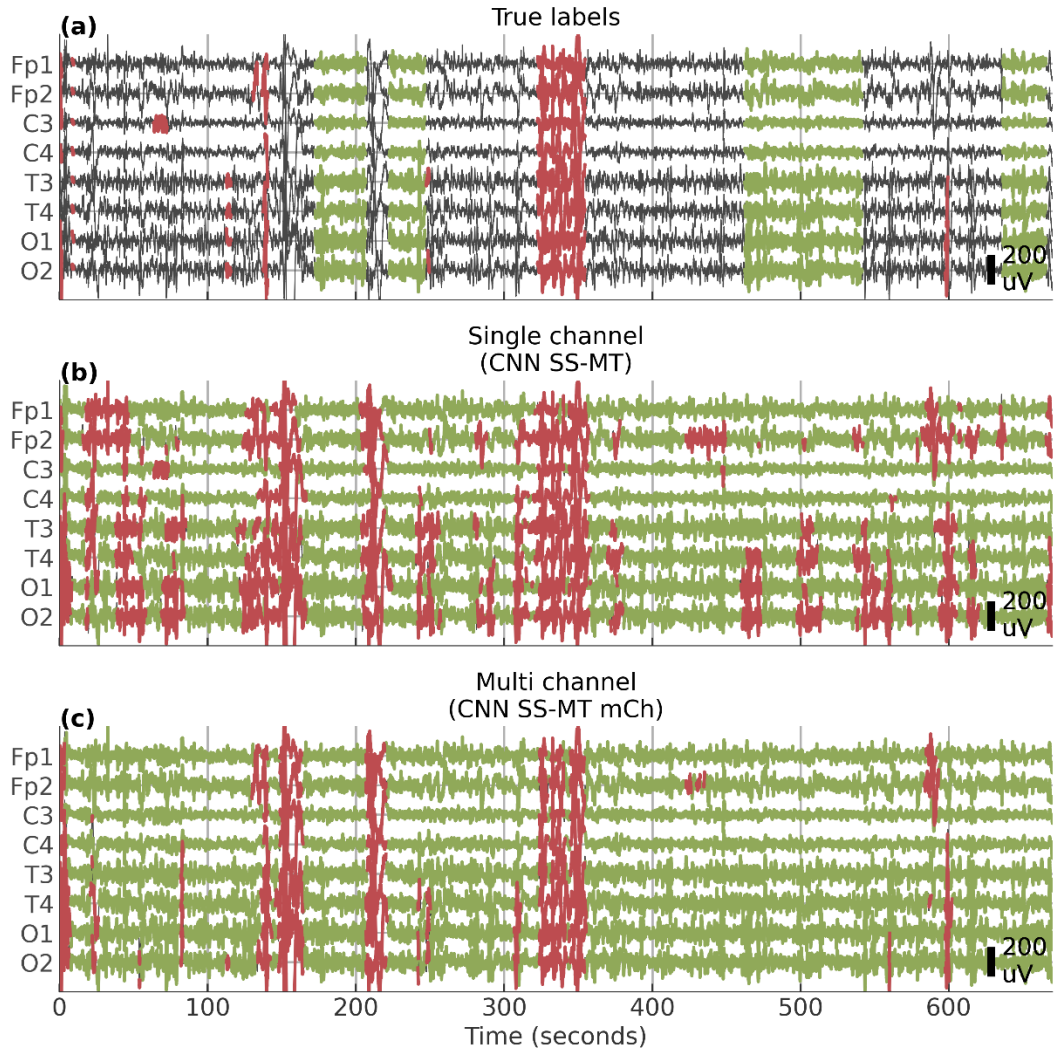

Figure 3 Example of the artefact detection models for a recording from the test set. (a): true annotations; (b): output of single-channel semi-supervised multi-task CNN model; (c): output of multi-channel semi-supervised multi-task CNN model. Black: unlabelled; red: artefact; green: clean.

Table 2 compares the runtimes of the models during test time. More specifically, it shows the time it takes to identify the artefacts in 1 h of 8-channel EEG. As expected, the CNN models are faster compared to the feature-based model (SVM and GMM). The runtime of the GMM is mainly occupied by the computation of the features. The SVM requires the same amount of time for feature extraction, but additionally requires a substantial amount of time to predict the output class from the extracted features.

Table 2 Test times of the models for processing 1 h of 8-channel EEG (from D1). The mean and standard deviation of 10 repetitions are reported.

Model	Run time (s)
CNN SS-MT mCh	2.9 ± 0.1
CNN S mCh	2.9 ± 0.2
CNN SS-MT	2.8 ± 0.2
CNN SS-AD	2.4 ± 0.2
CNN S	2.9 ± 0.2
GMM	29.3 ± 1.9
SVM	102.9 ± 3.2

Figure 4 shows the effect of the amount of labelled training data on the performance of the different models. Note that the F1 scores at 100% of labels used are the same as those in Table 1. Figure 4(a) shows the results obtained with D1. When excluding more and more labels, especially the fully supervised deep learning models (CNN S and CNN S mCh) drop in performance to an F1 below 30% when only 12.5% of the labels are used. In contrast, the feature-based and the semi-supervised deep learning methods are much less affected by a reduced amount of labelled data and yield models with F1 scores around 70%, even if only 12.5% of the available labels were used. For D2, the effect on reducing the number of labels is smaller since this dataset contains more labels to start with (Figure 4(b)). Nevertheless, with 12.5% of the labels used, the supervised multi-channel CNN (CNN S mCh) drops significantly in performance, while the F1 of the semi-supervised multi-channel CNN (CNN SS-MT mCh) remains near 90%.

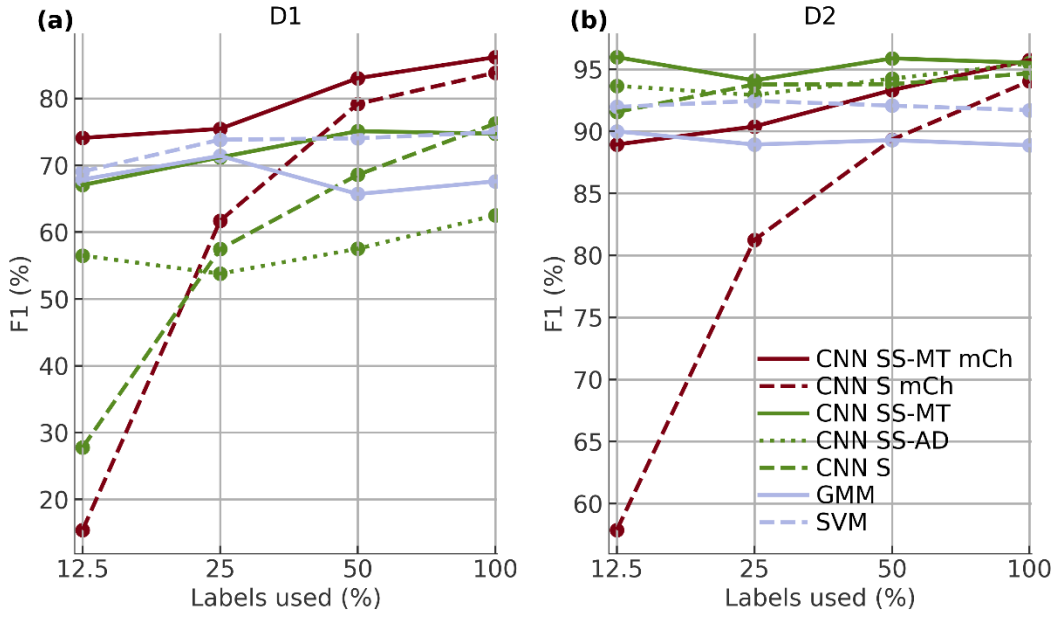

Figure 4 Performance of the models (F1 score) when less labelled data is used. (a): results for dataset D1; (b): results for dataset D2. CNN: convolutional neural network, mCh: multi-channel, S: supervised model, SS-MT: semi-supervised multi-task model, SS-AD: semi-supervised anomaly detection, GMM: Gaussian Mixture Model (semi-supervised, feature-based), SVM: Support Vector Machine (supervised, feature-based).

The last analysis investigates the relationship between the error of an FBA model and the level of artefact contamination (as defined by the percentage of artefacts in an EEG segment according to the automated

artefact detection model). Figure 5 clearly shows that, for both the single- and multi-channel CNN SS-MT models and despite their difference in classification performance, the error of the FBA estimation increases with the increasing level of artefact contamination in the EEG data.

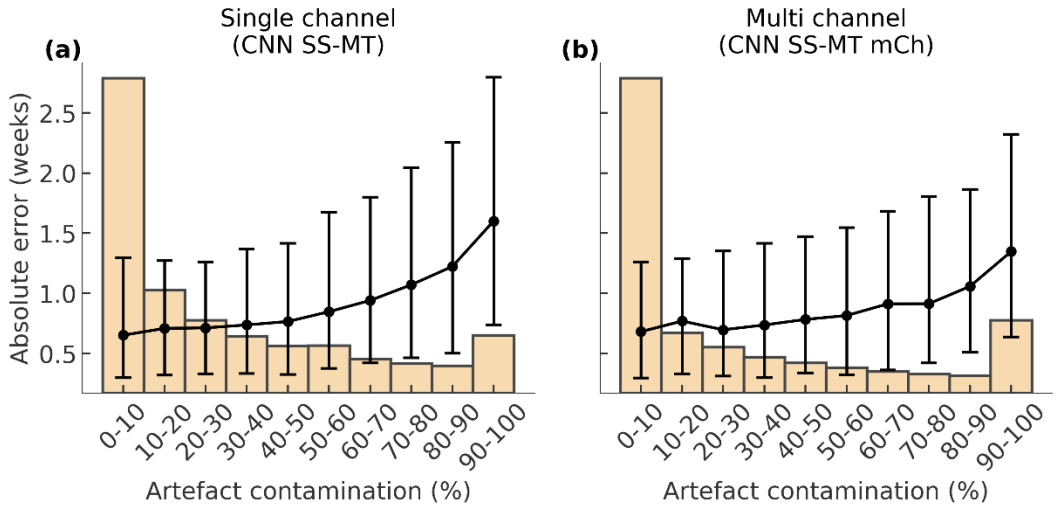

Figure 5 Relationship between artefact contamination of segments (as detected by an artefact detection model) and the error of a functional brain age (FBA) prediction model. (a): artefact contamination computed using the single- channel model (CNN SS-MT); (b): artefact contamination computed using the multi-channel model (CNN SS-MT mCh). The left-most point in each graph represents the error that the FBA model makes on EEG segments with very low amounts of detected artefacts (0%-10%). The right-most point in each graph is the error that the FBA model makes on EEG segments with very high amounts of detected artefacts (90%-100%). The points indicate the median and the error bars indicate the 25 and 75 percentiles among the recordings. The histogram in the background shows the proportional distribution of the artefact contamination levels among all 30-second segments.

4. Discussion

A novel deep multi-task artefact detection model was proposed that classifies each second of a neonatal EEG signal as either clean or artefact by using a convolutional neural network (CNN). By jointly training an autoencoder and a classifier that share the encoder, we could leverage the large amount of unlabelled neonatal EEG data that is typically available besides a limited labelled dataset. Our results show that the proposed CNN models can outperform state-of-the-art feature-based methods. Moreover, the multi-task CNN models outperform the supervised CNN models when limited labelled data is used for training.

The main novelty of our approach lies in the training of the network, that combines an unsupervised objective with a supervised objective. Wen et al. have proposed to train an EEG autoencoder for unsupervised feature learning and showed that the features extracted by such a trained autoencoder can be used as input for a classification model [25]. In that approach, the unsupervised feature learning and the supervised classification tasks are disjoint and therefore the feature extraction is solely learnt from unlabelled data. In contrast, we proposed a model that joins the feature learning and classification tasks together, requiring only one training process. In this way, deep feature extraction is learnt from both unlabelled and labelled data simultaneously, which enables the model to learn better features for the given supervised task in a semi-supervised way. With such a training strategy, the unsupervised task of the model (autoencoder) helps steer the training process of the supervised task (classification). Moreover, the unsupervised loss can be regarded as a sort of regularisation of the model that can prevent overfitting. In the future, the proposed multi-task approach could also be extended with other or additional unsupervised or supervised tasks. For neonatal EEG, there is namely a need for various models that perform different

tasks, such as sleep classification, seizure detection and functional brain age prediction. For each of these tasks, the availability of labelled data is limited, which could impede the development of a deep learning algorithm. A multi-task approach as presented in this paper, when configured properly, makes it possible to pool all the labelled datasets together and train one multi-output deep learning model that jointly learns to do the different tasks. In such a set-up, the shared part of the multi-output learns from a larger dataset, potentially improving accuracy and generalisability.

Besides the multi-task training, another novelty of the proposed artefact detection method lies in the architecture of the neural network. Whereas classifiers commonly map a single input (e.g., an EEG segment) to a single prediction (e.g., clean/artefact), the proposed classifier outputs one prediction for each second of EEG. Due to the convolutional nature of the model, the input EEG can have any length and the output resolution is not dependent on the input length. Furthermore, by using convolutional kernels that span several neighbouring time samples, temporal context (of about 15 seconds) is incorporated in our model (see the layer indicated by [1] in Figure 2). Making predictions on a per second basis instead of a per input basis is a major difference with another recently published supervised deep neural network proposed by Webb at el., who developed a residual CNN that predicts a single label for a 4-second EEG input [22]. To incorporate temporal context, Webb et al. smoothed the outputs of consecutive EEG segments using a moving average. In contrast, in our proposed model, the temporal resolution of the predictions is higher (1 every second vs. 1 every 4 seconds) and the temporal context is incorporated into the neural network.

We analysed our methods on two different datasets (D1 and D2) by training and testing the methods on each dataset. For D1, the single-channel semi-supervised model (CNN SS-MT) was on par with a recently published baseline model (a supervised feature-based SVM), which we tailored and retrained to our dataset and problem statement. The results obtained with D2 show that all CNN models outperformed the feature-based models. This indicates that deep learning models can outperform feature-based methods, depending on the dataset used. Besides this, the deep learning model has several other advantages. One practical advantage of the CNN over the SVM is its computational speed. On a laptop without GPU, processing 1 h of 8-channel EEG data took only 2.8±0.2 seconds using the CNN SS-MT method, compared to 102.9±3.2 seconds using the SVM method. Another advantage of deep learning models is that there is no need for feature-engineering, as the feature extraction is incorporated in the training and is thus completely data-driven. Moreover, deep learning is an active field of research in which many developments continue to be made, including semi-supervised training procedures such as reinforcement learning, pseudo-labelling and generative adversarial networks (GANs).

The versatility of the proposed network allowed for a straightforward extension to a multi-channel model, which outperformed all single-channel models when using D1. Please note that, for this multi-channel approach, although the information across the channels is shared in the classification layers, predictions are still made for each channel separately. This makes it possible for the model to analyse all channels in a multi-channel EEG segment at once and identify the channels that contain artefacts. The multi-channel approach performs better than the single-channel approach (see Table 1). Due to the exchange of channel information in the first layer of the classifier (annotated by [2] in Figure 2), we allow the model to learn cross-channel information. Such cross-channel information provides spatial context that is useful information for the task of artefact detection. E.g., artefacts often occur on all channels simultaneously (e.g., due to body movements), which possibly is the main reason for the improved accuracy of the multi-channel model. The downside of the multi-channel approach is that, once trained on a specific channel montage, it can only be applied to EEG data with that same montage. Therefore, further developments are needed if a multi-channel model is desired that works for any montage. Nonetheless, within one NICU, the routine EEG acquisition typically follows a fixed EEG protocol and montage, which minimises the necessity of a versatile multi-channel model.

Besides comparing our CNN methods to state-of-the-art methods, we compared our semi-supervised multi-task approach to a fully supervised CNN approach. The proposed semi-supervised CNN had similar performance scores as the fully supervised implementation of our CNN. We hypothesised that the absence of a significant improvement of the semi-supervised over the supervised method can be attributed to the large amount of labelled data that was available. More concretely, in D1, there were 150,529 seconds of labelled data in the training set, which might be sufficient to make the addition of unlabelled data redundant. This hypothesis was confirmed when we excluded a part of the training labels and observed a significant performance decrease in the supervised CNN model, whereas the same reduction in the training labels affected the semi-supervised CNN to a much lesser extent (see Figure 4(a)). This shows that the semi-supervised multi-task approach is especially useful when the amount of labelled data is limited. With this semi-supervised approach, the model learns to extract meaningful deep latent features representations using the unsupervised autoencoder and learns how to classify these deep latent features using the labelled data. For D2, the effect of reducing the amount of labelled data was smaller due to the larger total amount of labelled data in D2. Still, for the supervised multi-channel CNN we can observe a similar decrease in performance, while the semi-supervised multi-channel CNN retains a similar performance level when decreasing the amount of labels.

We further compared our novel semi-supervised multi-task approach (CNN SS-MT) to another semi-supervised deep learning approach, namely, deep semi-supervised anomaly detection (CNN SS-AD). For D1, the CNN SS-AD was less accurate than all other methods (Table 1). A possible explanation for the reduced accuracy of the SS-AD method is that SS-AD was proposed for anomaly detection, with the assumption that anomalies are rare and that any unlabelled data is predominantly clean. However, the large amount of unlabelled data in D1 may have contained too many artefacts, violating the assumption of the SS-AD method. If artefacts occur too often in the unlabelled data, the SS-AD will learn to recognise them as clean (i.e., not anomalous). In contrast, the proposed CNN SS-MT method does not make any assumptions on the unlabelled data to learn from it: it merely learns to compress and decompress the data. This assumption-less characteristic of the multi-task approach could explain why it is a more effective deep semi-supervised method for artefact detection than the SS-AD approach. Nevertheless, the results from dataset D2 indicate that the SS-AD approach can be as effective as our CNN SS-MT approach if the data is well-labelled. The more labels are available, the more similar this SS-AD method becomes to supervised classifiers, as the unsupervised part of this method is only determined by the unlabelled data. In contrast, with our multi-task approach, the unsupervised part is determined by the entire (unlabelled + labelled) dataset.

To develop a method for D1 with higher accuracy, we hypothesise that the bottleneck is not the model design, but the labels. In the results section, only the metrics computed on the test dataset are shown, but we noticed a large gap between the test and validation scores for all models, as shown in Table A5 in the Appendix (see supplementary materials). This reveals that the model performs very well on both the training and validation data without overfitting, while performing significantly worse on the test data. Additional analysis of the recordings and their labels showed that the model works especially well for detecting high-amplitude artefacts, and less well for low amplitude artefacts, see Appendix Figure A1. We know that the training and validation sets – given that this data was labelled by non-experts – contained mainly labels for artefacts that are easier to recognise, such as high-amplitude variations. Conversely, the test set - labelled by an expert - included not only high-amplitude artefacts but also less obvious types of artefacts that the model thus has not seen during the training phase. Therefore, the labellers, and thereby the types of artefacts that are labelled, have a major influence on what the model learns and on the reported accuracy and this should be taken into consideration when interpreting the results. To support this hypothesis, we applied our method to a dataset with a larger number of labels from expert annotators (dataset D2). As expected, the CNN models reached higher F1 scores (near 95%) on the test set.

Furthermore, the large gap between validation and test scores that we observed in D1 was absent in D2 (see Table A6 in the Appendix).

Artefact detection itself is not an objective on its own. Instead, it can provide a gateway towards more robust and reliable applications of other automated EEG analyses. One example of such an application where artefact detection can improve the robustness of subsequent analyses, is functional brain age (FBA) estimation [10,11]. This FBA algorithm analyses the maturation of the brain based on the patterns observed in the neonatal EEG. In this paper, an existing model for the estimation of the FBA was used to predict the PMA of the neonate based on a 30-second segment of eight-channel EEG. As a proof of concept, we showed the FBA error for varying levels of artefact contamination in Figure 5. Artefact contamination levels computed with either the proposed single- or multi-channel artefact detection models were related to the error of the FBA estimation. EEG segments in which more artefacts were detected were related to larger FBA errors. This illustrates that the proposed artefact detection method could be used to select clean data epochs, which can improve the accuracy of subsequent automated EEG analysis. Additionally, the histogram in the background of Figure 5 shows that segments with a high artefact contamination level are not that rare. This stresses the importance of implementing automated methods for the identification of such low-quality EEG segments that affect the reliability of automated analyses.

It is important to consider the limitations of this research. First, it is worth considering that it is inevitable that there are uncertainties in the ground truth labels, given that labelling EEG data is subjective and experience-based, without clear guidelines [15]. Therefore, classification scores should be interpreted with care as they do not only depend on the accuracy of the model, but also on the accuracy of the ground truth labels. Future studies on the inter-rater agreement could provide insight regarding the uncertainties in the ground truth labels and provide a benchmark for these models.

A second limitation is that the models presented in this paper are trained and tested only on data coming from the same centre. We expect that the models do not generalize well to data from other centres that for example use different recording machines or protocols. One solution would be to train one model on multi-centre data to provide a one-model-fits-all solution. However, in practice it is often not easy to share sensitive personal data between centres. Alternatively, dedicated centre-specific models could be developed, as we did for D1 and D2. With tools such as transfer learning, a pretrained model could be tailored to a specific centre. Moreover, the fact that the proposed multi-task training approach does not require as much labelled training data as a fully supervised end-to-end approach makes the proposed method more feasible for centre-specific model development.

Finally, the small sample size has been a limiting factor. Even though we showed that our proposed approach can deal with limited amounts of labelled data in the training phase, a well-labelled dataset is required for elaborate testing and validation of the model. Future research with a larger and independent dataset is needed to address these limitations. More specifically, such research could focus on testing the generalisability of these models using data from different centres. Additionally, the sensitivity for detecting specific types of artefacts can be analysed using a well-labelled dataset to get a better insight into the strengths and shortcomings of the models.

With a reliable method for automated artefact detection, further research should focus on how to deal with the detected artefacts in application-specific cases. As a first step towards this, we showed that for brain age estimation, artefact detection could be used to select the cleanest possible data from a recording, which is associated with lower errors. However, for other applications such as sleep staging, a different approach might be needed to incorporate the results from artefact detection.

5. Conclusion

In conclusion, this paper proposes a semi-supervised deep convolutional neural network that can simultaneously learn from labelled and unlabelled data to classify EEG data as clean or artefact. The

proposed CNN model outperforms the state-of-the-art feature-based models. Additionally, the proposed semi-supervised multi-task training strategy proved to be more powerful than a fully supervised strategy when a low amount of labelled data is available. Therefore, this semi-supervised training strategy can be a solution for developing deep learning models for signals for which ground truth labels are scarce. Using FBA estimation as an example, we showed that the automated detection of artefacts can improve robustness and reliability in automated analysis of neonatal EEG.

Acknowledgements

This research received funding from the European Union under Grant Agreement #813483: H2020 MSCA-ITN-2018: 'INtegrating Functional Assessment measures for Neonatal Safeguard (INFANS)'; from the Horizon 2020 Framework Programme of the European Union as part of the COST action ‘Maximising impact of multidisciplinary research in early diagnosis of neonatal brain injury’ (AI-4-NICU) CA20124; from the Research Foundation—Flanders in the frame of the FWO Research Project: 'Deep, personalized epileptic seizure detection', G0D8321N; and from the Flemish Government (AI Research Program). SVH, MDV and TH are affiliated to Leuven.AI - KU Leuven institute for AI, B-3000, Leuven, Belgium.

References

[1] Abdelhameed A M and Bayoumi M 2019 Semi-supervised EEG signals classification system for epileptic seizure detection *IEEE Signal Process. Lett.* **26** 1922–6

[2] Ansari A H, Cherian P J, Caicedo A, Naulaers G, Vos M D and Huffel S V 2019 Neonatal Seizure Detection Using Deep Convolutional Neural Networks *Int. J. Neural Syst.* **29** 1850011

[3] Becker T, Vandecasteele K, Chatzichristos C, Van Paesschen W, Valkenborg D, Van Huffel S and De Vos M 2021 Classification with a deferral option and low-trust filtering for automated seizure detection *Sens. Switz.* **21** 1–18

[4] Raeisi K, Khazaei M, Croce P, Tamburro G, Comani S and Zappasodi F 2022 A graph convolutional neural network for the automated detection of seizures in the neonatal EEG *Comput. Methods Programs Biomed.* **222** 106950

[5] Dereymaeker A, Pillay K, Vervisch J, Van Huffel S, Naulaers G, Jansen K and De Vos M 2017 An automated quiet sleep detection approach in preterm infants as a gateway to assess brain maturation *Int. J. Neural Syst.* **27**

[6] Pillay K, Dereymaeker A, Jansen K, Naulaers G, Van Huffel S and De Vos M 2018 Automated EEG sleep staging in the term-age baby using a generative modelling approach *J. Neural Eng.* **15**

[7] Ansari A H, De Wel O, Pillay K, Dereymaeker A, Jansen K, Van Huffel S, Naulaers G and De Vos M 2020 A convolutional neural network outperforming state-of-the-art sleep staging algorithms for both preterm and term infants *J. Neural Eng.* **17** 016028

[8] Lavanga M, Smets L, Bollen B, Jansen K, Ortibus E, Van Huffel S, Naulaers G and Caicedo A 2020 A perinatal stress calculator for the neonatal intensive care unit: An unobtrusive approach *Physiol. Meas.* **41** 75012

[9] Lavanga M, De Wel O, Caicedo A, Jansen K, Dereymaeker A, Naulaers G and Van Huffel S 2018 A brain-age model for preterm infants based on functional connectivity *Physiol. Meas.* **39** 1–9

[10] Pillay K, Dereymaeker A, Jansen K, Naulaers G and De Vos M 2020 Applying a data-driven approach to quantify EEG maturational deviations in preterms with normal and abnormal neurodevelopmental outcomes *Sci. Rep.* **10** 1–14

[11] Stevenson N J, Oberdorfer L, Tataranno M-L, Breakspear M, Colditz P B, Vries L S, Benders M J N L, Klebermass-Schrehof K, Vanhatalo S and Roberts J A 2020 Automated cot-side tracking of functional brain age in preterm infants *Ann. Clin. Transl. Neurol.* **7** 891–902

[12] Duffy F H and Als H 2012 A stable pattern of EEG spectral coherence distinguishes children with autism from neuro-typical controls - a large case control study *BMC Med.* **10** 1–18

[13] Peters J M, Taquet M, Vega C, Jeste S S, Fernández I S, Tan J, Nelson C A, Sahin M and Warfield S K 2013 Brain functional networks in syndromic and non-syndromic autism: A graph theoretical study of EEG connectivity *BMC Med.* **11** 1–16

[14] Lavanga M, De Ridder J, Kotulska K, Moavero R, Curatolo P, Weschke B, Riney K, Feucht M, Krsek P, Nabbout R, Jansen A C, Wojdan K, Domanska-Pakieła D, Kaczorowska-Frontczak M, Hertzberg C, Ferrier C H, Samueli S, Jahodova A, Aronica E, Kwiatkowski D J, Jansen E, Jó´zwiak J, Lagae L, Van Huffel S and Caicedo A 2021 Results of quantitative EEG analysis are associated with autism spectrum disorder and development abnormalities in infants with tuberous sclerosis complex *Biomed. Signal Process. Control* **68** 102658

[15] Malfilâtre G, Mony L, Hasaerts D, Vignolo-Diard P, Lamblin M-D and Bourel-Ponchel E 2021 Technical recommendations and interpretation guidelines for electroencephalography for premature and full-term newborns *Neurophysiol. Clin.* **51** 35–60

[16] Khlif M S, Mesbah M, Boashash B and Colditz P 2010 Influence of EEG artifacts on detecting neonatal seizure *10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010)* 2010 10th International Conference on Information Sciences, Signal Processing and their Applications (ISSPA) (Kuala Lumpur, Malaysia: IEEE) pp 500–3

[17] Vos M D, Deburchgraeve W, Cherian P J, Matic V, Swarte R M, Govaert P, Visser G H and Huffel S V 2011 Automated artifact removal as preprocessing refines neonatal seizure detection *Clin. Neurophysiol.* **122** 2345–54

[18] Matic V, Cherian P J, Jansen K, Koolen N, Naulaers G, Swarte R M, Govaert P, Van Huffel S and De Vos M 2016 Improving Reliability of Monitoring Background EEG Dynamics in Asphyxiated Infants *IEEE Trans. Biomed. Eng.* **63** 973–83

[19] Sadiya S, Alhanai T and Ghassemi M M 2021 Artifact Detection and Correction in EEG data: A Review *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)* 2021 10th International IEEE/EMBS Conference on Neural Engineering (NER) (Italy: IEEE) pp 495–8

[20] Stevenson N J, O'Toole J M, Korotchikova I and Boylan G B 2014 Artefact detection in neonatal EEG *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (Chicago, IL: IEEE) pp 926–9

[21] Bhattacharyya S, Biswas A, Mukherjee J, Majumdar A K, Majumdar B, Mukherjee S and Singh A K 2013 Detection of artifacts from high energy bursts in neonatal EEG *Comput. Biol. Med.* **43** 1804–14

[22] Webb L, Kauppila M, Roberts J A, Vanhatalo S and Stevenson N J 2021 Automated detection of artefacts in neonatal EEG with residual neural networks *Comput. Methods Programs Biomed.* **208** 106194

[23] Kauppila M, Vanhatalo S and Stevenson N J 2018 Artifact detection in neonatal EEG using Gaussian mixture models *EMBEC & NBC 2017* ed H Eskola, O Väisänen, J Viik and J Hyttinen (Singapore: Springer Singapore) pp 221–4

[24] Yang X, Song Z, King I and Xu Z 2021 A Survey on Deep Semi-supervised Learning

[25] Wen T and Zhang Z 2018 Deep Convolution Neural Network and Autoencoders-Based Unsupervised Feature Learning of EEG Signals *IEEE Access* **6** 25399–410

[26] Ruff L, Vandermeulen R A, Görnitz N, Binder A, Müller E, Müller K-R and Kloft M 2019 Deep Semi-Supervised Anomaly Detection

[27] Stevenson N J, Tapani K, Lauronen L and Vanhatalo S 2019 A dataset of neonatal EEG recordings with seizure annotations *Sci. Data* **6** 190039

[28] TensorFlow Developers 2022 TensorFlow

[29] Chollet F 2015 Keras

[30] Ansari A H, Pillay K, Dereymaeker A, Jansen K, Van Huffel S, Naulaers G and De Vos M 2022 Brain-age as an estimator of neurodevelopmental outcome: A deep learning approach for neonatal cot-side monitoring (in preparation)

[31] Ansari A H, Pillay K, Dereymaeker A, Jansen K, Van Huffel S, Naulaers G and De Vos M 2022 A Deep Shared Multi-Scale Inception Network Enables Accurate Neonatal Quiet Sleep Detection With Limited EEG Channels *IEEE J. Biomed. Health Inform.* **26** 1023–33